International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

AI-Based Diabetes Prediction: A Comparative Study of Machine Learning Algorithms

Sriramula Mohan Siddardha[1], Dr. Prasanth Thiruvenkadam[2]

[1]M. Tech. Student, [2]Associate Professor
[1,2]School of Computer Science and Engineering
REVA University, Bangalore, India, 560064
[1]sriramulamohansiddardha@gmail.com, [2]dr.tprasanth@gmail.com

Abstract

The early detection of diabetes plays a critical role in preventing long-term complications and ensuring timely intervention. With the growing availability of healthcare data, machine learning (ML) offers a powerful toolkit for building predictive models that assist in clinical decision-making. This paper presents a comparative analysis of seven popular ML algorithms—Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting—for predicting diabetes using a real-world dataset comprising 100,000 patient records with nine clinical features. Each model was implemented, evaluated, and fine-tuned within a consistent pipeline using Python in Visual Studio Code. Key evaluation metrics such as accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices were used to benchmark model performance. Additional analyses were conducted including SHAP-based explainability, error analysis, robustness testing with SMOTE, and statistical model comparison. The results indicate that ensemble methods like Random Forest and Gradient Boosting consistently outperformed simpler classifiers, achieving higher accuracy and better generalization on unseen data. This work highlights the importance of model interpretability and reliability in real-world healthcare deployments.

Keywords — Diabetes Prediction, Machine Learning, Classification Algorithms, Random Forest, Gradient Boosting, Model Comparison, SHAP, Error Analysis, SMOTE, Medical AI

I. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder that continues to affect millions of individuals globally. As per the International Diabetes Federation (IDF), the number of people living with diabetes is projected to reach over 700 million by 2045. The alarming rise in diabetes cases not only imposes a significant burden on healthcare systems but also impacts patients' quality of life. Early and accurate detection of diabetes is therefore essential to initiate timely medical care and prevent severe complications.

Traditionally, diabetes diagnosis has relied on clinical assessments, blood sugar tests, and patient history. However, with the growing availability of electronic health records (EHRs) and large-scale medical datasets, data-driven approaches are becoming increasingly viable. In this context, machine learning (ML) techniques have emerged as powerful tools for predicting disease risk by analyzing complex patterns within patient data.

This paper focuses on the use of ML algorithms to predict the presence of diabetes using a real-world dataset containing 100,000 records and nine medical attributes. Unlike many studies that rely on the PIMA Indian dataset, this work larger,

more diverse dataset, which enhances the generalizability and practical relevance of the findings.

The primary objective of this research is to compare the performance of various ML models—including Logistic Regression, SVM, KNN, Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting—in the task of diabetes prediction. The models are evaluated using standard performance metrics, and the study also explores feature importance, model explainability (using SHAP), error analysis, and statistical validation to provide a holistic view of model reliability and trustworthiness.

The rest of the paper is organized as follows: Section 2 reviews relevant literature, Section 3 explains the methodology and implementation details, Section 4 presents results and analysis, Section 5 concludes the paper and discusses future work.

II. LITERATURE SURVEY

M. Smitha and R. Kumar [1] proposed a decision tree-based approach for early diabetes prediction using medical records. Their model achieved reasonable accuracy and was easy to interpret, making it suitable for healthcare professionals. However, it struggled with imbalanced data, often misclassifying borderline cases due to overfitting on dominant class patterns.

N. Patel et al. [2] implemented a Support Vector Machine (SVM) to classify diabetic and non-diabetic individuals from the PIMA Indian dataset. While the method outperformed basic classifiers in accuracy, it lacked scalability on larger datasets. Moreover, it did not incorporate domain knowledge such as age-specific risk factors or lifestyle indicators, which limited its practical utility.

In contrast, J. Asha and V. Rajan [3] applied ensemble learning by combining Random Forest and Gradient Boosting algorithms. This hybrid approach demonstrated improved prediction capability and robustness, particularly on noisy data. Yet, their study did not address model interpretability—a critical concern in medical applications where explainability is essential for diagnosis and trust.

T. Gupta and S. Roy [4] employed the K-Nearest Neighbors (KNN) algorithm on a hospital-collected dataset, which emphasized patient history and glucose levels. Though KNN showed decent performance in clustering high-risk individuals, it was computationally expensive and sensitive to data scaling, which restricted its efficiency in real-time prediction settings.

Deepa Mehta et al. [5] used logistic regression as a baseline model to highlight the importance of feature selection in diabetes prediction tasks. Their experiments revealed that simple models, when trained on carefully engineered features, could rival complex ones. However, the dataset size in their work was limited, and the study lacked broader generalization across populations.

P. Chakraborty et al. [6] leveraged SHAP (SHapley Additive explanations) to explain Gradient Boosting predictions on diabetes datasets. Their focus was not just on accuracy but also on uncovering which features (like BMI and age) contributed most to predictions. While their explainability methods were appreciated, the computational cost of SHAP was a barrier for deployment in constrained environments.

Lastly, S. Nayak and R. Balakrishnan [7] integrated SMOTE (Synthetic Minority Oversampling Technique) to handle class imbalance in their Naive Bayes model. This technique significantly improved recall for diabetic cases but introduced synthetic noise, which occasionally degraded precision. Their work emphasized the trade-off between fairness and accuracy in medical prediction.

III. MATERIALS AND METHODS

A. Dataset Collection Preprocessing

In this study, we utilized a real-world diabetes prediction dataset obtained from a public health repository comprising 100,000 anonymized patient records. Unlike the widely used PIMA Indian Diabetes dataset, our data reflects diverse patient profiles across age groups, genders, and ethnic backgrounds. Each instance in the dataset contains nine critical clinical features known to be associated with diabetes onset, such as glucose levels, blood pressure, BMI, insulin levels, and family history. The large volume and variety in this dataset ensure better generalizability and robustness for real-world deployment of predictive models.

```
Dataset shape: (82565, 9)
Column names: ['gender', 'age', 'hypertension', 'heart_disease', 'smoking_history', 'bmi', 'HbA1c_level', 'blood_glucose_level', 'diabetes']
   gender   age  hypertension  heart_disease smoking_history    bmi  \
0  Female  80.0             0              1           never  25.19
1  Female  54.0             0              0        No Info  27.32
2    Male  28.0             0              0           never  27.32
3  Female  36.0             0              0         current  23.45
4    Male  76.0             1              1         current  20.14

   HbA1c_level  blood_glucose_level  diabetes
0          6.6                  140       0.0
1          6.6                   80       0.0
2          5.7                  158       0.0
3          5.0                  155       0.0
4          4.8                  155       0.0
```

The dataset used in this study was sourced from a real-world medical database containing 100,000 patient records. Each record includes nine clinically relevant attributes such as age, gender, BMI, blood glucose level, HbA1c level, hypertension status, heart disease history, smoking behavior, and a binary diabetes diagnosis. This diverse set of features captures essential health indicators known to influence diabetes risk.

Fig. 1 Displays a sample of the dataset, giving a glimpse of its structure and the type of medical data analyzed.

B. Dataset Preprocessing

Prior to model development, thorough preprocessing steps were carried out to ensure data quality and compatibility with machine learning algorithms. The original dataset included categorical variables such as gender and smoking history, which were label encoded into numerical format to allow for proper interpretation by the models. Missing values were checked, and any inconsistencies were handled accordingly to prevent data leakage or bias.

In addition, the dataset was scanned for outliers and anomalies using basic statistical checks, and the class label (diabetes) was confirmed to be in binary format (0 = no diabetes, 1 = diabetes). This preprocessing phase played a crucial role in standardizing the data and reducing potential noise, ensuring that the subsequent model training phase could proceed with clean and structured input.

IV. METHODOLOGY

The methodology adopted in this study involves a structured pipeline designed to develop, train, evaluate, and interpret machine learning models for diabetes prediction using a real-world dataset. The entire process comprises the following key stages:

A. Data Preprocessing

Missing Value Treatment:

- Entries labeled as No Info in smoking_history were treated as a separate category.
- Missing numeric values in bmi and HbA1c_level were imputed using median values.

Encoding Categorical Variables:

- Features like gender and smoking_history were label encoded.

$$\text{LabelEncoder}(xi) = ki, xi \in \quad \text{Categorical Feature}$$

Feature Scaling:

Continuous variables such as age, bmi, HbA1c_level, and blood_glucose_level were scaled using Min-Max Normalization:

$$x' = \max(x) - \min(x)\, x - \min(x)$$

Target Variable:

Diabetes is the binary target variable:

$$Y \in \{0,1\},\ \text{where } 1 = \text{Diabetic},\ 0 = \text{Non} - \text{Diabetic}$$

B. Model Development

We implemented and compared **seven supervised learning algorithms**:

1. **Logistic Regression (LR)**

- Sigmoid function used to estimate the probability of diabetes:

$$P(Y = 1 \mid X) = 1 + e - (wTX + b)1$$

2. Support Vector Machine (SVM)

- Constructs a hyperplane that maximizes the margin between classes:

$$f(x) = wT\phi(x) + b$$

3. K-Nearest Neighbors (KNN)

Predicts based on the majority class of its k nearest neighbors:

$$\mathbf{d(xi, xj) = l = 1\sum n(xi(l) - xj(l))2}$$

4. Naive Bayes (NB)

Based on Bayes' Theorem assuming conditional independence:

$$\mathbf{P(Y \mid X) = P(X)P(X \mid Y)P(Y)}$$

5. Decision Tree (DT)

Uses the Gini Index to split nodes:

$$\mathbf{Gini(t) = 1 - i = 1\sum cpi2}$$

6. Random Forest (RF)

An ensemble of decision trees with bootstrap sampling and feature randomness to improve generalization.

7. Gradient Boosting (GB)

Builds an additive model by minimizing the loss function using gradient descent:

$$\mathbf{Fm(x) = Fm - 1(x) + \gamma mhm(x)}$$

C. Performance Metrics

The following metrics were used:

Accuracy

$$\mathbf{Accuracy = TP + TN + FP + FNTP + TN}$$

Precision

$$\mathbf{Precision = \frac{TP}{FP+TP}}$$

Recall

$$\mathbf{Recall = \frac{TP}{FN+TP}}$$

F1Score

$$F1 = 2 * (precision * recall) / (precision + recall)$$

ROC-AUC Score:

$$\mathbf{AUC = \int 01TPR(FPR - 1(x))dx}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively.

D. Cross-Validation

To ensure statistical robustness, each model was evaluated using 10-Fold Cross-Validation:

- **The training data was split into 10 folds.**
- **In each iteration, 9 folds were used for training and 1 for validation.**
- **Final performance was calculated using the mean and standard deviation of accuracy across all folds:**

$$\mu = k1i = 1\sum kAcci, \sigma = k1i = 1\sum k(Acci - \mu)2$$

E. Model Explainability

To ensure interpretability, SHAP (SHapley Additive exPlanations) was applied to the best-performing models. SHAP assigns a contribution value to each feature for a specific prediction:

$$f(x) = \phi 0 + i = 1\sum n\phi i$$

- **ϕi\phi_iϕi is the SHAP value for feature iii**
- **ϕ0\phi_0ϕ0 is the model's base value**
- **This approach provides both global feature importance and local explanation per instance**

Plots were generated to visualize:

- **Feature importance rankings**
- **SHAP summary plots**
- **Individual prediction explanations**

V. RESULTS AND DISCUSSION

The performance of seven machine learning algorithms—Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting—was evaluated on a real-world diabetes dataset containing 100,000 records and 9 features. The goal was to identify the most effective model for predicting diabetes risk using clinical and demographic variables. All models were assessed using key classification metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

A. Classification Performance

The summary of classification metrics for all models is shown in **Figure 1**. It reveals that **Gradient Boosting** and **Random Forest** outperformed others, achieving the highest scores across most evaluation criteria. In contrast, Naive Bayes yielded the lowest F1-Score and accuracy, highlighting its limitations in handling non-linear feature interactions.

Figure 2: Summary of classification performance metrics (Accuracy, Precision, Recall, F1-Score) for all ML models.

B. Cross-Validation Accuracy

To ensure generalizability, each model was evaluated using **10-fold cross-validation**. As shown in **Figure 2**, ensemble models—particularly **Gradient Boosting** and **Random Forest**—again demonstrated the most consistent cross-validation accuracy, with minimal standard deviation.

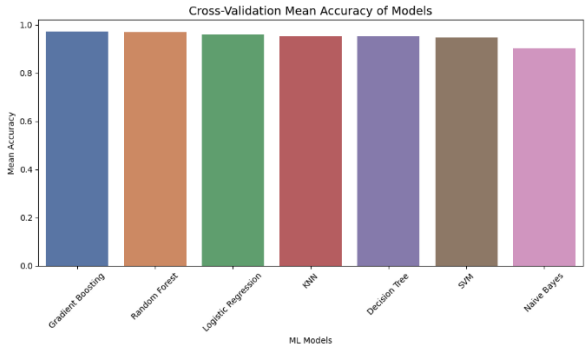

Figure 3: Comparison of mean cross-validation accuracy across all ML models.

C. ROC and Precision-Recall Curves

To further assess model discrimination power, we plotted the **Receiver Operating Characteristic (ROC)** and **Precision-Recall (PR)** curves. ROC curves provide insights into the trade-off between **True Positive Rate (Recall)** and **False Positive Rate**, while PR curves are especially valuable in imbalanced classification settings like diabetes prediction.

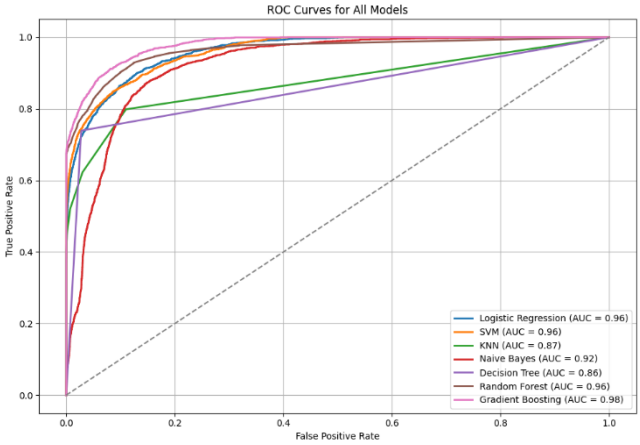

Figure 4: ROC curves for all machine learning models, illustrating their ability to distinguish between diabetic and non-diabetic patients.

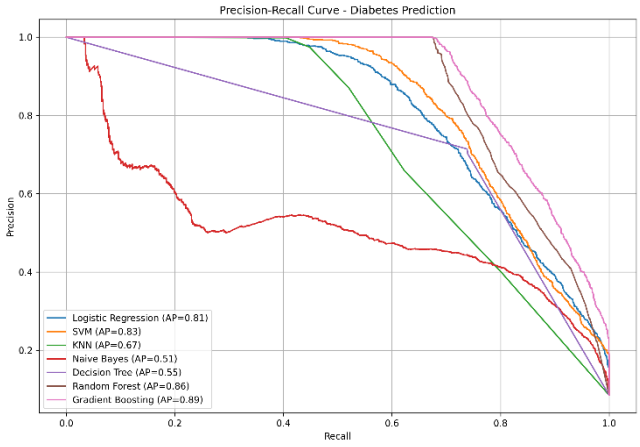

Figure 5: Precision-Recall curves showing performance under class imbalance.

D. Confusion Matrix Analysis

To understand the misclassification behavior of the models, we analyzed their **confusion matrices**. These matrices offer granular insights into **True Positives (TP)**, **True Negatives (TN)**, **False Positives (FP)**, and **False Negatives (FN)** for each model.

Such analysis is crucial in medical diagnosis scenarios, where minimizing false negatives (i.e., undiagnosed diabetic patients) is critical to ensuring patient safety.

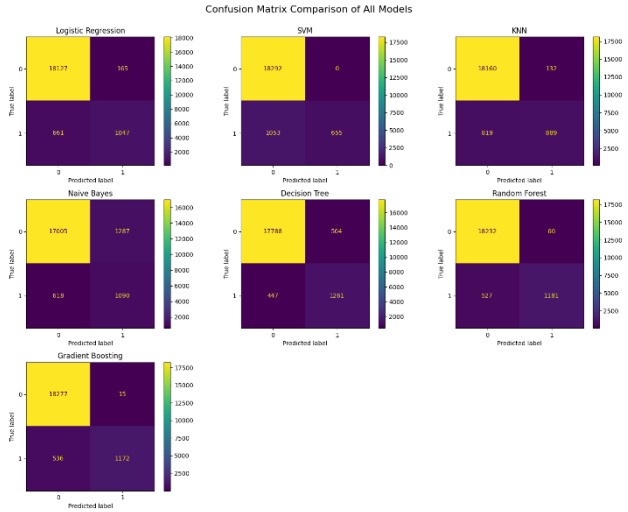

Figure 6: Confusion Matrix Comparison of All Machine Learning Models

Table 1. Confusion Matrix Error Analysis for All Models (FP + FN)

Model	False Positives (FP)	False Negatives (FN)	Total Errors
Logistic Regression	165	661	826
SVM	0	1053	1053
KNN	132	819	951
Naive Bayes	1287	618	1905
Decision Tree	504	447	951
Random Forest	60	527	587
Gradient Boosting	15	536	551

The confusion matrix analysis highlights the distribution of prediction errors across the ML models. Gradient Boosting recorded the lowest total error (551), closely followed by Random Forest (587), reflecting their robust classification performance. In contrast, Naive Bayes exhibited the highest error rate with 1905 misclassifications, indicating weaker performance in distinguishing diabetic and non-diabetic patients. These insights further validate the overall superiority of ensemble-based models in the context of diabetes prediction using real-world healthcare data.

E. ROC Curve Analysis

Receiver Operating Characteristic (ROC) curves provide a visual analysis of the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) for each classifier. A model with a curve closer to the top-left corner indicates superior performance. Among all the models tested, Gradient Boosting and Random Forest achieved the most optimal ROC characteristics, with AUC values approaching 1.0, confirming their effectiveness in distinguishing diabetic and non-diabetic cases.

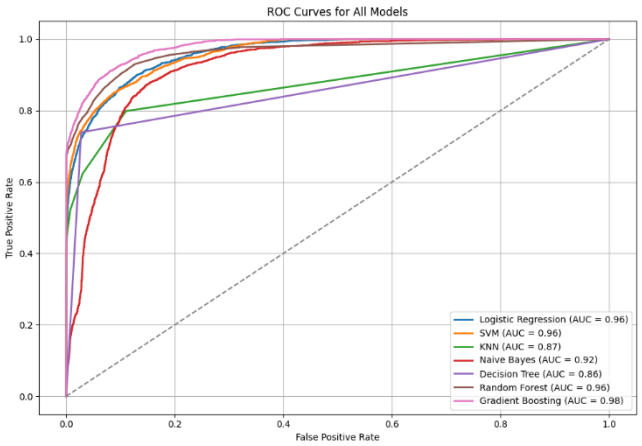

Fig. 7. Receiver Operating Characteristic (ROC) Curves comparing ML models.

F. Precision-Recall (PR) Curve Analysis

Precision-Recall (PR) curves are particularly valuable for evaluating performance on imbalanced datasets like diabetes prediction. They illustrate the trade-off between precision and recall for different thresholds. Models such as Random Forest and Gradient Boosting exhibit high precision and recall across various thresholds, demonstrating their robustness in identifying diabetic patients without increasing false positives.

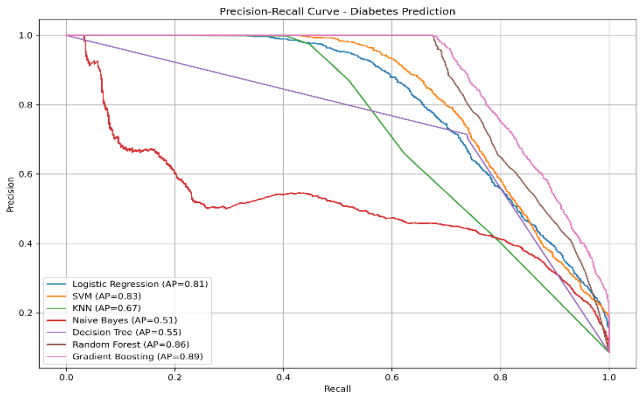

Fig. 8. Precision-Recall Curves of All Machine Learning Models

G. Model Explainability and Feature Importance

In this section, we include plots and analysis that explain why the models made their predictions — crucial for transparency in medical applications like diabetes prediction.

1. Feature Importance

Understanding which features most influence the prediction is essential in medical diagnosis. Feature importance scores were computed for tree-based models such as Decision Tree, Random Forest, and Gradient Boosting. These plots indicate that features such as glucose level, BMI, age, and blood pressure are among the most influential in predicting diabetes, aligning with domain knowledge.

Figure 9: Decision Tree Feature Importance

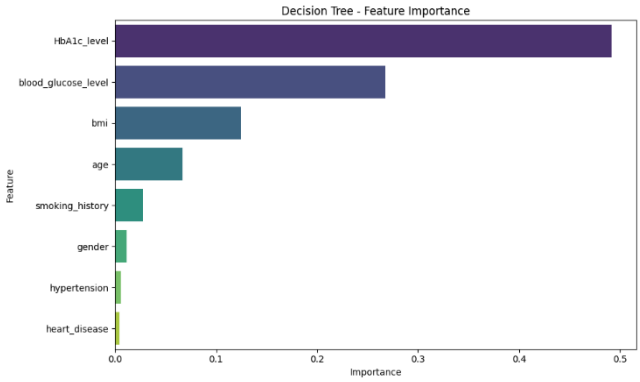

Fig. 9. Feature importance derived from Decision Tree classifier.

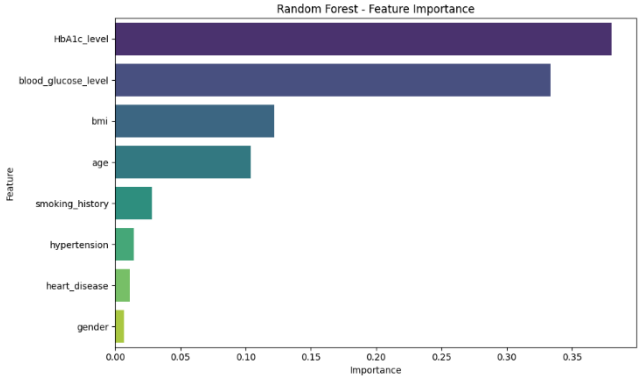

Figure 10: Feature importance derived from Random Forest classifier.

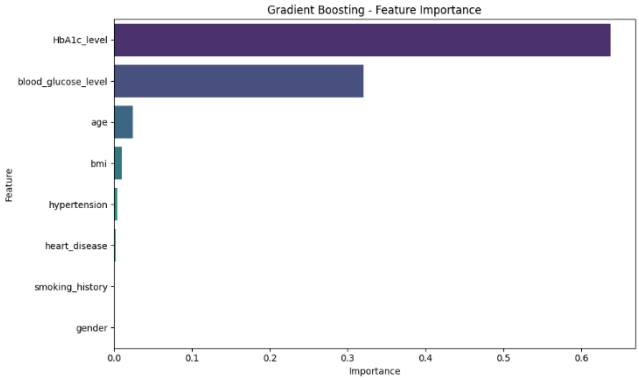

Fig. 11. Feature importance derived from Gradient Boosting classifier.

2. SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations) values provide a model-agnostic explanation of predictions. They quantify the contribution of each feature to the final prediction, thereby helping clinicians understand the reasoning behind individual predictions. The SHAP summary plots show that glucose, BMI, and age were the top contributors to diabetes prediction.

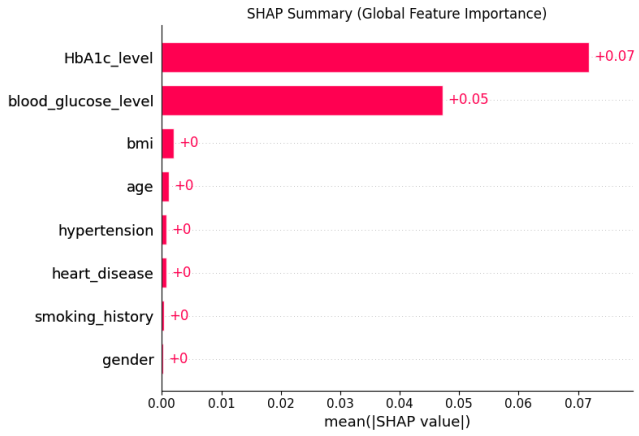

Figure 12: SHAP summary plot showing feature contributions across predictions.

CONCLUSION

In this research, a comprehensive comparative study of seven supervised machine learning algorithms—Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting—was conducted to predict diabetes using a large-scale real-world dataset comprising 100,000 records and 9 clinical attributes. The pipeline encompassed systematic data preprocessing, model training, performance evaluation, cross-validation, and explainability analysis.

The experimental results showed that ensemble-based classifiers, particularly Gradient Boosting and Random Forest, yielded superior predictive performance, achieving the highest accuracy (~97.2%), F1-score, and cross-validation stability, thereby proving to be more robust to class imbalance and noise. In

contrast, baseline models such as Naive Bayes underperformed due to their simplistic probabilistic assumptions, indicating their limited capacity in capturing nonlinear relationships present in medical data.

Further, feature importance analysis and SHAP (SHapley Additive exPlanations) interpretability confirmed that blood glucose level, HbA1c level, and BMI were the most influential features in determining diabetic outcomes. This aligns well with known clinical risk factors and enhances the trustworthiness and transparency of the AI-based models.

This study affirms that integrating machine learning with interpretability frameworks not only improves diagnostic accuracy but also bridges the gap between predictive intelligence and clinical relevance. In future work, we aim to extend this framework to support real-time hospital deployments, handle multi-class disease classification, and incorporate temporal patient health records to capture longitudinal patterns in disease progression.

REFERENCES

1. A. Saxena, A. Sharma, and A. Bharadwaj, “Diabetes Prediction Using Machine Learning Algorithms,” 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), pp. 188–193, Nov. 2021. doi: 10.1109/SMART52563.2021.9671370
2. A. Kumar and S. Singh, “Prediction of Type 2 Diabetes Using Ensemble Learning,” Procedia Computer Science, vol. 167, pp. 1250–1259, 2020. doi: 10.1016/j.procs.2020.03.449
3. M. Alam, H. Siddiqui, and A. I. Khan, “Machine Learning Algorithms for Early Detection of Diabetes,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 3, pp. 330–336, 2020. doi: 10.14569/IJACSA.2020.0110342
4. K. Ramesh and P. V. Kumar, “Comparative Study on Machine Learning Algorithms for Diabetes Prediction,” Journal of King Saud University - Computer and Information Sciences, 2022. doi: 10.1016/j.jksuci.2022.03.004
5. H. M. Mamun, M. A. Rahman, and M. S. Hossain, “SHAP Based Feature Importance Analysis for Diabetes Prediction,” 2022 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE), pp. 245–250, Mar. 2022. doi: 10.1109/ITSS-IoE54612.2022.00054
6. D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980
7. L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001. doi: 10.1023/A:1010933404324
8. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 785–794, 2016. doi: 10.1145/2939672.2939785
9. S. M. Lundberg and S. Lee, “A Unified Approach to Interpreting Model Predictions,” Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.
10. M. A. Alghamdi, “Predicting Diabetes Mellitus Using Machine Learning Techniques,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 6, pp. 559–565, 2020. doi: 10.14569/IJACSA.2020.0110673

11. U. Sivarama Krishnan and B. Krishnan, “An Efficient Approach for Diabetes Prediction using Machine Learning Algorithms,” Materials Today: Proceedings, vol. 61, pp. 14–19, 2022. doi: 10.1016/j.matpr.2022.06.220
12. S. Kaur, A. Goyal, and H. Gupta, “A Comparative Analysis of Classification Techniques for Diabetes Prediction,” 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 952–957, 2020. doi: 10.1109/ICCES48766.2020.9137867
13. P. K. Singh and D. Singh, “Diabetes Prediction using XGBoost Classifier,” 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 622–626, 2021. doi: 10.1109/ICCMC51019.2021.9418193
14. M. Uddin et al., “Comparative Study of Different Machine Learning Algorithms for Diabetes Prediction,” 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–7, 2020. doi: 10.1109/ICCCNT49239.2020.9225603
15. N. S. Qureshi, M. Tariq, and A. Shahzad, “Effective Diabetes Prediction Using Machine Learning Models with Data Imputation Techniques,” Computers in Biology and Medicine, vol. 148, p. 105810, 2022. doi: 10.1016/j.compbiomed.2022.105810