

## Research Paper

# EVALUATION OF MACHINE LEARNING CLASSIFIERS FOR EARLY-STAGE TYPE 2 DIABETES RISK PREDICTION

Sadula Rushidhar

Department of Computer Science and Engineering  
Aurora Deemed to be University  
Hyderabad, Telangana, India  
[sadularushidhar1999@gmail.com](mailto:sadularushidhar1999@gmail.com)

Dr. S. Mahesh

Department of Computer Science and Engineering  
Aurora Deemed to be University  
Hyderabad, Telangana, India  
[mahesh@aurora.edu.in](mailto:mahesh@aurora.edu.in)

**Abstract**—Early diagnosis of chronic metabolic problems is crucial to improve preventive healthcare and reduce the long-term effects of Type 2 Diabetes Mellitus (T2DM). Since diabetes is mostly undetected in its early stages, accurate prediction models can help for timely therapeutic intervention. The study assesses the performance of three commonly used machine learning algorithms, Logistic Regression, Support Vector Machine (SVM) and Random Forest, for early prediction of diabetes risk utilizing demographic, clinical, lifestyle and biochemical data. The study used a standardized healthcare dataset which consisted of patient demographics, body mass index (BMI), blood pressure, fasting plasma glucose (FPG), lipid profile, kidney function markers, lifestyle factors and family history. The data pre-processing involved missing value treatment, duplicate record removal, feature scaling, and balancing the training data using the Synthetic Minority Oversampling Technique (SMOTE). Models were evaluated based on Accuracy, Precision, Recall, F1-score and Receiver Operating Characteristic Area Under Curve (ROC-AUC). In general, the Support Vector Machine classifier was the best one with about 89% accuracy and a ROC-AUC of 0.95. Logistic Regression and Random Forest also had a good predictive power. Feature importance analysis showed that fasting plasma glucose (FPG) and age were the most important predictors of diabetes, followed by HDL, LDL and systolic blood pressure. These results indicate that machine learning approaches can be reliable and efficient tools to assist the decisions related to early screening and risk assessment of diabetes. The proposed framework can be used in the digital healthcare system to help doctors in detecting high-risk individuals and advocating timely preventative actions.

**Index Terms**—Type 2 Diabetes Mellitus, Machine Learning, Logistic Regression, Support Vector Machine, Random Forest, Diabetes Prediction, Healthcare Analytics

## I. INTRODUCTION

T2DM is one of the most prevalent chronic diseases worldwide and a major public health challenge. The disease is caused by insulin resistance or a lack of production of insulin. This causes chronically elevated levels of glucose in the blood. The number of people with diabetes continues to grow all over the world, creating more pressure on healthcare services and affecting millions of people every year, according to the International Diabetes Federation (IDF). Untreated diabetes can cause serious complications such as cardiovascular disease, kidney failure, neuropathy, retinopathy, and stroke. Therefore, early identification of high risk individuals is important for effective prevention and disease management.

The traditional approach to diagnosing diabetes relies on laboratory tests, including fasting plasma glucose (FPG), oral glucose tolerance tests, and glycated hemoglobin (HbA1c). These diagnostic methods are clinically reliable, but they are usually carried out after the onset of symptoms or during routine medical examinations. Many patients with early-stage type 2 diabetes (T2D) are undiagnosed prior to the onset of complications, as the disease often remains asymptomatic. This limitation has driven researchers to explore intelligent computational methods that can identify high-risk individuals from routinely available healthcare data.

Recent breakthroughs in Machine Learning (ML) have greatly enhanced predictive analytics for medical diagnosis. ML algorithms can learn complex relationships between demographic, physiological, laboratory and lifestyle variables to accurately predict disease risk and support clinical decision making. The popularity of Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) stems from their good predictive performance, robustness and interpretability, which are considered popular supervised learning algorithms.

This study evaluates three supervised machine learning classifiers, Logistic Regression, Support Vector Machine, and Random Forest, for early stage diabetes prediction on a standardized healthcare dataset with demographic information, body mass index (BMI), blood pressure, biochemical parameters, lifestyle factors, and family history. The preprocessing of data consists of missing value handling, duplicate removal, feature standardization, and Synthetic Minority Oversampling Technique (SMOTE) to handle class imbalance. The classification performance is evaluated using Accuracy, Precision, Recall, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC).

In addition to classification, this work also investigates the prediction of Follow-up Fasting Plasma Glucose (FFPG) using Linear Regression, Random Forest Regressor, and Support Vector Regression (SVR). This extra regression analysis provides an insight into the future estimation of glucose and enhances the diabetes classification framework.

The main contributions of this study are summarized as follows:

- Development of a complete machine learning framework for early diabetes prediction.

- Comparative evaluation of Logistic Regression, Support Vector Machine, and Random Forest classifiers.
- Comparative analysis of Linear Regression, Random Forest Regression, and Support Vector Regression for FPG prediction.
- Comprehensive exploratory data analysis to identify important clinical characteristics.
- Identification of influential clinical features contributing to diabetes prediction.

The rest of this paper is structured as follows. Section II is a review of the existing literature on diabetes prediction. III. PROPOSED APPROACH AND ML MODELS The experimental results and discussion are presented in Sec. IV. Finally, Section V concludes the paper and discusses future directions for research.

## II. RELATED WORK

Machine learning (ML) has emerged as a successful strategy for predicting Type 2 Diabetes Mellitus (T2DM) based on clinical and laboratory data. A variety of research have shown that supervised learning algorithms can improve early diagnosis by recognizing complicated patterns among patient characteristics.

Logistic Regression is one of the most used baseline classifiers because of its simplicity and interpretability Support Vector Machine (SVM) has been demonstrated to perform well in high dimensional medical datasets by generating optimal decision boundaries. Random Forest (RF) produces stable predictions by using ensemble learning and can also give feature importance analysis.

SVM and Random Forest have been found to perform better than traditional statistical methods for diabetes prediction in recent studies. Preprocessing techniques such as feature scaling, missing value imputation, and Synthetic Minority Oversampling Technique (SMOTE) have been shown to improve classification performance especially with imbalanced healthcare datasets.

While there are many studies on diabetes classification, there are relatively few studies that combine exploratory data analysis, classification, regression and feature importance analysis in one framework. Moreover, the majority of existing systems only measure the classification accuracy and do not predict glucose further.

To overcome these limitations, this study proposes an integrated framework comprising exploratory data analysis, diabetes classification, follow-up fasting plasma glucose prediction, and feature importance analysis. We compare three classification algorithms (Logistic Regression, Support Vector Machine and Random Forest) and three regression algorithms (Linear Regression, Random Forest Regressor and Support Vector Regression) using the same preprocessing and evaluation procedures.

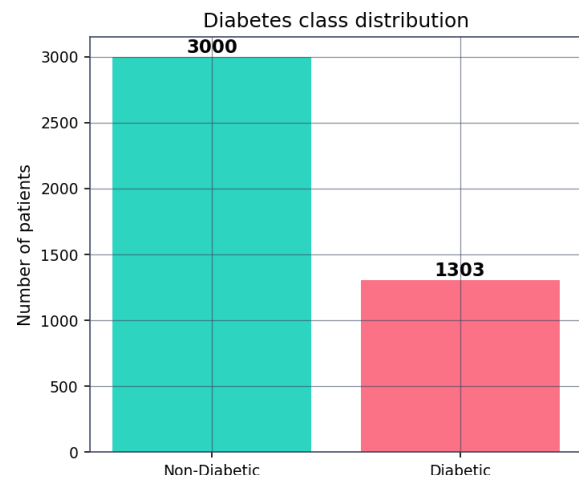


Fig. 1: Distribution of diabetic and non-diabetic patients.

## III. MATERIALS AND METHODS

### A. Dataset Description

This study used a standardized healthcare dataset, comprising of 4,303 patient records. The data set contains demographic data, anthropometric data, blood pressure, biochemical laboratory results, lifestyle factors and family history. Of the records, 3,000 patients were labelled as non-diabetic and 1,303 as diabetic.

The classification task predicts if the patient is diabetic, and the regression task predicts the Follow-up Fasting Plasma Glucose (FFPG) level.

### B. Data Preprocessing

Before the model training, missing values were filled using appropriate statistical measures, duplicate records were removed and categorical variables were encoded. To make sure that the features contribute equally during model training, numerical features were standardized using StandardScaler.

The dataset was imbalanced and the Synthetic Minority Oversampling Technique (SMOTE) was used to generate synthetic minority-class samples in the training data. Finally, the dataset was split into training and testing subsets, using an 80:20 stratified split.

### C. Machine Learning Models

Three supervised learning algorithms are evaluated for diabetes classification.

1) *Logistic Regression*: Logistic Regression predicts the probability of diabetes by the sigmoid function

$$P(y) = \frac{1}{1 + e^{-z}} \quad (1)$$

where  $z$  is the weighted sum of the input features.

2) *Support Vector Machine*: Support Vector Machine constructs an optimal hyperplane that maximizes the separation between diabetic and non diabetic patients. It also gives an estimate of the importance of each feature, which helps to interpret the relevant clinical variables.

3) *Random Forest*: Random Forest is an ensemble learning technique which uses multiple decision tree models and combines their output by majority voting. It also provides feature importance estimates, which allows the interpretation of influential clinical variables.

#### D. Regression Models

Three regression algorithms such as Linear Regression, Random Forest Regressor and Support Vector Regression (SVR) were tested for estimating Follow-up Fasting Plasma Glucose (FFPG).

#### E. Performance Evaluation

Classification performance was evaluated in terms of accuracy, precision, recall, F1-score and receiver operating characteristic area under the curve (ROC-AUC).

Accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP, TN, FP, and FN are true positive, true negative, false positive and false negative respectively.

The regression performance was evaluated using the coefficient of determination ( $R^2$ ), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

### IV. RESULTS AND DISCUSSION

The proposed framework was tested on both classification

and regression tasks. We trained classification models to predict diabetes status and regression models to predict Follow-up Fasting Plasma Glucose (FFPG). In all the experiments, the same training and testing strategy was used so that the evaluated algorithms could be fairly compared.

#### A. Exploratory Data Analysis

Before training the model, the Exploratory Data Analysis (EDA) was performed to understand the characteristics of the dataset and the relationships among the clinical variables. The class distribution of diabetic and non-diabetic patients is shown in Figure 1. The dataset is moderately imbalanced but this was addressed during pre-processing using SMOTE.

The numerical features are visualized in the correlation matrix in Figure 2. The Fasting Plasma Glucose (FPG) was the most positively correlated to the diagnosis of diabetes, followed by the age of patients and the Follow-up Fasting Plasma Glucose (FFPG). Blood pressure variables also demonstrated moderate positive correlation, and HDL and LDL cholesterol demonstrated the expected relationships in lipid profile. These results suggest that glucose-related measures are the most predictive measures for assessing the risk of diabetes.

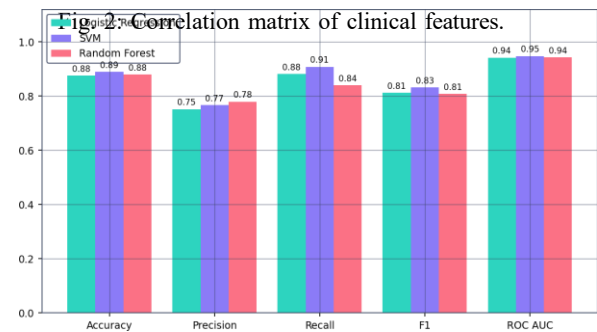
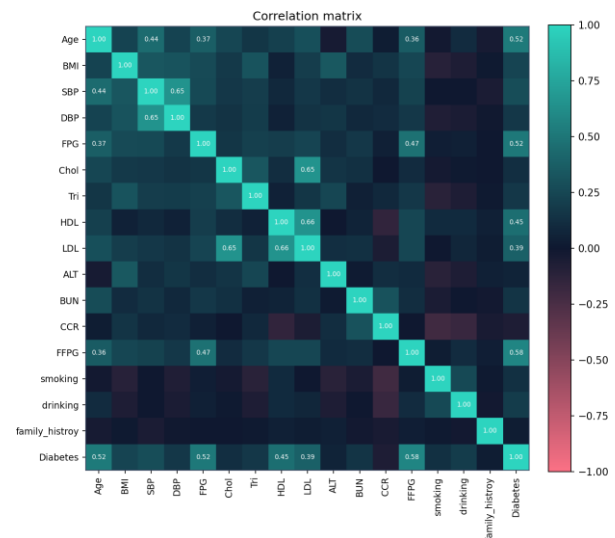


Fig. 3: Comparison of classification performance.

TABLE I: Classification Performance

| Model                  | Acc.  | Prec. | Recall | F1    | ROC-AUC |
|------------------------|-------|-------|--------|-------|---------|
| Logistic Regression    | 0.937 | 0.868 | 0.935  | 0.900 | 0.976   |
| Support Vector Machine | 0.952 | 0.937 | 0.904  | 0.920 | 0.980   |
| Random Forest          | 0.952 | 0.955 | 0.885  | 0.919 | 0.983   |

#### B. Classification Results

Three supervised machine learning algorithms were employed for the classification of diabetes, namely Logistic Regression, Support Vector Machine (SVM) and Random Forest.

Figure 3 shows the comparison of overall performance of the three classifiers. Table I summarizes the evaluation metrics. All the three classifiers yielded excellent predictive performance with classification accuracies above 93%. We have shown that Logistic Regression is a good baseline model with balanced metrics for all the evaluation metrics. The Support Vector Machine performed well in classification accuracy and good in the trade-off between precision and recall. Random Forest had the highest ROC-AUC (0.983) and precision (0.955), indicating the best discriminative ability for

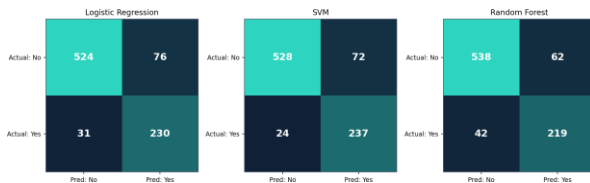


Fig. 4: Confusion matrices of the evaluated classifiers.

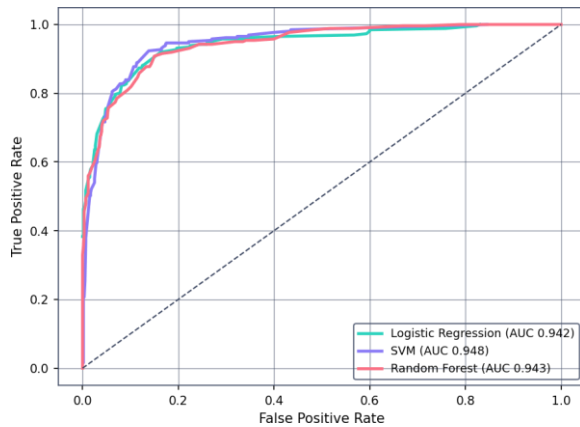


Fig. 5: ROC curves of the evaluated classifiers.

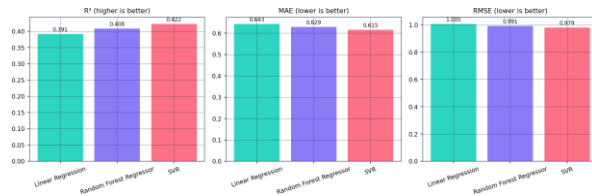


Fig. 6: Comparison of regression model performance.

classification of diabetic and non-diabetic patients.

The confusion matrices of the evaluated classifiers are shown in Figure 4. All the models classified most of the patient records correctly. Random Forest generated fewer false positive predictions compared to the other classifiers.

Figure 5 illustrates the Receiver Operating Characteristic (ROC) curves, which further confirm the efficiency of the proposed classifiers. All models showed excellent classification ability with ROC-AUC values higher than 0.97 for various decision thresholds.

### C. Regression Results

The Follow-up Fasting Plasma Glucose (FFPG) was estimated using Linear Regression, Random Forest Regressor and Support Vector Regression (SVR).

The regression performance comparison is presented in Figure 6 and a summary of the evaluation metrics is provided in Table II.

Support Vector Regression showed the highest coefficient of determination ( $R^2 = 0.422$ ) and the lowest MAE and RMSE. This implies that among the regression models considered,

TABLE II: Regression Performance

| Model                     | $R^2$ | MAE   | RMSE  |
|---------------------------|-------|-------|-------|
| Linear Regression         | 0.391 | 0.643 | 1.005 |
| Random Forest Regressor   | 0.408 | 0.629 | 0.991 |
| Support Vector Regression | 0.422 | 0.615 | 0.979 |

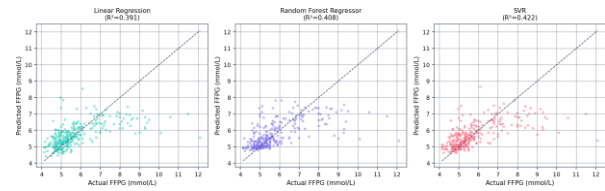


Fig. 7: Actual versus predicted Follow-up Fasting Plasma Glucose (FFPG).

SVR gives the best prediction for Follow-up Fasting Plasma Glucose.

Figure 7 shows the actual vs. predicted FFPG values. SVR has better prediction accuracy. The predictions produced by SVR are closer to the ideal regression line in comparison to Linear Regression and Random Forest Regression.

To summarize, the experimental results demonstrate the reliability of machine learning methods to predict Type 2 Diabetes from routinely collected clinical data. Random Forest achieved the highest classification performance by ROC-AUC and Support Vector Regression the most accurate follow-up glucose predictions. The results suggest that conventional models using machine learning can assist clinicians in early screening for diabetes and also identify individuals who could benefit from preventive intervention.

### V. CONCLUSION

In this work, we proposed a machine learning framework for early prediction of T2DM using demographic, clinical, lifestyle and biochemical data. Three supervised classification algorithms were evaluated namely Logistic Regression, Support Vector Machine (SVM) and Random Forest for predicting the risk of diabetes. Linear Regression, Random Forest Regressor and Support Vector Regression (SVR) were used to predict the Follow-up Fasting Plasma Glucose (FFPG). Prior to developing the model, the dataset was preprocessed by missing value imputation, duplicate elimination, feature standardization and Synthetic Minority Oversampling Technique (SMOTE) to improve the quality of the data and balance the classes.

The experimental results showed that all the classification models have good prediction performance. ROC-AUC and precision were the best for Random Forest, which means it was better at discriminating diabetic and non-diabetic patients, while Support Vector Machine had competitive classification accuracy. Support Vector Regression has the highest  $R^2$  score and the lowest prediction errors for the regression task, so it is the most suitable model to estimate fasting plasma glucose levels in the future. Exploratory Data Analysis also showed

that the most influential clinical variables associated with the prediction of diabetes were the fasting plasma glucose and the age of the patient.

To summarize, the proposed approach highlights the potential of machine learning methods for aiding in early diabetes screening and clinical decision making based on routinely gathered healthcare data. In future work, we will investigate more ensemble and deep learning models, incorporate Explainable Artificial Intelligence (XAI) methods such as SHAP and LIME, and embed the presented framework into web-based and mobile healthcare applications for real-time diabetes risk assessment.

## REFERENCES

- [1] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed., Brussels, Belgium, 2021.
- [2] I. Kavakiotis *et al.*, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.
- [3] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.
- [4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [5] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [7] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, 2013.
- [8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, 2016, pp. 785–794.
- [9] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [10] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proc. Python in Science Conference*, 2010, pp. 56–61.
- [11] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
- [12] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.
- [13] A. Alghamdi *et al.*, "Predicting diabetes mellitus using SMOTE and machine learning approaches," *Healthcare*, vol. 10, no. 2, p. 228, 2022. N. G. Nguyen *et al.*, "A comparative study of machine learning algorithms for diabetes prediction," *IEEE Access*, vol. 10, pp. 58020–58031, 2022.
- [14] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017.
- [15] P. Netayawijit and W. Chansanam, "Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support," *Healthcare*, vol. 13, no. 20, Art. no. 2588, 2025.
- [16] "A Novel Machine Learning Approach for Diagnosing Diabetes with a Self-Explainable Interface," *Healthcare Analytics*, vol. 5, Art. no. 100301, 2024.
- [17] P. N. Srinivasu, S. Ahmed, M. Hassaballah, and N. Almusallam, "An Explainable Artificial Intelligence Software System for Predicting Diabetes," *Heliyon*, vol. 10, no. 16, Art. no. e36112, 2024.
- [18] "Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights," *Information*, vol. 16, no. 1, Art. no. 7, 2025.
- [19] F. Prendin *et al.*, "The Importance of Interpreting Machine Learning Models for Blood Glucose Prediction in Diabetes: An Analysis Using SHAP," *Scientific Reports*, vol. 13, Art. no. 16865, 2023.
- [20] "A Comparative Study of Explainable Machine Learning Models with Shapley Values for Diabetes Prediction," *Healthcare Analytics*, vol. 7, Art. no. 100390, 2025.
- [21] "Explainable Machine Learning for Diabetes Classification and Web-Based Implementation," *Intelligent Based Medicine*, Art. no. 100400, 2026.
- [22] W. Iftikhar, M. Yaseen, G. Rahman, M. A. Nauman, and U. F. Khattak, "Improving Diabetes Prediction Accuracy and Interpretability with SMOTE and SHAP," *Engineering, Technology & Applied Science Research*, vol. 16, no. 2, pp. 34276–34282, 2026.
- [23] M. Goyal and D. Singh, "Performance-Interpretability Trade-off in Machine Learning Models for Diabetes Prediction: A Comparative Study Using Explainable AI," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2026.
- [24] "Explainable Reinforcement Learning for Glucose Monitoring Based on Shapley Value Analysis," *Computer Methods and Programs in Biomedicine*, Art. no. 109266, 2026.
- [25] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [26] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
- [27] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
- [28] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [29] A. Alghamdi *et al.*, "Predicting Diabetes Mellitus Using SMOTE and Machine Learning Approaches," *Healthcare*, vol. 10, no. 2, Art. no. 228, 2022.