Comparative Analysis of Machine Learning Algorithms for Early Detection of Diabetes

[1] Park Ji Hyun, [2] Rohan Sharma
[1] Yonsei University, Seoul, South Korea
[2] Indian Institute of Technology (IIT) Bombay, Mumbai, India
Corresponding E-mail: park.hyun@interviauniversity.com

Abstract

Diabetes is a common health problem that affects people in many parts of the world. Detecting it early is important because timely action can help control the condition and reduce the risk of serious complications. Traditionally, diabetes is diagnosed through clinical examination and laboratory tests. With the growing amount of healthcare data available, machine learning can also be used to find patterns that may help identify people who are at risk. This paper compares four machine learning algorithms for the early detection of diabetes: Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. The general process includes data preprocessing, feature preparation, model development, and performance evaluation. The models can use patient information such as glucose level, blood pressure, body mass index, age, and insulin level. Their performance can be assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Each method has its own strengths. Logistic Regression is easy to understand, SVM is useful for more complex classification problems, Random Forest can handle nonlinear relationships, and XGBoost is a powerful method for structured data.

Keywords: Machine Learning, Diabetes Detection, Logistic Regression, Support Vector Machine, Random Forest, XGBoost, Healthcare, Classification

1. Introduction

Diabetes mellitus is a long-term metabolic disease that develops when the body does not produce enough insulin or is unable to use it properly. This causes blood glucose levels to rise and remain high over time. If diabetes is not identified and managed properly, it can lead to serious health problems such as heart disease, kidney damage, vision problems, and nerve

damage. For this reason, finding diabetes at an early stage is an important part of prevention and management.

Type 2 diabetes is especially important because many of its risk factors are connected to lifestyle and demographic characteristics. Age, obesity, lack of physical activity, family history, blood pressure, and glucose levels can all increase the likelihood of developing the disease. Some people may not notice clear symptoms in the early stages. Because of this, methods that can identify people who may be at higher risk could be useful for screening and early intervention [1].

Machine learning is a part of artificial intelligence that enables computers to learn patterns from data and make predictions without having to be programmed for every individual case [2]. In healthcare, it has been explored for tasks such as disease prediction, medical diagnosis, risk assessment, medical image analysis, and clinical decision support.

Diabetes prediction is a suitable area for machine learning because patient records usually contain several variables that may be related to the disease. For example, glucose level, body mass index, age, blood pressure, insulin level, and pregnancy history can provide useful information when considered together. Instead of looking at each factor separately, a machine learning model can learn how different variables are related to diabetic and non-diabetic cases[3] .

Different algorithms can give different results even when they are trained on the same healthcare dataset. Choosing an appropriate algorithm is therefore an important part of building a reliable diabetes detection system. Logistic Regression is a simple and interpretable method that can be used as a baseline. SVM can find more complex boundaries between classes. Random Forest combines several decision trees and can model nonlinear relationships, while XGBoost improves a group of decision trees step by step and often works well with structured data.

The purpose of this research is to compare these four machine learning algorithms for the early detection of diabetes. The study looks at their general methodology, preprocessing needs, characteristics, and evaluation measures. The comparison is intended to highlight the

strengths and limitations of each method and to identify which approaches may be useful for a diabetes screening system[4].

2. Methodology

The study uses a supervised machine learning classification approach. The main goal is to predict whether a person belongs to the diabetic or non-diabetic group using available health-related information.

The overall process consists of five main stages:

1. Data collection
2. Data preprocessing
3. Feature preparation
4. Machine learning model development
5. Model evaluation and comparison

Four algorithms are considered in the comparison: Logistic Regression, Support Vector Machine, Random Forest, and XGBoost.

2.1 Research Approach

The study uses a supervised machine learning classification approach. The main goal is to predict whether a person belongs to the diabetic or non-diabetic group using available health-related information.

The overall process consists of five main stages:

1. Data collection
2. Data preprocessing
3. Feature preparation
4. Machine learning model development
5. Model evaluation and comparison

Four algorithms are considered in the comparison: Logistic Regression, Support Vector Machine, Random Forest, and XGBoost[5].

2.2 Dataset

For this type of study, a publicly available diabetes dataset can be used for experimental analysis. A commonly used dataset contains patient information along with several health-related measurements. These features may include the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index, diabetes pedigree function, and age.

The target variable indicates whether or not the patient has diabetes. This makes the task a binary classification problem because the model has to predict one of two possible outcomes.

The selected features are useful because diabetes is influenced by a combination of physiological and demographic factors. Glucose concentration is usually one of the most informative variables, while BMI, age, blood pressure, insulin, and family-related risk information may also contribute to the prediction [6].

2.3 Data Preprocessing

Data preprocessing is an important part of any machine learning project because the quality of the input data can directly affect the results. Before training the models, the dataset should be checked for missing values, duplicate records, unusual values, and inconsistent measurements.

In some healthcare datasets, a value of zero may appear for a variable where zero would not normally be a realistic physiological measurement. These values need to be examined carefully and, when appropriate, treated as missing data. Suitable imputation methods can then be used to fill the missing information [7].

Feature scaling is also important for some machine learning algorithms. Logistic Regression and SVM can be influenced by differences in the numerical ranges of the input variables. Standardization can therefore be used to place the features on comparable scales.

Tree-based methods such as Random Forest and XGBoost generally do not depend on feature scaling to the same extent. Even so, using a consistent preprocessing approach is helpful when comparing different models.

After preprocessing, the data should be divided into training and testing sets. The training set is used to build the models, while the testing set is kept separate so that performance can be

checked on data the models have not seen before. Cross-validation can also be used during training to get a more dependable estimate of model performance[8].

2.4 Logistic Regression

Logistic Regression is a statistical machine learning method that is commonly used for binary classification. In this study, it can be used to estimate the probability that a patient belongs to the diabetic group.

One of the main benefits of Logistic Regression is that it is simple and relatively easy to interpret. Its coefficients provide information about the relationship between the input variables and the prediction. This is especially useful in healthcare, where medical professionals may want to understand the factors behind a model's prediction.

The main limitation is that Logistic Regression works best when the relationship between the predictors and the outcome is reasonably simple. If the relationships between patient characteristics are highly nonlinear, the model may not be able to capture all of the important patterns[9].

2.5 Support Vector Machine

Support Vector Machine is a supervised learning algorithm that tries to find a boundary that separates different classes. It focuses on important observations, known as support vectors, which help determine the position of the classification boundary.

SVM can also use kernel functions to deal with nonlinear relationships. This gives it more flexibility than a basic linear classifier. However, its performance can depend heavily on choosing suitable parameters and scaling the input features[10].

For diabetes detection, SVM can be useful when the relationship between clinical variables and diabetes status is complicated. Its ability to create nonlinear decision boundaries may help it recognize patterns that simpler classification methods might miss.

2.6 Random Forest

Random Forest is an ensemble learning method that combines the predictions of many decision trees. Each tree is built using different subsets of the observations and features, and the final prediction is made by combining the results of all the trees [11].

A major advantage of Random Forest is its ability to handle nonlinear relationships. Diabetes risk can depend on interactions between factors such as age, BMI, glucose level, and blood pressure. Random Forest can learn these interactions without requiring the researcher to define them manually.

Another useful feature is that Random Forest can provide information about feature importance. This can help researchers understand which variables are most useful for making predictions. However, feature importance should be viewed as an indication of predictive relevance and not as proof that a particular variable directly causes diabetes.

2.7 XGBoost

XGBoost is a gradient-boosting algorithm based on decision trees. Unlike methods that build trees independently, XGBoost creates them one after another. Each new tree is designed to improve the errors made by the previous trees.

The algorithm includes regularization and optimization techniques that can help improve prediction performance and reduce overfitting when the parameters are chosen properly. Because it works well with structured or tabular data, XGBoost is a suitable method to include in a diabetes prediction comparison.

The main drawback of XGBoost is its complexity. Compared with Logistic Regression, it is harder to interpret directly. It can also require careful hyperparameter tuning to obtain reliable results.

3. Results and Discussion

The comparison considers Logistic Regression, SVM, Random Forest, and XGBoost for diabetes classification. Since the four algorithms use different learning approaches, each one has its own strengths and weaknesses [12].

Logistic Regression is simple and easy to interpret, which makes it a useful baseline model. However, it may struggle with complicated nonlinear relationships. SVM can deal with nonlinear classification by using kernel functions, but its performance depends on suitable parameter settings and feature scaling.

Random Forest is able to capture nonlinear relationships and interactions among different health-related variables. It is also fairly robust and can provide information about feature importance. XGBoost is a powerful boosting method that can learn complex patterns in structured datasets, although it needs careful tuning to reduce the risk of overfitting [13].

For diabetes detection, accuracy by itself is not enough to judge a model. Incorrectly classifying a person with diabetes as non-diabetic could delay further medical attention. Therefore, precision, recall, F1-score, and ROC-AUC should also be considered. Recall is especially important because it shows how well the model identifies people who actually have diabetes[14].

Overall, there is no single algorithm that can be considered the best choice for every diabetes detection problem. The most suitable model depends on the dataset, the required level of performance, and how important interpretability is. Data quality and proper validation are also important because a model that works well on one dataset may not perform in the same way on a different population [15].

Machine learning can be a useful support tool for early diabetes screening, but its predictions should assist professional medical assessment rather than replace a proper medical diagnosis[16].

4. Conclusion

This study compared four machine learning algorithms—Logistic Regression, Support Vector Machine, Random Forest, and XGBoost—for the early detection of diabetes. The discussion covered the main steps involved in the machine learning process, including dataset preparation, preprocessing, feature preparation, model development, and evaluation.

Each algorithm offers different advantages. Logistic Regression is simple and easy to interpret, which makes it a useful baseline when transparency is important. SVM provides flexible classification and can deal with nonlinear relationships. Random Forest can capture complex interactions between health-related variables and can also provide information about feature importance. XGBoost is a powerful method for structured datasets, although it requires careful tuning and validation.

The comparison also shows that accuracy alone is not enough when evaluating a healthcare prediction model. Precision, recall, F1-score, and ROC-AUC should also be taken into account. In particular, recall is important for early diabetes detection because failing to identify someone who may have diabetes could delay further medical evaluation.

Overall, machine learning has strong potential as a tool for supporting early diabetes screening. However, dependable results require good-quality data, appropriate preprocessing, proper model validation, and careful interpretation of predictions. Future work could test these four algorithms on larger and more diverse datasets, carry out more systematic hyperparameter optimization, and validate the models using independent patient populations. With appropriate validation and responsible use, machine learning can provide useful decision support for identifying people who may be at increased risk of diabetes.

References:

[1] J. Barach, "Federated Learning for Privacy-Preserving Employee Performance Analytics," *IEEE Access,* vol. 13, 2025, doi: 10.1109/ACCESS.2025.3591360.

[2] D. Bodra and S. Khairnar, "Accelerating and analyzing performance of shortest path algorithms on GPU using CUDA platform: Bellman-Ford, Dijkstra, and Floyd-Warshall algorithms," *Научно-технический вестник информационных технологий, механики и оптики,* vol. 25, no. 5, pp. 866-875, 2025.

[3] R. Ramadugu, "A Formalized Approach to Secure and Scalable Smart Contracts in Decentralized Finance," in *2025 International Conference on Engineering, Technology & Management (ICETM)*, 2025: IEEE, pp. 1-6.

[4] E. Ahmed, "Student performance prediction using machine learning algorithms," *Applied computational intelligence and soft computing,* vol. 2024, no. 1, p. 4067721, 2024.

[5] F.-Z. Nakach, A. Idri, and E. Goceri, "A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification," *Artificial Intelligence Review,* vol. 57, no. 12, p. 327, 2024.

[6] A. Mohammed, "Artificial Intelligence-Driven Cybersecurity Crimes: Emerging Threats, AI Bots, and Advanced Defense Strategies," *Euro Vantage journals of Artificial intelligence,* vol. 3, no. 1, pp. 46-65, 2026, doi: https://doi.org/10.65923/c2sxrx21.

[7] A. Mohammed, "Transforming SOC Operations: Harnessing the Power of AI and ML for Enhanced Threat Detection," *INTERNATIONAL JOURNAL OF RESEARCH CULTURE SOCIETY Monthly Peer-Reviewed, Refereed, Indexed,* vol. 8, pp. 11-21, 2024, doi: 10.2017/IJRCS/ICCDMP01.

[8] M. Beseiso, "Enhancing student success prediction: A comparative analysis of machine learning technique," *TechTrends,* vol. 69, no. 2, pp. 372-384, 2025.

[9] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga, and N. Kuruwitaarachchi, "Real-time credit card fraud detection using machine learning," in *2019 9th international conference on cloud computing, data science & engineering (Confluence)*, 2019: IEEE, pp. 488-493.

[10] A. Vladyko, A. Khakimov, A. Muthanna, A. A. Ateya, and A. Koucheryavy, "Distributed edge computing to assist ultra-low-latency VANET applications," *Future Internet,* vol. 11, no. 6, p. 128, 2019.

[11] S. Khairnar and D. Bodra, "Analysis and Evaluation of Modern Lightweight Cryptographic Algorithms: Standards, Hardware Implementation, and Security Considerations," *International Journal of Computer Applications,* vol. 975, p. 8887, 2025.

[12] A. Qatawneh, "The influence of data mining on accounting information system performance: a mediating role of information technology infrastructure," *Journal of Governance and Regulation/Volume,* vol. 11, no. 1, pp. 141-151, 2022, doi: https://doi.org/10.22495/jgrv11i1art13.

[13] J. Barach, "Enhancing Ransomware Resilience in Cloud-Based HR Systems through Moving Target Defense," *Computers, Materials and Continua,* vol. 86, no. 2, pp. 1-23, 2025, doi: https://doi.org/10.32604/cmc.2025.071705.

[14] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory,* vol. 55, pp. 1-9, 2015.

[15] A. M. Qatawneh, "The effect of electronic commerce on the accounting information system of Jordanian banks," *International Business Research,* vol. 5, no. 5, p. 158, 2012, doi: 10.5539/ibr.v5n5p158.

[16] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via improved machine learning techniques," *Procedia Computer Science,* vol. 174, pp. 433-442, 2020.