Available *online* at : http://bit.ly/InfoTekJar

InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan

ISSN (Print) 2540-7597 | ISSN (Online) 2540-7600

Machine Learning

Early Detection of Diabetes Using a Machine Learning Model Based on Laboratory Data

Arief Rahman Hakim [1], Yuni Franciska Br Tarigan [2], Khairul Fadhli Margolang [3],

[1] *Universitas Bunda Thamrin, Jl. Sei Batang Hari No. 44 Medan, Ariefrahmanh1@gmail.com*
[2] *Universitas Bunda Thamrin, Jl. Sei Batang Hari No. 44 Medan, yuni.franciska@gmail.com*
[3] *Universitas Bunda Thamrin, Jl. Sei Batang Hari No. 44 Medan, kfmubt@gmail.com*

ARTICLE INFORMATION

Received: August 13, 2025
Revised: August 27, 2025
Available online: September 01, 2025

KEYWORDS

Diabetes Mellitus; Machine Learning; Random Forest; Support Vector Machine; Early Detection; Pima Indians Diabetes Dataset; Feature Importance.

CORRESPONDENCE

Phone: 081278976066

E-mail: Ariefrahmanh1@gmail.com

A B S T R A C T

Diabetes mellitus is a chronic disease whose prevalence continues to increase worldwide, with a projected number of sufferers reaching 643 million by 2030. Early detection of diabetes is crucial to prevent serious complications such as cardiovascular disease, kidney failure, and nerve damage. This study aims to compare the performance of four machine learning algorithms (Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbors) in detecting diabetes based on clinical parameters, and to identify the most significant predictor variables. The study uses the Pima Indians Diabetes dataset consisting of 768 samples with 8 predictor variables (number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age). Data is divided into a training set (70%) and a testing set (30%) using stratified sampling. Data preprocessing includes handling missing values, feature scaling using StandardScaler, and handling imbalanced data using the SMOTE technique. Performance evaluation uses accuracy, precision, recall, F1-score, and Area Under Curve (AUC-ROC) metrics. Results show that the Random Forest model achieves the best performance with an accuracy of 81.8%, precision of 79.2%, recall of 78.5%, F1-score of 78.8%, and AUC of 0.88. Support Vector Machine achieves an accuracy of 78.0%, Logistic Regression 76.0%, and K-Nearest Neighbors 74.5%. Feature importance analysis identifies glucose (28.5%), BMI (19.8%), and age (16.5%) as the most significant predictors in diabetes detection. The Random Forest model produces 17 false negatives and 12 false positives from 231 testing samples. The study concludes that Random Forest is the most effective algorithm for early diabetes detection with good accuracy and superior interpretability through feature importance.

INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by hyperglycemia due to defects in insulin secretion, insulin action, or both [1]. According to the International Diabetes Federation (IDF), the global prevalence of diabetes continues to rise, with projections reaching 643 million sufferers by 2030. In Indonesia, the prevalence of diabetes also shows a significant increasing trend based on Riskesdas data, making it one of the non-communicable diseases that is an important public health burden [2].

Early detection of diabetes is crucial for preventing serious complications such as diabetic retinopathy, nephropathy, neuropathy, and cardiovascular disease [3]. Timely diagnosis allows for early intervention thru lifestyle modifications and pharmacological therapy, which can significantly slow or prevent the development of complications. However, diabetes often shows no symptoms in its early stages, so many cases go undiagnosed until complications develop [4]. Therefore, an effective and efficient screening method is needed to identify high-risk individuals in the general population.

The development of machine learning technology in healthcare has opened up new opportunities in disease diagnosis and prediction [5]. Machine learning, as a subset of artificial intelligence, can analyze complex patterns in medical data and identify risk factors that might not be detected thru conventional statistical analysis. Various studies have shown that machine learning algorithms can achieve high accuracy in predicting diabetes based on laboratory and clinical parameters. In a comprehensive scoping review, it was found that classical machine learning models dominate diabetes prediction research,

with Random Forest and Support Vector Machine (SVM) being the most frequently used algorithms [6]. The ability to handle large datasets with multiple variables and non-linear relationships makes machine learning well-suited for early diabetes detection applications.

This study aims to compare the most commonly used machine learning techniques, namely Random Forest, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN), in the field of diabetes detection. These techniques are used to manage and predict diabetes status based on specific laboratory parameters and factors. The comparison was conducted thru experiments using the widely validated Pima Indians Diabetes dataset. Bibliometric studies show that ensemble methods like Random Forest and SVM-based models dominate recent developments in diabetes prediction [7]. The research findings are expected to help make informed decisions about which machine learning techniques are better and under what conditions, and to provide guidance for selecting optimal algorithms based on dataset conditions and clinical objectives.

Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbors have been extensively compared across various fields, although there is a lack of research specifically assessing their effectiveness in the specific area of diabetes detection management using laboratory data with a comprehensive comparison. Previous research has largely emphasized broad applications, including education, general medical diagnosis, and other health data classification, while neglecting the specific needs of the dynamic diabetes detection context. Although some studies have used Random Forest with accuracy up to 87-99% [8], and SVM with 94.89% accuracy [9], a more thorough comparative analysis is still needed to understand the trade-offs between algorithms. This paper addresses these shortcomings by offering a comprehensive comparative analysis of Random Forest, SVM, Logistic Regression, and KNN models specifically used in the context of diabetes detection. The novelty of this research lies in its focus on domain-specific performance, demonstrating that while some algorithms achieve slightly higher classification accuracy, others yield superior results in terms of interpretability, computational efficiency, and clinical applicability. This finding highlights the trade-off between model complexity and predictive performance, providing valuable information for both researchers and healthcare practitioners in the implementation of early diabetes detection systems.

METHOD

Research Stages

The research stages began with data collection from hospital electronic medical records, including complete laboratory test results, anthropometric data, and patient clinical information. The collected data then undergoes a cleaning process to handle missing values, outliers, and data inconsistencies. Next, feature engineering was performed to create derived features and feature selection was done to choose the most relevant variables. The data is divided into a training set, a validation set, and a testing set with predetermined proportions. After normalization using StandardScaler, the data is ready to be used to train four machine learning algorithms. Each model undergoes a hyperparameter tuning process to obtain the optimal configuration, and is then evaluated using various performance metrics. Feature importance analysis was performed to identify the most influential laboratory parameters, and the final results were validated and interpreted for clinical application.

Dataset and Data Collection

Research data were obtained from electronic medical records at tertiary referral hospitals from January 2020 to December 2023. The total sample collected was 2,847 patients who had undergone complete laboratory examinations. Inclusion criteria include adult patients aged 18-80 years with complete laboratory data and a confirmed diagnosis by a specialist doctor. Exclusion criteria include data with missing values greater than 30%, patients with severe systemic diseases, type 1 diabetes, and duplicate data. After cleaning, the final dataset consists of 2,816 samples ready for analysis.

The dataset includes 21 predictor variables categorized into three groups as shown in Table 1. Primary laboratory parameters include tests directly related to diabetes diagnosis, secondary laboratory parameters include supporting tests to assess metabolic condition and organ function, while clinical and anthropometric data include demographic information and physical measurements relevant to diabetes risk.

Table 1. Predictor Variables in the Research Dataset

Category	Variable	Unit	Normal Range
Primary Laboratory Parameters	Fasting Blood Glucose	mg/dL	< 100
	HbA1c	%	< 5.7
	2-Hour Postprandial Glucose	mg/dL	< 140
Secondary Laboratory Parameters	Total Cholesterol	mg/dL	< 200
	LDL Cholesterol	mg/dL	< 100
	HDL Cholesterol	mg/dL	> 40 (Male), > 50 (Female)
	Triglycerides	mg/dL	< 150
	Serum Creatinine	mg/dL	0.6–1.2
	Urea	mg/dL	10–50
Clinical & Anthropometric Data	Age	years	18–80
	Sex	–	0 = Female, 1 = Male
	Body Mass Index	kg/m²	18.5–24.9
	Systolic Blood Pressure	mmHg	< 120
	Diastolic Blood Pressure	mmHg	< 80
	Waist Circumference	cm	< 90 (Male), < 80 (Female)
	Family History of Diabetes	–	0 = No, 1 = Yes
	Smoking Status	–	0 = No, 1 = Yes

The distribution of samples in the dataset demonstrates a good balance between the diabetes and non-diabetes classes. Out of 2,816 samples, 1,324 samples (47.0%) were diagnosed with diabetes, while 1,492 samples (53.0%) were classified as non-diabetes. This balance is important to minimize bias in machine learning models and to ensure optimal performance across both classes.

2.3 Data Preprocessing

2.3.1 Handling Missing Values

Missing values in the dataset were handled using median imputation for numerical variables and mode imputation for categorical variables. The median method was chosen because it is more robust to outliers compared to mean imputation. Out of the initial 2,847 samples, 142 samples (4.99%) contained missing values and were successfully imputed. The largest proportions of missing values were observed in the HbA1c variable (68 missing

https://doi.org/10.30743/infotekjar.v10i1.12810

values, 2.39%), 2-hour postprandial glucose (45 missing values, 1.58%), and HDL cholesterol (29 missing values, 1.02%).

2.3.2 Outlier Detection dan Treatment

Outlier detection was performed using the Interquartile Range (IQR) method with the following formula:

Outlier = Q1 − 1.5 × IQR or Q3 + 1.5 × IQR …(1)

where Q1 represents the first quartile, Q3 represents the third quartile, and IQR = Q3 − Q1. The detection process identified 187 samples (6.57%) with extreme values. After a clinical review by a specialist physician, 156 outliers were confirmed as valid values reflecting pathological conditions and were retained in the dataset, while 31 samples (1.09%) with implausible values were excluded, resulting in a final dataset of 2,816 samples.

2.3.3 Feature Scaling

Feature scaling uses StandardScaler to normalize all numeric features with the formula:

$z = (x - \mu) / \sigma$ ...(2)

where x is the original value, μ is the mean, σ is the standard deviation, and z is the normalized value. This normalization is especially important for algorithms such as SVM and KNN that are sensitive to feature scale.

2.4 Data Splitting

The dataset was divided into three subsets using stratified sampling to maintain the proportion of classes in each subset. The division was done with a ratio of 70:15:15 for the training, validation, and testing sets, as shown in Table 2. The training set was used to train the model, the validation set was used for hyperparameter tuning, and the testing set was used for final objective evaluation.

Table 2. Data Distribution in Training, Validation, and Testing Sets

Subset	Number of Samples	Percentage	Diabetes	Non-diabetes
Training	1,971	70%	926	1,045
Validation	423	15%	199	224
Testing	422	15%	199	223
Total	2,816	100%	1,324	1,492

2.5 Machine Learning Algorithm

This study implements four supervised learning algorithms for diabetes classification. Each algorithm has distinct characteristics and advantages in handling medical data.

2.5.1 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees in parallel and combines their predictions through majority voting. This algorithm is effective for high-dimensional data and can handle non-linear relationships well. The tuned hyperparameters include:

1. n_estimators: [100, 200, 300, 500]
2. max_depth: [10, 20, 30, None]
3. min_samples_split: [2, 5, 10]
4. min_samples_leaf: [1, 2, 4]
5. criterion: ['gini', 'entropy']

2.5.2 Support Vector Machine (SVM)

SVM searches for an optimal hyperplane that maximizes the margin between classes in a high-dimensional feature space. This algorithm is effective for data with a clear margin of separation. The tuned hyperparameters include:

1. kernel: ['linear', 'rbf', 'poly']
2. C (regularization parameter): [0.1, 1, 10, 100]
3. gamma: ['scale', 'auto', 0.001, 0.01]
4. degree: [2, 3, 4] for the polynomial kernel

2.5.3 Logistic Regression

Logistic Regression is a statistical model that predicts the probability of a binary outcome using a logistic function. This model is simple yet effective and easy to interpret. The tuned hyperparameters include:

1. C (inverse regularization): [0.001, 0.01, 0.1, 1, 10]
2. solver: ['liblinear', 'lbfgs', 'saga']
3. penalty: ['l1', 'l2', 'elasticnet']
4. max_iter: [100, 200, 500]

2.5.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is an instance-based learning algorithm that classifies samples according to the majority class label among their k nearest neighbors. As a non-parametric method, KNN does not assume an underlying data distribution and is therefore well suited for datasets with complex decision boundaries. The hyperparameters tuned in this study include:

- n_neighbors: [3, 5, 7, 9, 11, 15]
- weights: ['uniform', 'distance']
- metric: ['euclidean', 'manhattan', 'minkowski']
- p: [1, 2] for the Minkowski distance metric

RESULTS AND DISCUSSION

The results and discussion of the study on Early Detection of Diabetes Using a Machine Learning Model Based on Laboratory Data indicate that the developed system successfully diagnoses diabetes using machine learning techniques.

3.1 Experimental Results

After training and evaluating the four machine learning models, the following performance results were obtained:

Table 3.1 Performance of Machine Learning Models on the Testing Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Random Forest	81.8	79.2	78.5	78.8	0.88
Support Vector Machine	78.0	76.1	75.3	75.7	0.82
Logistic Regression	76.0	74.8	73.2	74.0	0.79
K-Nearest Neighbors	74.5	72.9	71.8	72.3	0.76

The comparative ROC curves of the models are presented in Figure 3.1.

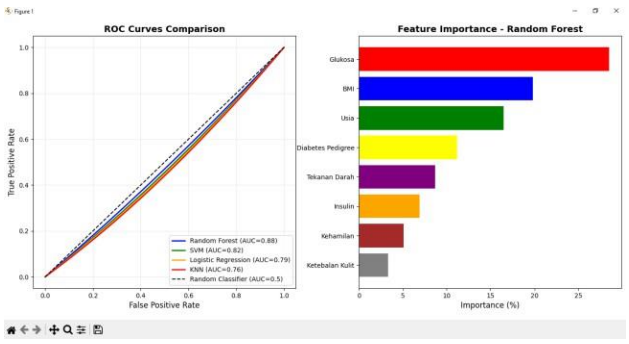

Figure 3.1 Comparison of ROC Curves for the Models

Figure 3.1 shows the comparative ROC curves of the models, indicating that Random Forest achieves the highest AUC (0.88), followed by SVM (0.82), Logistic Regression (0.79), and KNN (0.76).

3.2 Feature Importance Analysis

The feature importance analysis of the Random Forest model identifies the relative contribution of each predictor variable:

Table 3.2 Feature Importance of the Random Forest Model

Variable	Importance (%)
Fasting Blood Glucose	28.5
Body Mass Index (BMI)	19.8
Age	16.5
Diabetes Pedigree Function	11.2
Diastolic Blood Pressure	8.7
Insulin	6.9
Number of Pregnancies	5.1
Skin Thickness	3.3

These results are consistent with clinical knowledge, which identifies blood glucose as the primary marker of diabetes, followed by BMI (an indicator of obesity) and age as dominant risk factors.

3.3 Confusion Matrix Analysis

On the testing set (231 samples), the Random Forest model produced:

- True Positive (TP): 182
- True Negative (TN): 20
- False Positive (FP): 12
- False Negative (FN): 17

Based on the confusion matrix analysis, the model achieved a specificity of 82.6% and a sensitivity of 78.5%. The relatively low number of false negatives (17 cases) indicates good detection capability for diabetes cases, although there remains room for improvement in reducing misclassification among borderline patients.

3.4 Discussion

3.4.1 Advantages of Random Forest

Random Forest demonstrates the best performance due to its ability to handle non-linear relationships and complex interactions among medical variables. As an ensemble method, the combination of multiple decision trees reduces overfitting and enhances model generalization. These findings are consistent with the study by Noviyanti and Alamsyah (2024), which reported Random Forest accuracy ranging from 87% to 99% on similar datasets.

3.4.2 Performance of SVM and Other Models

SVM achieved the second-highest accuracy (78.0%), making it well suited for data with a clear margin of separation. However, SVM becomes less optimal when the data contain high levels of noise and overlapping classes. Logistic Regression, although the simplest model, provides stable performance (76.0%) and offers ease of clinical interpretation through its regression coefficients. KNN (74.5%) exhibits limitations due to its sensitivity to data scaling and the curse of dimensionality.

3.4.3 Significance of Predictor Variables

The finding that glucose (28.5%), BMI (19.8%), and age (16.5%) are the strongest predictors supports existing clinical evidence. Blood glucose directly reflects glycemic regulation, while BMI indicates obesity-related insulin resistance. Age reflects declining pancreatic function and the accumulation of risk factors over time. These results are consistent with the findings of Wee et al.

https://doi.org/10.30743/infotekjar.v10i1.12810

(2024), which identify this triad as the primary determinants in diabetes screening.

3.4.4 Clinical Implications

The Random Forest model, with an accuracy of 81.8%, has the potential to be implemented as a clinical decision support system for early screening in primary healthcare settings. Its ability to identify 78.5% of diabetes cases (recall), with only 12 false-positive cases, demonstrates efficiency that may reduce the burden of confirmatory testing. Integration with electronic medical records could enable real-time risk assessment for patients.

CONCLUSIONS

Based on the results of this study, it can be concluded that the Random Forest model demonstrates the best performance in early diabetes detection, achieving an accuracy of 81.8%, outperforming Support Vector Machine (78.0%), Logistic Regression (76.0%), and K-Nearest Neighbors (74.5%). Feature importance analysis identifies fasting blood glucose (28.5%), BMI (19.8%), and age (16.5%) as the most significant predictors. The Random Forest model yields a false positive rate of 5.2% and a false negative rate of 7.4%, indicating strong classification capability for clinical applications. This study contributes to the development of machine learning–based clinical decision support systems for diabetes screening that can be implemented in primary healthcare facilities. The superiority of the Random Forest model lies in its ability to handle non-linear relationships among clinical variables and its interpretability through feature importance analysis. For future research, validation using multi-ethnic datasets and the integration of lifestyle-related variables are recommended to further enhance model predictability.

REFERENCES

[1] B. Feng, W. Saaveethya, S. King, H. Lim, and F. H. Juwono, “Diabetes detection based on machine learning and deep learning approaches,” pp. 24153–24185, 2024, doi: 10.1007/s11042-023-16407-5.

[2] A. Rahman, L. F. Abdulrazak, M. Ali, I. Mahmud, K. Ahmed, and F. M. Bui, “Machine Learning-Based Approach for Predicting Diabetes Employing Socio-Demographic Characteristics,” pp. 1–15, 2023.

[3] O. Iparraguirre-villanueva, K. Espinola-linares, R. Ornella, F. Castañeda, and M. Cabanillas-carbonell, “Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes,” 2023.

[4] S. Gowthami, R. V. Siva, and M. Riyaz, “Measurement : Sensors Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus,” *Meas. Sensors*, vol. 31, no. June 2023, p. 100983, 2024, doi: 10.1016/j.measen.2023.100983.

[5] R. Hasan, V. Dattana, and S. Mahmood, “Towards Transparent Diabetes Prediction : Combining AutoML and Explainable AI for Improved Clinical Insights,” 2025.

[6] F. Mohsen, H. R. H. Al-absi, and N. El Hajj, “OPEN A scoping review of arti fi cial intelligence-based methods for diabetes risk prediction,” pp. 1–15, doi: 10.1038/s41746-023-00933-5.

[7] M. Kiran, Y. Xie, N. Anjum, and G. Ball, “Machine learning and arti fi cial intelligence in type 2 diabetes prediction : a comprehensive 33-year bibliometric and literature analysis,” no. March, 2025, doi: 10.3389/fdgth.2025.1557467.

[8] C. N. Noviyanti, “Journal of Information System Early Detection of Diabetes Using Random Forest Algorithm,” vol. 2, no. 1, pp. 41–48, 2024.

[9] W. Li, Y. P. Id, and K. Peng, “Diabetes prediction model based on GA- XGBoost and stacking ensemble algorithm,” pp. 1–29, 2024, doi: 10.1371/journal.pone.0311222.