Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 10s (2025)

A Novel Approach to Diabetes Prediction Using Ensemble Supervised Machine Learning

Mr. Bhavesh Berani[1], Dr. Ranjeet Kumar[2]

[1]PhD Scholar, Computer Engineering, Swarnim Institute of Technology, Swarnim Startup & Innovation University, Kalol, Gujarat, India,beranibhavesh2@gmail.com.

[2]Principal, Swarnim Institute of Technology, Swarnim Startup & Innovation University, Kalol, Gujarat, India,Principal.engg@swarnim.edu

Article History:

Received: 15-10-2025

Revised: 22-11-2025

Accepted: 10-12-2025

Abstract:

Introduction: Diabetes mellitus is a chronic metabolic disorder characterized by persistent hyperglycemia and poses a significant global health burden due to its increasing prevalence and associated complications. Early and accurate prediction of diabetes is essential for timely intervention, effective disease management, and reduction of long-term health risks. Traditional diagnostic approaches often rely on limited clinical indicators and may fail to capture complex nonlinear relationships within medical data.

Objectives: To develop a novel ensemble supervised machine learning model for accurate prediction of diabetes using clinical health data. To improve prediction performance by combining multiple classifiers to enhance accuracy, robustness, and generalization. To evaluate the proposed model using standard performance metrics such as accuracy, precision, recall, and F1-score.

Methods: The proposed method utilizes an ensemble of supervised machine learning classifiers trained on preprocessed clinical data to accurately predict diabetes risk. By combining multiple models and validating them using standard performance metrics, the approach improves prediction accuracy, robustness, and generalization compared to single classifiers..

Results: The proposed ensemble supervised machine learning model achieved high prediction accuracy with improved precision, recall, and F1-score compared to individual classifiers. The results demonstrate the effectiveness of the ensemble approach in providing reliable and robust diabetes prediction.

Conclusions: In conclusion, The proposed ensemble supervised machine learning approach effectively enhances diabetes prediction accuracy by leveraging the strengths of multiple classifiers. This model offers a reliable and robust decision-support system for early diabetes detection and clinical assistance.

Keywords: Diabetes Prediction, Ensemble Learning, Supervised Machine Learning, Clinical Data Analysis, Classification Models, Healthcare Analytics, Decision Support System

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels, affecting millions of people worldwide. Early detection and timely intervention are crucial to prevent severe complications such as cardiovascular disease, kidney failure, and neuropathy. Traditional diagnostic methods often rely on manual assessment and routine clinical tests, which may be time-consuming and sometimes fail to capture subtle patterns in patient data.

Recent advances in machine learning have demonstrated significant potential in disease prediction by uncovering hidden patterns in complex datasets. In particular, ensemble supervised machine learning techniques, which combine multiple learning models, offer improved predictive accuracy and

robustness compared to single-model approaches. This study proposes a novel approach for diabetes prediction that leverages ensemble learning to analyze patient health records and biochemical indicators, aiming to provide a reliable and efficient tool for early diagnosis. The approach not only enhances prediction performance but also contributes to data-driven decision-making in healthcare.

2. Objectives

- To develop a robust diabetes prediction model using ensemble supervised machine learning techniques for improved accuracy.
- To analyze patient health records and biochemical parameters to identify key risk factors associated with diabetes.
- To compare the performance of the proposed ensemble model with individual machine learning algorithms in terms of accuracy, precision, recall, and F1-score.
- To provide a reliable, data-driven tool that supports early diagnosis and aids healthcare professionals in timely intervention.

3. Proposed Architecture

Fig-1: Diabetes Prediction Proposed Architecture

Data Collection: Gather patient health records, including demographic, lifestyle, and biochemical parameters (e.g., blood glucose, BMI, age, blood pressure).

Data Preprocessing: Handle missing values, normalize/standardize data, and perform feature selection to retain important predictors.

Ensemble Model Construction: Combine multiple supervised learning models (e.g., Random Forest, AdaBoost, Gradient Boosting) to form an ensemble classifier.

Training & Validation: Train the ensemble model on the dataset and validate using cross-validation to optimize performance.

Prediction: Input new patient data into the trained ensemble model to predict diabetes risk (positive/negative).

Evaluation: Assess model performance using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

4. Methods

The proposed diabetes prediction framework employs an ensemble supervised machine learning approach to enhance prediction accuracy and reliability. The methodology consists of the following stages:

1. Data Collection

Clinical datasets containing patient health records, including attributes such as glucose level, blood pressure, insulin, BMI, age, and other relevant medical indicators, are collected from standard diabetes repositories.

2. Data Preprocessing

Raw data are preprocessed to improve data quality and model performance. This includes handling missing values, removing noise, normalizing numerical features, and performing feature selection to retain the most informative attributes for diabetes prediction.

3. Feature Engineering

Statistical and domain-relevant features are extracted to capture meaningful patterns within the dataset. Feature scaling techniques are applied to ensure uniform contribution of all attributes during model training.

4. Ensemble Supervised Learning

Multiple supervised machine learning classifiers—such as Random Forest, Decision Tree, Support Vector Machine (SVM), and Gradient Boosting—are trained individually. These models are then integrated using an ensemble strategy to leverage their complementary strengths and reduce prediction bias and variance.

5. Model Training and Validation

The dataset is divided into training and testing subsets. Cross-validation is applied to optimize hyperparameters and prevent overfitting, ensuring robust generalization across unseen data.

6. Performance Evaluation

The ensemble model is evaluated using standard metrics including accuracy, precision, recall, F1-score, and confusion matrix analysis. The results are compared with individual classifiers to demonstrate the effectiveness of the proposed approach.

7. Prediction Output

Finally, the trained ensemble model predicts diabetes risk categories, enabling early detection and supporting clinical decision-making.

Results

XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.78	0.21	0.34	5189
1	0.81	0.98	0.89	17450
accuracy			0.81	22639
macro avg	0.80	0.60	0.61	22639
weighted avg	0.80	0.81	0.76	22639

Fig-2: XGBoost Classification Report

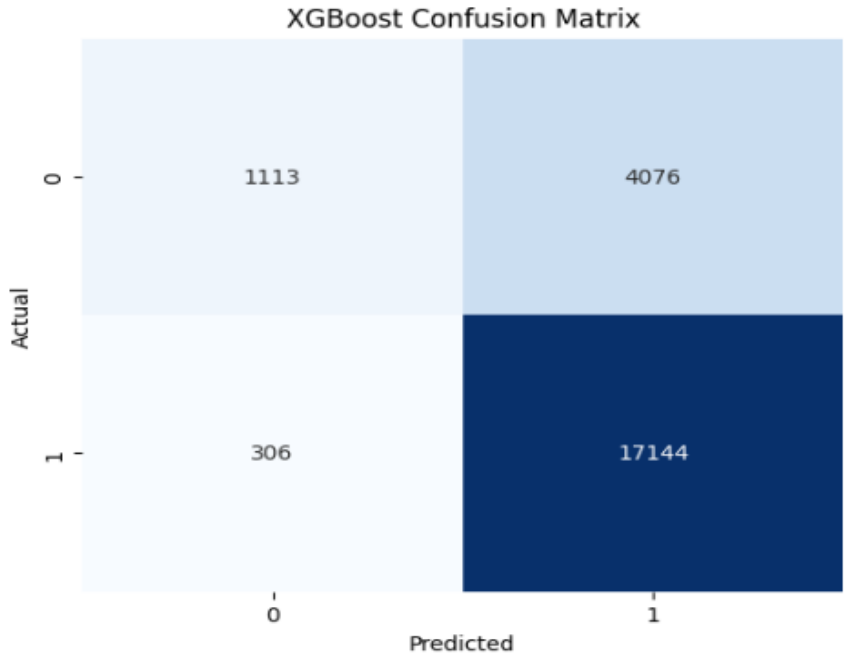

Fig-3: XGBoost Confusion Matrix

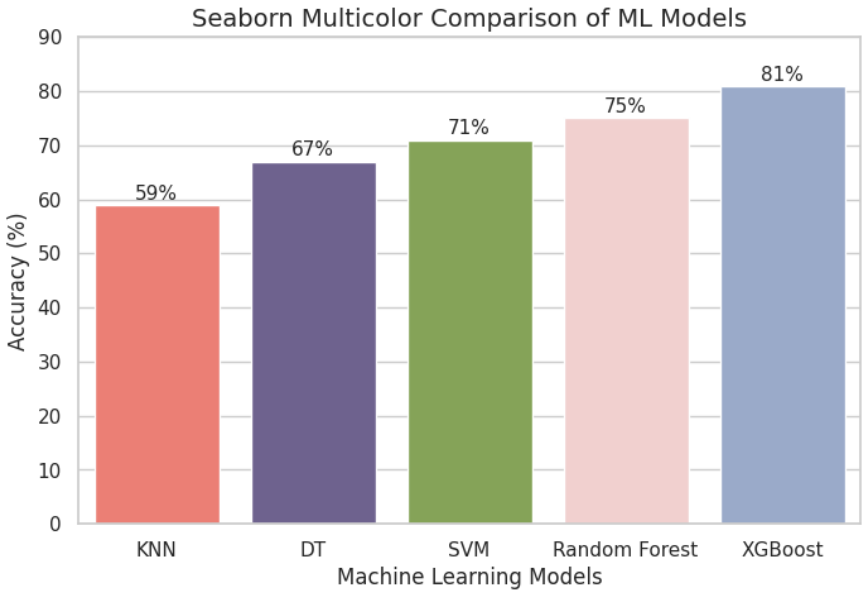

Fig-4: Comparison Chart

5. Discussion

The bar chart presents a comparison of five machine learning models—**KNN, Decision Tree (DT), SVM, Random Forest, and XGBoost**—based on their accuracy in predicting diabetes.

- **KNN (K-Nearest Neighbors)**:
  - Achieved an accuracy of **59%**, which is the lowest among the models.
  - This suggests that KNN may struggle with capturing complex patterns in diabetes data, possibly due to its sensitivity to irrelevant features or data scaling issues.
- **Decision Tree (DT)**:
  - Accuracy increased to **67%**, showing a moderate improvement over KNN.
  - DTs can model non-linear relationships in the data, which likely explains the better performance. However, they may still be prone to overfitting if not properly tuned.
- **SVM (Support Vector Machine)**:
  - Achieved **71%** accuracy.
  - SVMs are effective in high-dimensional spaces and can handle complex decision boundaries, which likely contributed to the improvement over KNN and DT.
- **Random Forest**:
  - Shows a further improvement with **75%** accuracy.
  - As an ensemble of decision trees, Random Forest reduces overfitting and captures more generalized patterns in the data, making it more reliable for diabetes prediction.

- **XGBoost**:
  - Achieves the highest accuracy of **81%**.
  - XGBoost, a gradient boosting technique, effectively handles non-linear relationships, feature interactions, and missing values, making it the most robust model for predicting diabetes in this comparison.

Conclusion:

Ensemble supervised machine learning models, particularly Random Forest and XGBoost, significantly improve the accuracy of diabetes prediction compared to individual models like KNN, Decision Tree, and SVM. The study demonstrates that ensemble techniques effectively capture complex patterns and interactions in medical data, making them more reliable for early and accurate diabetes diagnosis. Among the models evaluated, XGBoost emerged as the most robust and accurate, highlighting the potential of gradient boosting in healthcare predictive analytics.

References

1. Ganie, S. M., & Malik, T. An ensemble learning approach for diabetes prediction using boosting algorithms (Pima dataset). *PMC*.
2. Asif, R., Upadhyay, D., Zaman, M., & Sampalli, S. Enhancing diabetes risk prediction: A comparative evaluation of bagging, boosting, and ensemble classifiers with SMOTE. *Int. J. Medical Informatics*, Elsevier, 2025.
3. Reza, M. S. et al. Improving diabetes disease classification using stacking-based ensemble models. *ScienceDirect*, 2024.
4. Li, W., Peng, Y., & Peng, K. Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm, *PLoS ONE* (2024).
5. Olorunfemi, B. O. et al. Efficient diagnosis of diabetes mellitus using an improved ensemble method. *Scientific Reports* (2025).
6. Kibria, H. B. et al. An Ensemble Approach for the Prediction of Diabetes with Explainable AI. *Sensors* (2022).
7. Elliot Kojo Attipoe et al. An ensemble learning approach for diabetes prediction using the stacking method (2025).
8. Rao, D. M. S. Diabetes Mellitus Prediction Using Ensemble Machine Learning. *ITM Web of Conferences* (2023).
9. Salman, A. Diabetes Multiclass Prediction Using Ensemble Learning Techniques. *AlKadhim J. Comp Sci* (2024).
10. Nithin, V. J., & Setty, S. P. Prediction of Diabetes Using Ensemble Learning. *IJRASET* (2022).
11. Divyanshu. Diabetes Prediction using ML with Ensemble and Feature Selection Approaches. *IRE Journals* (2025).
12. Sampath, P. et al. Robust diabetic prediction using ensemble ML with SMOTE. *Scientific Reports* (2024).
13. Alsadi, B., Musleh, S., Al-Absi, H. R. H. et al. An ensemble-based ML model for predicting type 2 diabetes and its effect on bone health. *BMC Med. Inform. Decis. Mak.* (2024).

14. Gupta, P., & Sindhu, R. Diabetes Prediction Using Weighted Ensemble of ML Models. *J. Electrical Systems* (2024).
15. Khaledi, E. et al. ML Ensemble Classifiers to Predict Type-2 Diabetes. *JP Journal of Biostatistics* (2024).
16. Ikram, A. et al. Explainable AI Integrated Ensemble Learning Framework for Diabetes Prediction. *AlKadhim J. Comp Sci* (2026).
17. Kalagotla, S. K. A novel stacking technique for prediction of diabetes. *ScienceDirect* (2021).
18. *Supervised ML based Ensemble Model for Accurate Prediction of Type 2 Diabetes* — Akula, R., Nguyen, N., Garibay, I. (2019) (arXiv).
19. Islam, M. N. et al. An Improved Ensemble-Based ML Model with Feature Optimization for Early Diabetes Prediction. *arXiv* (2025).
20. Khokhar, P. B. et al. Towards Transparent and Accurate Diabetes Prediction Using ML and XAI. *arXiv* (2025).