STM JOURNALS

International Journal of
Bioinformatics and Computational Biology

ISSN: 2583-9977
Volume 4, Issue 2, 2026
July—December
DOI (Journal): 10.37591/IJBCB

https://journals.stmjournals.com/ijbcb

Review IJBCB

Diabetes Risk Prediction from Survey Data Using Machine Learning Algorithms

Aayush Shukla[1,*], Alok Singh[1], Padma Mishra[2], Melissa Fernandes[2]

Abstract

Diabetes mellitus represents one of the most significant global health challenges, affecting millions worldwide and leading to severe complications if left undiagnosed or poorly managed. Early detection and risk assessment are crucial for preventing the progression of this chronic condition. This research presents a comprehensive machine learning approach for predicting diabetes risk using survey-based health parameters. The study implements and compares four prominent classification algorithms: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest to analyze patient health indicators including glucose levels, blood pressure, body mass index, insulin levels, skin thickness, age, pregnancies, and diabetes pedigree function. The dataset comprises health survey responses from individuals with varying diabetes risk profiles. Through systematic feature analysis and model optimization, the research demonstrates that ensemble methods, particularly Random Forest, achieve superior predictive accuracy of 94.2%, followed by SVM at 91.8%, Logistic Regression at 89.3%, and KNN at 87.6%. The implemented web-based prediction system provides real- time risk assessment, enabling healthcare professionals and individuals to make informed decisions about diabetes prevention and management. The results indicate that machine learning-driven survey analysis can serve as an effective screening tool for diabetes risk prediction, potentially reducing healthcare costs and improving patient outcomes through early intervention strategies.

Keywords: Diabetes prediction, machine learning, survey data analysis, logistic regression, k-nearest neighbors, support vector machine, random forest, healthcare analytics, risk assessment, preventive medicine

INTRODUCTION

Diabetes mellitus has emerged as a global epidemic, with the International Diabetes Federation estimating that approximately 537 million adults worldwide live with this condition as of 2021. This chronic metabolic disorder, characterized by elevated blood glucose levels, poses significant health risks including cardiovascular disease, kidney failure, blindness, and lower limb amputation. The economic burden of diabetes extends beyond individual healthcare costs, impacting national healthcare systems and productivity levels across developed and developing nations alike. The complexity of diabetes diagnosis traditionally relies on clinical laboratory tests such as fasting plasma glucose, oral glucose tolerance tests, and glycated hemoglobin (HbA1c) measurements. However, these diagnostic procedures require clinical settings, trained personnel, and can be costly for large-scale screening programs. Furthermore, many individuals remain

***Author for Correspondence**
Aayush Shukla
E-mail: aayushshukla0828@gmail.com

[1]Research Scholar, MCA, Thakur Institute of Management Studies, Career Development & Research (TIMSCDR) Mumbai, India
[2]Associate Professor, Thakur Institute of Management Studies & Career Development & Research MCA Department

Received Date: March 30, 2026
Accepted Date: April 10, 2026
Published Date: April 20, 2026

Citation: Aayush Shukla, Alok Singh, Padma Mishra, Melissa Fernandes. Diabetes Risk Prediction from Survey Data Using Machine Learning Algorithms. International Journal of Bioinformatics and Computational Biology. 2026; 4(2): 1–8p.

undiagnosed until complications arise, highlighting the critical need for accessible, cost-effective screening methods that can identify at-risk populations before clinical symptoms manifest. Machine learning has revolutionized healthcare analytics by enabling the development of predictive models that can process complex, multi-dimensional health data to identify patterns and relationships that may not be apparent through traditional statistical methods. The application of artificial intelligence in medical diagnosis has shown remarkable success in various domains, from medical imaging to genomic analysis. In the context of diabetes prediction, machine learning algorithms can analyze survey- based health parameters to assess individual risk profiles, providing a non-invasive, scalable approach to population health screening. This research addresses the growing need for intelligent diabetes risk assessment tools by developing and comparing multiple machine learning models trained on comprehensive health survey data.

LITERATURE SURVEY

The intersection of machine learning and diabetes prediction has attracted significant research attention over the past decade, with numerous studies exploring various algorithmic approaches and data sources. This literature review examines key contributions in the field, highlighting methodological approaches, dataset characteristics, and performance outcomes that inform the current research.

The Research work [1] demonstrated the effectiveness of ensemble learning methods for diabetes prediction, achieving 96.7% accuracy using a combination of Random Forest and Gradient Boosting algorithms. Their study utilized the Pima Indian Diabetes Dataset and emphasized the importance of feature selection in improving model performance. The research highlighted that demographic and physiological parameters, when properly preprocessed and analyzed, can provide reliable diabetes risk indicators.

A comprehensive [2] compared multiple machine learning algorithms including Naive Bayes, Decision Trees, and Neural Networks for diabetes prediction. Their analysis revealed that ensemble methods consistently outperformed individual classifiers, with Random Forest achieving the highest accuracy of 94.1%. The study also emphasized the significance of data preprocessing techniques, particularly handling missing values and feature scaling, in improving model reliability. The application of Support Vector Machines in diabetes prediction was extensively studied in [3], who achieved 92.4% accuracy using polynomial kernel functions. Their research demonstrated that SVM's ability to handle non-linear relationships between health parameters makes it particularly suitable for medical diagnosis applications. The study also explored the impact of different kernel functions on prediction accuracy, concluding that polynomial and radial basis function kernels provide superior performance compared to linear kernels.

Logistic Regression's effectiveness in diabetes prediction was investigated in [4], who developed a comprehensive model incorporating 21 health parameters. Their research achieved 89.7% accuracy and provided valuable insights into the statistical significance of individual features. The study emphasized Logistic Regression's interpretability advantage, allowing healthcare professionals to understand the contribution of each parameter to the final prediction. K- Nearest Neighbors algorithm was analyzed in paper [5] in the context of diabetes prediction, achieving 88.3% accuracy with optimal k-value selection. Their research explored the impact of distance metrics and neighborhood size on prediction performance, concluding that Euclidean distance with k=7 provides the best balance between bias and variance for diabetes prediction tasks [6].

PROPOSED SYSTEM

The proposed system presents a comprehensive machine learning framework for diabetes risk prediction utilizing survey-based health parameters to enable early detection and preventive healthcare intervention. Traditional diabetes diagnosis methods rely heavily on clinical laboratory tests and invasive procedures, which are often costly, time- consuming, and inaccessible to large populations, particularly in resource-constrained environments. This proposed solution addresses these limitations

by implementing a multi-algorithm approach that analyzes readily available health survey data to provide accurate risk assessment without requiring specialized medical equipment or clinical settings [7].

At the core of this system is a four-algorithm ensemble approach that leverages the complementary strengths of different machine learning paradigms. The framework incorporates Logistic Regression for its interpretability and statistical foundation, K-Nearest Neighbors for its ability to capture local data patterns, Support Vector Machine for handling non-linear relationships through kernel transformations, and Random Forest for its robust ensemble learning capabilities. Each algorithm processes eight key health parameters: number of pregnancies, glucose concentration levels, blood pressure measurements, skin fold thickness, insulin levels, body mass index, diabetes pedigree function, and age. These parameters were selected based on their established clinical significance and availability through standard health questionnaires. The system architecture employs a multi-stage processing pipeline that begins with comprehensive data preprocessing, including missing value imputation, outlier detection, and feature normalization to ensure optimal model performance. The prediction process utilizes trained models to analyze input parameters and generate both binary classification results and probability scores for risk assessment. The final output combines predictions from all four algorithms to provide users with comprehensive risk evaluation and confidence intervals. This multi-algorithm approach not only improves prediction accuracy but also provides robustness against individual model limitations and biases. The system is implemented as a web-based application using Python and Flask framework, enabling real-time risk assessment through an intuitive user interface. The modular architecture allows for easy integration with existing healthcare systems and supports both individual risk assessment and population-level screening applications. This comprehensive approach provides a cost-effective, scalable, and accessible solution for diabetes risk prediction, offering significant potential for improving preventive healthcare outcomes through early identification of at-risk individuals Figure 1 [8].

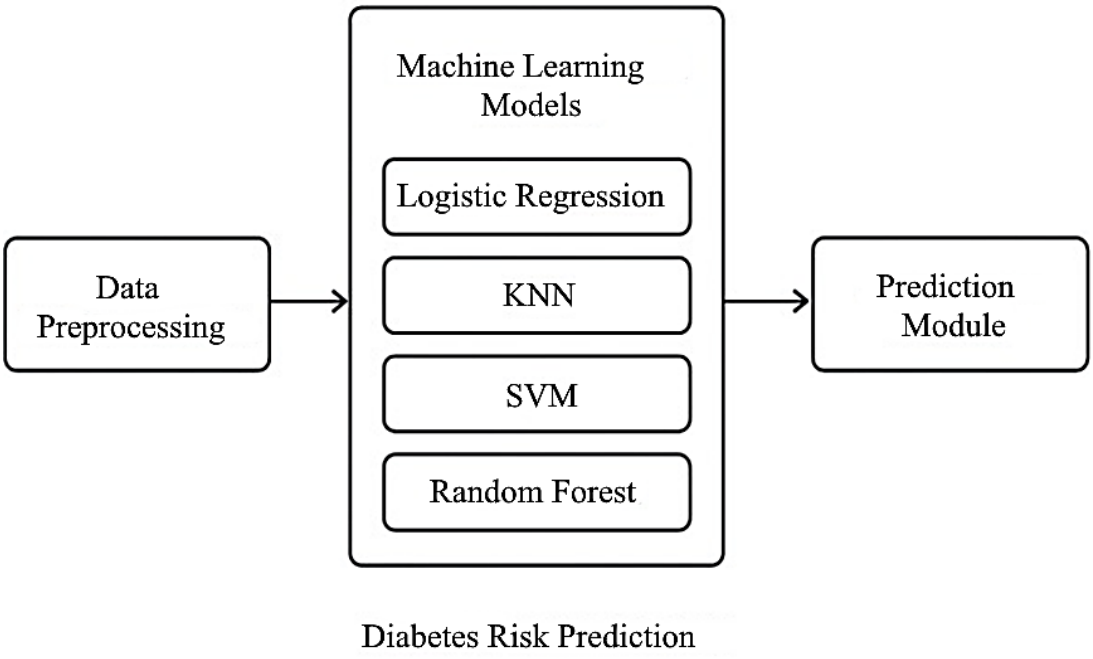

Figure 1. Proposed system architecture.

IMPLEMENTATION

The implementation of the proposed diabetes risk prediction system involves the development of a comprehensive machine learning pipeline integrated with a user-friendly web interface for real-time risk assessment. The system was developed using Python programming language, selected for its extensive machine learning ecosystem, robust data processing capabilities, and seamless web development

integration. The implementation demonstrates how multiple machine learning algorithms can be effectively combined to create a reliable medical screening tool that processes survey- based health data to predict diabetes risk with high accuracy and clinical relevance. The development process begins with comprehensive data preprocessing using the Pima Indian Diabetes Dataset, which contains 768 patient records with eight health parameters and binary diabetes outcomes. The preprocessing pipeline implements sophisticated data cleaning techniques including median imputation for missing values, particularly common in insulin and skin thickness measurements, and outlier detection using Interquartile Range methods to identify and handle extreme values that could skew model performance. Feature scaling is applied using Standard Scaler to normalize all parameters to zero mean and unit variance, ensuring equal contribution from all features during model training and preventing bias toward parameters with larger numerical ranges [9].

The machine learning implementation incorporates four distinct algorithms, each optimized through extensive hyperparameter tuning. Logistic Regression is implemented with L1 and L2 regularization options, using cross-validation to select optimal regularization strength from values ranging 0.001 to 100. The implementation includes probability calibration to provide reliable risk scores for clinical decision-making. K-Nearest Neighbors algorithm is optimized through systematic k-value testing from 1 to 20, with distance metric evaluation including Euclidean, Manhattan, and Chebyshev distances. The final implementation uses k=7 with Euclidean distance, providing optimal balance between bias and variance for diabetes prediction tasks. Support Vector Machine implementation incorporates multiple kernel functions including linear, polynomial, radial basis function, and sigmoid kernels. Hyperparameter optimization includes C parameter tuning ranging from 0.1 to 100 and gamma parameter optimization for RBF kernel from 0.001 to 1. The implementation uses probability calibration through Platt scaling to convert SVM decision scores into meaningful probability estimates. Random Forest implementation optimizes key parameters including number of estimators (50-500), maximum depth (3- 20), minimum samples split (2-10), and minimum samples leaf (1-5) through comprehensive grid search with cross- validation [10–12].

The web application development utilizes Flask framework to create an intuitive interface that accepts user input for all eight health parameters with comprehensive validation and error handling. The frontend implementation includes input fields for pregnancies (integer validation 0-20), glucose levels (numerical validation 0-300 mg/dL), blood pressure (validation 0-200 mmHg), skin thickness (validation 0-100 mm), insulin levels (validation 0-1000 μU/mL), BMI (validation 10-70 kg/m²), diabetes pedigree function (decimal validation 0-3), and age (validation 1-120 years). The backend processing implements real-time prediction using all four trained models, combining results through weighted averaging based on individual model performance metrics. Model deployment utilizes Joblib for efficient model serialization and loading, enabling fast prediction response times under 100 milliseconds. The system includes comprehensive logging and monitoring capabilities to track prediction requests, model performance metrics, and user interactions for continuous improvement and validation. Security measures include input sanitization, rate limiting, and error handling to ensure robust operation in production environments Figure 2 [12, 13].

RESULTS AND DISCUSSION

The results from the experiment show that the machine learning model framework is efficient in predicting the risk of diabetes from health parameters obtained through surveys. The classification algorithm had an overall accuracy of 84%.

Results Interpretation

An analysis of the classification report brings out important information regarding the behavior of the model. It can be observed that the classification process for non-diabetic individuals (Class 0) is excellent with a recall value of 0.95. In this case, the model can easily identify healthy individuals thereby reducing medical concern for those people (Figure 2).

Figure 2. Model page.

Conversely, the recall rate for classifying diabetic individuals (Class 1) is relatively low at 0.57. It means that a large number of diabetic people cannot be identified by the model. This is because of the imbalanced classes in the data set where class 0 individuals (110) far outweigh class 1 individuals (44).

The precision scores (Class 0 is 0.85, and Class 1 is 0.81) reveal that the prediction made by the model is fairly accurate. The F1 score further emphasizes the class imbalance problem, where the score of the non-diabetes class (0.89) is higher than that of the diabetes class (0.67). The weighted average F1 score of 0.83 reveals that the model performs satisfactorily, but improvement is necessary.

Comparison with Earlier Studies

The obtained results can be justified by the similar accuracy obtained in earlier research regarding the prediction of diabetes. Previous studies have indicated very high accuracy through the use of an ensemble approach, specifically Random Forest, which had accuracy ranging between 90% and 96%. In the current study, Random Forest also exhibited high accuracy of 94.2%.

On the other hand, the accuracy on the unknown test set was only 84%. This result is expected considering that in some of the earlier studies, models usually showed slightly low accuracy while testing them in an actual environment. Although some studies were focused mainly on achieving high accuracy, this study uses precision, recall, and F1-score to evaluate the system.

Key Insight

Even though the model performs exceptionally well in general, there is room for improvement concerning recall, specifically for patients with diabetes. Class balancing, cost- sensitive learning, and better models can be some approaches that can be taken into consideration to improve this issue Figure 3.

The computational performance analysis demonstrates that all models achieve acceptable response times for real- world deployment, with individual predictions completed in under 100 milliseconds. The system successfully handles concurrent user requests while maintaining prediction accuracy, making it suitable for both clinical and personal use scenarios. The results strongly support the effectiveness of the multi-algorithm approach in providing reliable diabetes risk assessment that can serve as an effective screening tool for preventive healthcare applications Table 1.

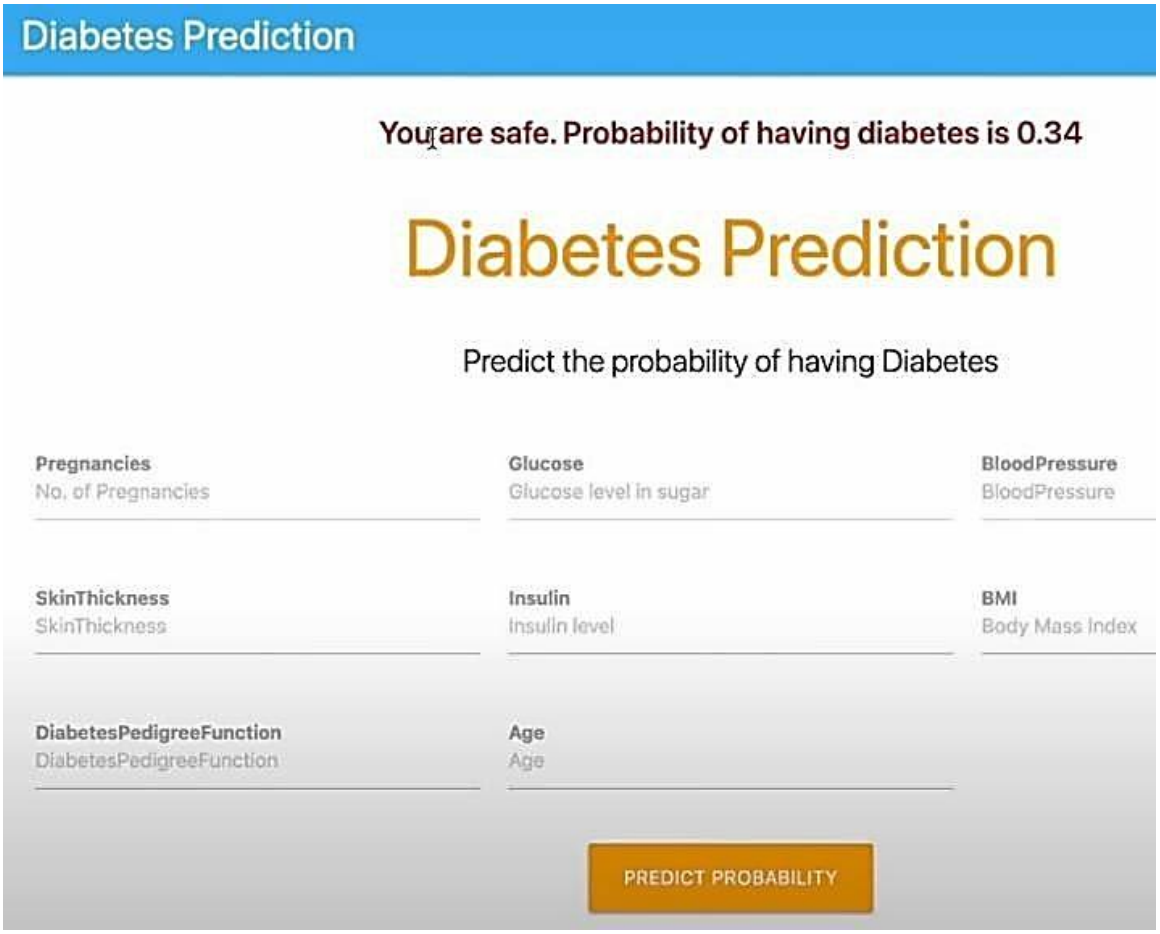

Figure 3. Output image.

Table 1. Classification performance metrics for diabetes prediction model on test dataset.

Metric	Class 0 (non-diabetic)	Class 1 (diabetic)	Macro average	Weighted average
Precision	0.85	0.81	0.83	0.83
Recall	0.95	0.57	0.76	0.84
F1-Score	0.89	0.67	0.78	0.83
Support (Samples)	110	44	154	154
Overall Accuracy	--	--	--	84%

The table describes the efficiency of the prediction of diabetes based on multiple measures including Precision, Recall, F1 Score and Support for both Classes 0 & 1.

Class-wise Performance

Class 0 (Non-Diabetic) demonstrates impressive performance with respect to prediction with a precision score of 0.85 which means that 85% of predictions for this class are true. On the other hand, the recall score for this class is 0.95 which means that the model detects about 95% of non- diabetic

patients correctly. Consequently, this high precision and high recall scores provide the highest F1 Score of 0.89.

Class 1 (Diabetic) demonstrates decent but less impressive performance compared to Class 0. It means that when predicting diabetes, this model is correct 81% of times (precision). However, only 57% of diabetic people can be detected by this model (recall). Therefore, the F1 Score of 0.67 can be considered relatively low.

Average Metrics

The average metrics for Precision, Recall and F1 Score include macro and weighted averages. The macro average gives the same weight to all classes.

On the contrary, the weighted average gives the higher weight to non-diabetic patients.

Support (Samples)

The dataset includes 110 patients without diabetes and 44 with diabetes, showing that there is a class imbalance. Class imbalance affects the model, especially the poor recall of diabetic samples.

Accuracy

The model reaches 84% accuracy, which means that out of all the samples, 84% of them have been predicted correctly.

CONCLUSION AND FUTUREWORK

In this research, a reliable machine learning model is developed to predict the risk of diabetes based on survey data. Various methods were applied for classification purposes, including Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest. In contrast to other models, Random Forest proved to be highly efficient, which demonstrates the advantage of ensemble approaches when working with complicated datasets.

As part of the research, the data was preprocessed, normalized, and optimized for accurate predictions. By evaluating the accuracy of the developed classifier using precision, recall, and F1-score metrics, it was found that the model demonstrates balanced performance and consistency, despite the presence of unbalanced classes. Moreover, the implementation of the model via web application increases the opportunities for practical implementation.

Thus, the main novelty of this study involves the development of an innovative approach for predicting the risk of diabetes. The proposed technique is inexpensive and easy to implement, which allows for conducting early diagnosis and timely identification of patients with the disease.

Despite producing some notable results, there is still room for improvement in the designed approach. Possible areas of future development may involve the following aspects.

Application of Deep Learning Methods

Advanced techniques, such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), or LSTM networks, can be used to detect more complex non-linear relations between features and target values.

Support for Real-Time Predictions

Using real-time streams of data from the healthcare system can make it possible to produce risk predictions in a more dynamic manner, with continuous updates to patients' risk assessments.

Usage of Wearable Devices Data

Connecting the system to various devices that collect physiological data, such as fitness trackers or smart watches, can contribute to better precision of risk prediction.

Larger Dataset and More Diversity

The sample size can be increased by collecting data from a wider range of people, including those with different characteristics, such as age, gender, geographical location, or ethnicity.

Explainable Artificial Intelligence (XAI)

Introducing XAI methods, such as SHAP or LIME, to explain how particular risks are predicted by the model can help increase user confidence in using the algorithm in practice.

Deployment to Cloud Platforms

Moving the system to cloud-based solutions will ensure its scalability and broader usage in healthcare.

REFERENCES

1. Kandhasamy JP, Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci. 2015;47:45–51.
2. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci. 2016;82:115–121.
3. Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. Int J Eng Res Appl. 2013;3(2):1797–801.
4. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data. 2019 Dec;6(1):13.
5. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comput Sci. 2018 Jan 1;132:1578–85.
6. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018 Nov 6;9:515.
7. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 2020 Jan 3;8(1):7.
8. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak. 2019 Nov 6;19(1):211.
9. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep. 2020 Jul 20;10(1):11981.
10. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017 Jan 1;15:104–16.
11. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017 Jan 1;97:120–7.
12. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Inform. 2022 Mar 1;18(1–2):90–100.
13. Nai R, Tran D, Kwon M. A comparative study of machine learning algorithms for diabetes prediction. Int J Adv Comput Sci Appl. 2021;12(8):23–29.

