International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Leveraging machine learning for early detection of diabetes: A dataset-driven comparative study

Taaha Ansari [1,*] and Vaishali M. Bagade [2]

[1] *PG Scholar, Department of Electronics and Tele Communication Engineering, Alamuri Ratnamala Institute of Engineering and Technology, Shahapur, India.*
[2] *Assistant Professor, Department of Electronics and Tele Communication Engineering, Alamuri Ratnamala Institute of Engineering and Technology, Shahapur, India.*

International Journal of Science and Research Archive, 2025, 16(03), 417–422

Publication history: Received on 13 July 2025; revised on 04 September 2025; accepted on 06 September 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.3.2538

Abstract

Diabetes mellitus is one of the most prevalent chronic diseases worldwide. This study investigates the application of machine learning algorithms for early diabetes prediction, a crucial step in preventing severe health complications. Two benchmark datasets, the Pima Indians Diabetes dataset and the Scikit-learn Diabetes dataset (converted to binary classification), were analyzed using Logistic Regression, Support Vector Machine (SVM with RBF kernel), and XGBoost. The methodology included preprocessing, train-test splitting, feature scaling, and model evaluation with metrics such as Accuracy, ROC-AUC, Precision, Recall, F1-score, and Confusion Matrices. Logistic Regression, serving as the baseline, performed best on the Pima dataset with 75.3% accuracy and a ROC-AUC of 0.815, demonstrating its strength in structured medical data. On the Scikit-learn dataset, XGBoost achieved the highest accuracy (76.4%), while SVM produced the best ROC-AUC (0.841), showcasing its ability to capture complex non-linear patterns. Findings highlight the importance of dataset-specific model selection and the integration of linear and non-linear approaches for reliable healthcare decision-support systems.

Keywords: Diabetes Prediction; Machine Learning; Logistic Regression; Support Vector Machine; Xgboost; Medical Diagnosis

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels (hyperglycemia), which can lead to severe complications affecting the heart, kidneys, eyes, blood vessels, and nerves. According to the World Health Organization (2022), diabetes prevalence has quadrupled since 1980, with over 422 million adults living with the disease globally. The International Diabetes Federation (IDF) further estimated that 589 million adults were diabetic in 2023, a number projected to rise to 783 million by 2045 without effective intervention. In India alone, over 77 million adults are affected, making it one of the countries with the largest diabetic populations. Type 2 diabetes, accounting for 90–95% of cases, is closely linked to lifestyle factors such as obesity, poor diet, and physical inactivity. Since traditional diagnostic tools like Fasting Plasma Glucose (FPG) and Oral Glucose Tolerance Tests (OGTT) are invasive and not feasible for mass screening, early and non-invasive prediction methods are essential [1].

Machine Learning (ML) techniques offer a promising solution to this challenge by analyzing historical health datasets and identifying hidden patterns in medical and demographic data. ML-based predictive systems can act as clinical decision-support tools, providing early diagnosis, personalized treatment recommendations, and large-scale screening capabilities. This research applies three ML classifiers Logistic Regression (LR), Support Vector Machine (SVM), and

* Corresponding author: Taaha Ansari; Email: taahaswork@gmail.com

Extreme Gradient Boosting (XGBoost) to benchmark datasets including the Pima Indians Diabetes Dataset and the Scikit-learn diabetes dataset, aiming to determine the most effective predictive model.

The motivation for adopting ML arises from the inadequacy of traditional methods to provide scalable, accurate, and personalized diagnostic solutions. ML algorithms can incorporate diverse patient data such as age, BMI, blood pressure, glucose levels, insulin concentration, and family history, delivering predictive insights that assist healthcare professionals and empower patients for proactive health management.

Diabetes manifests in three main forms: Type 1 diabetes, an autoimmune condition causing near-total insulin deficiency and affecting about 5–10% of cases; Type 2 diabetes, the most common type, developing gradually due to insulin resistance and lifestyle factors; and Gestational diabetes, a temporary condition in pregnancy that poses long-term risks of Type 2 diabetes for mothers. Symptoms often go unnoticed in the early stages and may include frequent urination, excessive thirst, unexplained weight loss, fatigue, blurred vision, slow wound healing, and recurrent infections. The subtle onset of these symptoms makes predictive models critical for timely detection and intervention.

The etiology of diabetes is multifactorial, with three major contributing factors. Genetic predisposition plays a significant role, with certain gene mutations linked to higher susceptibility, particularly in Type 1 diabetes. Lifestyle factors, including sedentary behavior, obesity, poor diet, and associated comorbidities like hypertension and polycystic ovarian syndrome (PCOS), are strongly correlated with Type 2 diabetes. Autoimmunity and environmental triggers such as viral infections can initiate immune responses that destroy insulin-producing pancreatic beta cells [2].

Developing an ML-based diabetes prediction system involves a structured workflow. First, data collection includes demographic, physiological, and lifestyle variables. Data preprocessing ensures consistency by handling missing values, balancing classes, and scaling features. Feature selection identifies key predictors like glucose and BMI to reduce dimensionality without losing predictive power. Model training applies algorithms such as LR, SVM, and XGBoost, while model validation evaluates their performance using metrics like Accuracy, Precision, Recall, F1-score, and ROC-AUC. Finally, the deployment phase integrates the best-performing model into a web-based application or decision-support system for real-time predictions.

The framework ensures scalability, interpretability, and reliability, bridging data science with clinical practice. Once deployed via platforms such as Flask, patients or clinicians can input real-time data to receive instant predictions, making healthcare more proactive and accessible. This integrated pipeline highlights how machine learning can transform traditional healthcare by providing robust, user-friendly solutions for large-scale diabetes prediction and early intervention.

2. Literature review

Diabetes prediction has attracted significant research interest due to the rising global prevalence of the disease. Early works largely relied on statistical methods such as Logistic Regression (LR), which offered interpretability but limited performance on complex datasets [Hosmer and Lemeshow, 2000]. LR has been widely used as a baseline model, achieving accuracies of around 75% on benchmark datasets such as the Pima Indians Diabetes Dataset (PIDD) [3].

With the advancement of machine learning, Support Vector Machines (SVMs) demonstrated superior capability in handling non-linear data. Studies have shown SVM with an RBF kernel achieving accuracy rates above 80%, highlighting its robustness against false negatives, a critical requirement in medical diagnosis [4].

Recent works emphasize ensemble models such as XGBoost, Random Forest, and Gradient Boosting Machines, which often outperform standalone algorithms. Hasan et al. (2020) reported that ensemble classifiers achieved AUC scores as high as 0.95 on the Pima dataset. Furthermore, hybrid and deep learning models (e.g., CNN-LSTM frameworks) have pushed performance even further, though at the expense of interpretability and computational cost [5].

Despite promising results, major research gaps persist

- Over-reliance on small benchmark datasets (PIDD, Sklearn).
- Limited comparative studies across datasets.
- The interpretability vs. accuracy trade-off in advanced models.
- Generalizability to real-world, heterogeneous hospital datasets.

This study addresses these gaps by systematically comparing LR, SVM, and XGBoost across two benchmark datasets, emphasizing model performance and practical deployment feasibility.

2.1. Problem statement

Diabetes is a chronic condition that can lead to severe health complications if not detected and managed early, making accurate prediction systems essential in modern healthcare. With the increasing availability of patient health data, machine learning (ML) offers significant potential for developing predictive models; however, challenges such as missing or imbalanced data, selecting the most relevant features, and ensuring interpretability remain critical. This study seeks to address these issues by applying and comparing three widely used ML algorithms—Logistic Regression, Support Vector Machine (SVM), and XGBoost—on benchmark diabetes datasets. The objective is to evaluate their performance using standard metrics, analyze their strengths and limitations, and identify the most effective model that achieves a balance between accuracy, interpretability, and computational efficiency for practical deployment in healthcare systems.

2.2. System architecture

Figure 1 illustrates the complete architecture of the proposed Diabetes Prediction System, which follows a structured and sequential pipeline designed to transform raw medical data into reliable predictions using machine learning (ML). The architecture integrates all essential stages of ML application data handling, model development, evaluation, and deployment ensuring that the system is not only technically sound but also clinically meaningful and user-friendly.

Figure 1 presents the architecture of the Diabetes Prediction System, which outlines the step-by-step pipeline for predicting diabetes using machine learning. The process starts with data collection, where demographic and clinical information such as age, glucose levels, BMI, insulin, and blood pressure are gathered. Since raw datasets often contain missing or inconsistent values, a preprocessing stage is carried out to clean the data, replace invalid entries, balance class distributions (using methods like SMOTE), and scale features for uniformity.

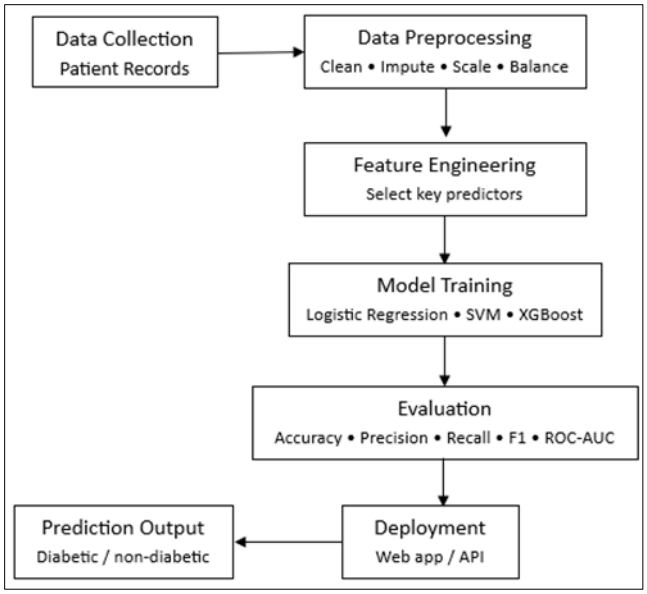

Figure 1 Diabetes Prediction System Architecture

The next stage is feature engineering, where statistically significant predictors such as glucose and BMI are identified. This reduces noise and ensures that only relevant attributes are used for training. Once features are prepared, three machine learning models Logistic Regression, Support Vector Machine (SVM), and XGBoost—are applied. Logistic Regression serves as an interpretable baseline, SVM effectively handles non-linear boundaries, and XGBoost captures complex interactions in the dataset.

These models are then evaluated using performance metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC. Comparative analysis ensures that the most effective model is chosen for deployment. Finally, the best model is integrated into a web application, allowing users to input medical data and receive real-time predictions (Diabetic/Non-Diabetic) with probability scores.

In summary, Figure 1 represents an end-to-end machine learning pipeline for diabetes prediction. It demonstrates how raw medical data progresses through preprocessing, feature engineering, and predictive modeling to finally deliver clinically useful outcomes. By integrating robust technical processes with user-oriented deployment, this architecture ensures that the system is both accurate in its predictions and practical for real-world healthcare decision-making.

3. Dataset selection and description

This study employs two benchmark datasets to train and evaluate the diabetes prediction system: the Pima Indians Diabetes Database (PIDD) and the Modified Scikit-learn Diabetes Dataset. Together, these datasets provide complementary strengths, ensuring robust analysis across different clinical conditions.

3.1. Pima Indians Diabetes Database (PIDD)

The primary dataset is sourced from the UCI Machine Learning Repository and contains medical records of 768 women of Pima Indian heritage, a group with a high prevalence of Type 2 diabetes. It includes 8 predictor variables such as pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. The target variable is binary (diabetic = 1, non-diabetic = 0). Challenges include imbalanced class distribution (~35% positive cases) and invalid zero entries in some features, making it a suitable test case for preprocessing and class balancing techniques.

3.2. Modified Scikit-learn Diabetes Dataset

The secondary dataset, originally designed for regression, was adapted into a binary classification task by splitting the target values at the median. Records above the median were labeled as high risk (1) and that below as low risk (0). This dataset contains 442 samples and 10 predictor variables, including age, sex, BMI, blood pressure, and six serum measurements. Unlike PIDD, it has no missing values and maintains a balanced class distribution after transformation.

In summary, while PIDD tests model robustness under imbalanced and noisy conditions, the Scikit-learn dataset evaluates performance on balanced, standardized data. Using both ensures broader reliability and generalizability of the proposed prediction system.

3.3. Proposed system

Figure 2 illustrates the decision-making flow of the proposed diabetes prediction model, which is based on machine learning algorithms. The flow begins with the input stage, where patient health-related attributes such as glucose, BMI, blood pressure, age, and insulin are provided to the system.

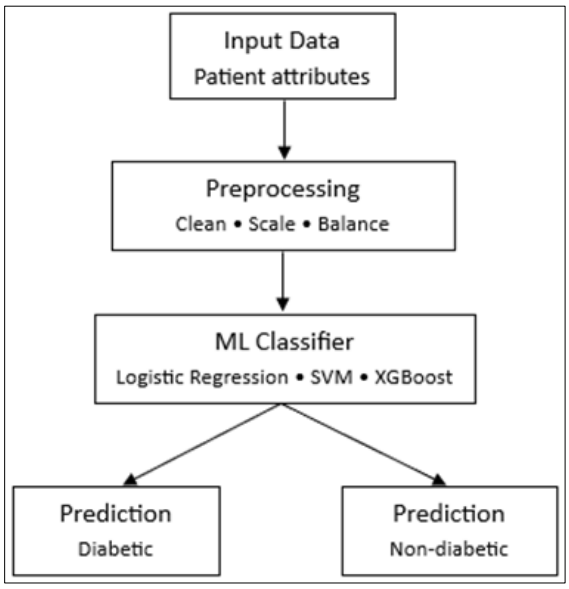

Figure 2 Diabetes Prediction System Architecture

The flow diagram highlights the step-by-step branching logic of classification: each node evaluates a feature-based condition (e.g., glucose level above a certain threshold), leading to further splits until the system arrives at a final classification outcome either Diabetic (Yes) or Non-Diabetic (No).

This hierarchical decision-making flow makes the prediction process transparent and interpretable, showing how different features influence the final result. It also emphasizes that the model can provide real-time, accurate predictions by systematically evaluating patient attributes through the decision pipeline.

Table 1 Comparative Analysis Of Models

Aspect	**Logistic Regression (LR)**	**Support Vector Machine (SVM, RBF Kernel)**	**XGBoost (Extreme Gradient Boosting)**
Algorithm Overview	Baseline linear model for binary classification. Estimates probability of diabetes occurrence using a logistic function. Highly interpretable.	Non-linear classifier that uses the RBF kernel to map data into higher dimensions, creating complex boundaries for classification.	Ensemble boosting algorithm that builds multiple decision trees sequentially to capture complex patterns and improve accuracy.
Model Configuration	• Solver: LBFGS • Max Iter: 1000 • Regularization: L2 • Class Weight: Balanced	• Kernel: RBF • Gamma: Auto (1/n_features) • Class Weight: Balanced	• Objective: Binary logistic • Eval Metric: Log-loss • Max Depth: 6 • Learning Rate: 0.3 • Estimators: 100 • Use Label Encoder: False
Advantages	• Simple and easy to interpret • Provides probability outputs • Computationally efficient • Well-accepted in medical research	• Effective in high-dimensional spaces • Robust against overfitting • Memory efficient (uses support vectors) • Versatile kernel functions	• Superior predictive accuracy • Handles missing values automatically • Built-in regularization to prevent overfitting • Provides feature importance • Fast and scalable (parallel computing)
Limitations	• Assumes linear relationships • Weak on non-linear patterns • Sensitive to feature scaling	• Computationally intensive • Requires feature scaling • Less interpretable • Sensitive to hyperparameter tuning	• Complex and computationally heavy • Many hyperparameters to tune • Less interpretable • Risk of overfitting if not validated

3.4. Comparative analysis

The performance of three machine learning models Logistic Regression (LR), Support Vector Machine (SVM), and XGBoost was evaluated on two datasets. The metrics considered were Accuracy, Precision, Recall, F1-score, and ROC-AUC, which reflect different aspects of prediction quality.

Table 2 Comparative results of proposed models on both datasets

Model	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC-AUC**
Logistic Regression (LR)	Pima Indians	0.7532 (75.3%)	0.81	0.80	0.81	0.81
	Sklearn (Modified)	0.7303 (73.0%)	0.77	0.73	0.75	0.84
Support Vector Machine (SVM)	Pima Indians	0.7338 (73.3%)	0.77	0.83	0.80	0.81
	Sklearn (Modified)	0.7416 (74.1%)	0.80	0.71	0.75	0.84
XGBoost	Pima Indians	0.7208 (72.0%)	0.82	0.73	0.77	0.78
	Sklearn (Modified)	0.7640 (76.4%)	0.80	0.76	0.78	0.81

4. Results and discussion

On the Pima Indians dataset, Logistic Regression achieved the best performance with accuracy of 75.3% and ROC-AUC of 0.815, demonstrating strong interpretability. On the Scikit-learn dataset, XGBoost attained the highest accuracy (76.4%), while SVM recorded the best ROC-AUC (0.841). The comparative analysis reveals that Logistic Regression is ideal for interpretable results, while XGBoost and SVM handle complex patterns better, though at the cost of interpretability. ROC curve analysis further confirmed dataset-dependent performance variations. This underlines the necessity of selecting models based on dataset characteristics in medical applications.

5. Conclusion and future scope

This research highlights the effectiveness of machine learning in predicting diabetes and emphasizes dataset-specific model selection. Logistic Regression remains a reliable choice for interpretable medical insights, while SVM and XGBoost excel in handling non-linearities and achieving higher predictive accuracy. Future work can explore deep learning approaches, integration with medical IoT devices, and adoption of explainable AI techniques to enhance interpretability. The deployment of a web application demonstrates the real-world applicability of this work in clinical decision-support systems.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Zhao, M., Yao, Z., Zhang, Y., et al. (2025). Predictive value of machine learning for the progression of gestational diabetes mellitus to type 2 diabetes: a systematic review and meta-analysis. BMC Medical Informatics and Decision Making.

[2] Iftikhar, K., Javaid, N., Ahmed, I., and Alrajeh, N. (2025). A Novel Explainable Deep Learning Framework for Accurate Diabetes Mellitus Prediction. Applied Sciences, 15(16), Article 9162.

[3] J. W. Smith et al., "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in Annu. Symp. Comput. Appl. Med. Care, 1998, pp. 261–265.

[4] N. Mohan and V. Jain, "Performance analysis of support vector machine in diabetes prediction," in Int. Conf. Electronics, Communication and Aerospace Technology, 2020.

[5] Ansari, G., Bhat, S., Ansari, M. D., and Shadab, A. (2025). Advanced Supervised Machine Learning Techniques for Accurate Prediction of Diabetes Mellitus Using Feature Selection. Frontiers in Medicine.

[6] Alzboon, M. S., Al-Batah, M., Alqaraleh, M., Abuashour, A., and Bader, A. F. (2025). A Comparative Study of Machine Learning Techniques for Early Prediction of Diabetes. arXiv.

[7] Khokhar, P. B., Pentangelo, V., Palomba, F., and Gravino, C. (2025). Towards Transparent and Accurate Diabetes Prediction Using Machine Learning and Explainable Artificial Intelligence. arXiv.

[8] Babar, K. H., Bothra, A. A., Desale, O. S., Doshi, P. M., Banu, E. A., and Vishwakarma, P. P. (2025). Type-2 Diabetes Detection using XGBoost with ADASYN Over SVM. Journal of Information Systems Engineering and Management, 10(49s).

[9] Petridis, P. D., Kristo, A. S., Sikalidis, A. K., and Kitsas, I. K. (2024). A Review on Trending Machine Learning Techniques for Type 2 Diabetes Mellitus Management. Informatics, 11(4), 70.