Journal of Technology and Informatics (JoTI)
Vol. 8, No. 1, Bulan April Tahun 2026
P-ISSN 2721-4842
E-ISSN 2686-6102

Comparative Performance of Machine Learning Algorithms for Diabetes Prediction

I Made Ardi Sudestra[1*], Adie Wahyudi Oktavia Gama[2], Gede Humaswara Prathama[3], I Gusti Ngurah Darma Paramartha[4], Musawer Hakimi[5]

[1*]Department of Computer Science, Universitas Pendidikan Ganesha, Bali, Singaraja, Indonesia
[2,3,4]Department of Information Technology, Universitas Pendidikan Nasional, Bali, Denpasar, Indonesia
[5]Department of Computer Science, Samangan University, Samangan, Afghanistan
e-mail: ardi.sudestra@student.undiksha.ac.id[1*], adiewahyudi@undiknas.ac.id[2], huma@undiknas.ac.id[3], ngurahdarma@undiknas.ac.id[4], musawer@adc.edu.in[5]

Article Information

Article History:
Received : 8 August 2025
Revised : 8 January 2026
Accepted : 26 March 2026
Published : 16 April 2026

***Correspondence:**
ardi.sudestra@student.undiksha.ac.id

Keywords:
Diabetes Mellitus, Machine Learning, Random Forest, Accuracy, Blood Glucose Levels

10.37802/joti.v8i1.1195

Journal of Technology and Informatics (JoTI)
P-ISSN 2721-4842
E-ISSN 2686-6102
https://e-journals.dinamika.ac.id/index.php/joti

***Abstract*:**

Early detection of diabetes mellitus is crucial to prevent severe complications. This study evaluates three machine learning algorithms for diabetes prediction using a quantitative comparative experimental design. The algorithms are k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), and Random Forest. These methods were chosen to compare distinct learning paradigms. k-NN is distance-based, SVM is margin-based, and Random Forest is an ensemble method. The goal is to find the optimal model for clinical use. The Pima Indians Diabetes dataset was used. It includes 390 patients and 15 clinical features. Performance was measured by accuracy, precision, recall, and F1-score. Random Forest had the highest accuracy (89.7%) and F1-score, providing the most balanced classification. SVM followed with 84.6%, and k-NN achieved 76.9%. Although k-NN had the highest recall (0.750), its precision was low (0.375), showing a high false-positive rate. Feature importance analysis pointed to blood glucose levels as the most significant predictor, which matches clinical knowledge. In summary, ensemble techniques like Random Forest offer the most reliable results. This highlights the importance of selecting the right algorithm for early diabetes detection in clinical applications.

INTRODUCTION

Diabetes mellitus is increasingly recognized as a chronic metabolic disorder that poses significant global health challenges. It is associated with severe complications, including

cardiovascular diseases, kidney failure, and neuropathy. The rising prevalence of diabetes underscores the urgent need for effective management strategies. Early detection and accurate diagnosis are critical to preventing progression and complications [1]. Machine learning (ML) technologies have emerged as valuable tools, providing more rapid and objective diagnoses compared to traditional methods [2]. Several ML algorithms—particularly k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Random Forest—have shown promising results in predicting diabetes onset [3], [4], [5], [6].

The comparative efficacy of these machine learning techniques varies based on the datasets, methodologies, and features used. For example, studies have shown that k-NN achieved about 80% accuracy. SVM and Random Forest reported accuracies of around 72% and 75%, respectively [3], [5]. There is a research gap in the comprehensive comparison between these algorithms and in understanding how model parameter selection affects outcomes. This study aims to bridge that gap by directly comparing the three ML algorithms in predicting diabetes using the Pima Indians Diabetes dataset [7].

Moreover, understanding the clinical features that drive prediction accuracy is essential for enhancing model interpretability and reliability. Research shows that features such as plasma glucose concentration, serum insulin resistance, and blood pressure are critical to diabetes prediction [3], [5]. Isolating these features yields valuable insight into underlying risk factors, supporting the development of targeted prevention strategies. Accordingly, this study implements and compares the efficacy of k-NN, SVM, and Random Forest algorithms in predicting diabetes, using standard evaluation metrics—accuracy, precision, recall, and F1-score—to identify the best-performing model.

The scientific contribution of this study is to provide a clearer evaluation of which ML algorithm is most effective for diabetes prediction and to identify the clinical features that significantly influence the predictive power of these algorithms. Practically, the findings are expected to serve as a foundation for developing more effective ML-based decision support systems in clinical settings for early diabetes detection, which can improve diagnostic quality and patient management. This research will strengthen the integration of ML technologies in clinical decision-making, with the goal of enhancing the accuracy and timeliness of diabetes diagnoses [8], [9].

METHOD

In our analysis of machine learning algorithms for diabetes prediction, we highlight three significant methods: k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), and Random Forest. The k-NN algorithm is favored for its simplicity and effectiveness in managing non-linear data without stringent assumptions about data distribution, making it suitable for complex medical datasets [10]. SVM excels in handling high-dimensional data, effectively discerning optimal separation boundaries between classes, which is critical in medical applications where data features are often numerous, as supported by various studies on its application in diabetes classification [11], [12]. Additionally, Random Forest offers superior performance in dealing with imbalanced datasets and feature-rich environments, greatly reducing the risk of overfitting through its ensemble learning approach [4][5]. The choice of these algorithms is substantiated by their empirical success in medical classification problems, particularly in diabetes-related studies [13], [14], [15].

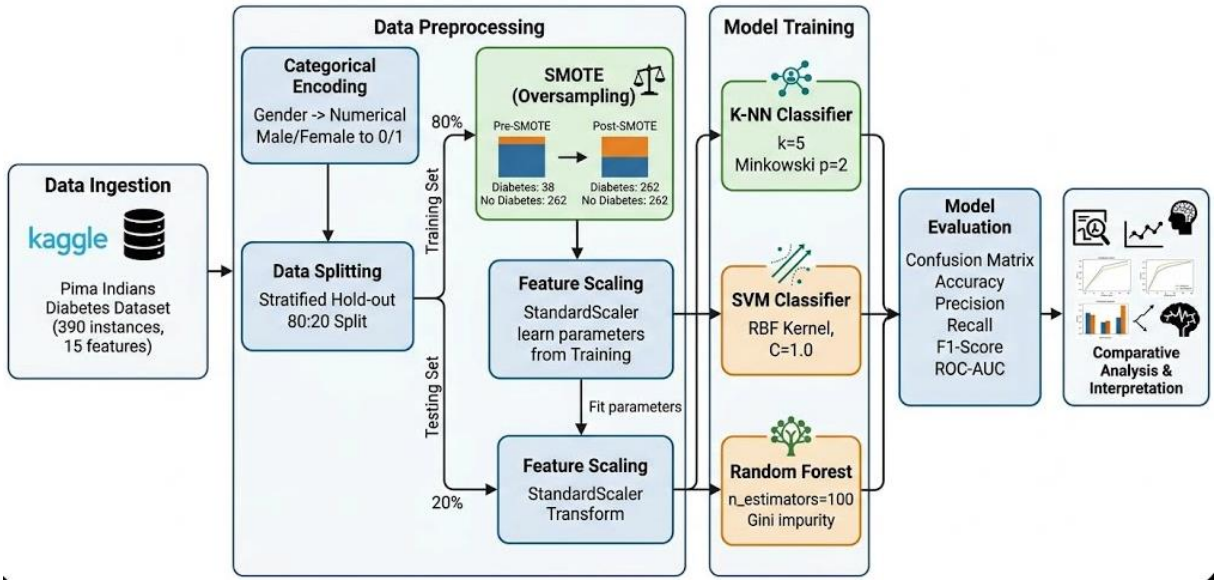

Figure 1. Research Methodology (Source: Author)

In our analysis of machine learning algorithms for diabetes prediction, we highlight three significant methods: k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), and Random Forest. To begin with, the k-NN algorithm is favored for its simplicity and effectiveness in managing non-linear data without stringent assumptions about data distribution, making it suitable for complex medical datasets [10]. Building upon this, SVM excels in handling high-dimensional data, effectively discerning optimal separation boundaries between classes. This capability is especially critical in medical applications where data features are often numerous, as supported by various studies on its application in diabetes classification [11], [12]. Furthermore, Random Forest offers superior performance in dealing with imbalanced datasets and feature-rich environments, greatly reducing the risk of overfitting through its ensemble learning approach [4], [5]. Ultimately, the choice of these algorithms is substantiated by their empirical success in medical classification problems, particularly in diabetes-related studies [13], [14], [15].

Dataset

This study uses the public Pima Indians Diabetes Database from Kaggle, which contains data from 390 patients and 15 clinical features relevant to predicting diabetes, such as cholesterol, glucose levels, age, BMI, and a target column for diabetes status. This widely used dataset is representative for diabetes prediction research. The study will test and compare the effectiveness of three popular classification algorithms in predicting diabetes.

Tabel 1. Patient Data from the Diabetes Dataset
(https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)

patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height	weight	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
1	193	77	49	3.9	19	female	61	119	22.5	118	70	32	38	0.84	No diabetes
2	146	79	41	3.6	19	female	60	135	26.4	108	58	33	40	0.83	No diabetes
3	217	75	54	4	20	female	67	187	29.3	110	72	40	45	0.89	No diabetes
4	226	97	70	3.2	20	female	64	114	19.6	122	64	31	39	0.79	No diabetes

patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height	weight	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
5	164	91	67	2.4	20	female	70	141	20.2	122	86	32	39	0.82	No diabetes
6	170	69	64	2.7	20	female	64	161	27.6	108	70	37	40	0.93	No diabetes
7	149	77	49	3	20	female	62	115	21	105	82	31	37	0.84	No diabetes
8	164	71	63	2.6	20	male	72	145	19.7	108	78	29	36	0.81	No diabetes
9	230	112	64	3.6	20	male	67	159	24.9	100	90	31	39	0.79	No diabetes
10	179	105	60	3	20	female	58	170	35.5	140	100	34	46	0.74	No diabetes

Data Partitioning and Validation Strategy

To objectively evaluate the predictive performance of the machine learning algorithms, the dataset was partitioned using a stratified hold-out validation strategy [16]. The data was divided into a training set and an independent testing set with an 80:20 ratio, allocating 80% of the instances for model training and the remaining 20% for testing. To mitigate the risk of sampling bias and ensure that the original class distribution of the target variable, diabetic versus non-diabetic cases, was strictly preserved across both subsets, a stratified sampling technique was employed during the splitting process. Furthermore, a fixed random seed *(random_state=42)* was initialized to guarantee the reproducibility of the data partition. It is important to note that the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the 80% training subset to prevent data leakage, ensuring that the 20% testing subset remained completely unseen and representative of real-world class distributions.

Data Preprocessing

Several preprocessing steps were conducted prior to model development. First, the categorical feature "gender" was converted into a numerical representation. To address class imbalance in the target variable, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training dataset. Initially, the minority class "Diabetes" consisted of 38 instances, while the majority class "No Diabetes" comprised 262 instances. SMOTE generated synthetic samples for the minority class to balance it with the majority. Lastly, all numerical features were standardized using StandardScaler to ensure uniform feature scales.

The SMOTE algorithm operates by generating synthetic samples for the minority class through linear interpolation between existing instances. Mathematically, for each sample x_i in the minority class, SMOTE selects k nearest neighbors (default k=5) and creates a new sample x_{nev} using the formula (1):

$$x_{nev} = x_i + \lambda \times (x_{in} - x_i) \quad (1)$$

where:
x_{in} represents one of the k-nearest neighbors of x_i
λ denotes a random value between 0 and 1

Table 2. Class Distribution Before and After SMOTE Application

Condition	Non-Diabetes Class	Diabetes Class
Pre-SMOTE	262	38
Post-SMOTE	262	262

This process iterates until the minority class sample count matches the majority class. In the current implementation, the parameter random_state=42 ensures result reproducibility. The increase in diabetic class samples from 38 to 262 demonstrates SMOTE's effectiveness in addressing class imbalance while preserving crucial information from original samples.

k-Nearest Neighbors (k-NN)

The k-Nearest Neighbors (k-NN) algorithm was implemented using the *KNeighborsClassifier* from the Scikit-learn library. The model was configured with default hyperparameters, setting k=5 to evaluate five nearest neighbors. Patient data point distances were computed using the Minkowski metric with p=2, equivalent to the Euclidean distance [7]. Uniform weighting ensured each point in the local neighborhood contributed equally to the majority vote [17].

Support Vector Machine (SVM)

The Support Vector Machine (SVM) model was specifically configured using the Radial Basis Function (RBF) kernel *(kernel='rbf'). This kernel maps* the clinical features into a higher-dimensional space and helps determine the optimal separating hyperplane [18]. The regularization parameter (C), which dictates the penalty for misclassification, was set to the standard value of 1.0. The kernel coefficient (Y) was configured to auto-scale. Notably, the *probability=True* parameter was enabled to allow the model to output class probability estimates for subsequent advanced analysis [19], [20]. Furthermore, the *random_state=42* parameter was defined to guarantee the computational stability and reproducibility of the generated support vectors [21].

Random Forest

As an ensemble learning method, the Random Forest algorithm was constructed by initializing 100 decision trees *(n_estimators=100)* [22]. This hyperparameter selection provides an optimal balance between classification performance and computational efficiency. The split quality at each node was evaluated using the Gini impurity criterion. The random_state=42 parameter was explicitly set to maintain the consistency of the bootstrap aggregating (bagging) process and the random feature sub-sampling at each tree split. The final predictive outcome regarding a patient's diabetes status was derived through a majority voting mechanism. This mechanism aggregated the predictions of all 100 constituent trees [23], [24].

RESULTS AND DISCUSSION

In this study, machine learning models were developed and evaluated using both Google Collaboratory (Google Collab) and a MacBook Air M1 as the local development environment. Google Collab served as a cloud-based platform for model execution. The MacBook Air M1 was used for initial data preprocessing, exploratory analysis, and local debugging. Google Collab was chosen for its support of cloud-based Python scripting and access to integrated GPUs in its free tier. These features facilitate faster computation. Several essential libraries were used in this experiment. Scikit-learn was used for modeling and evaluation, Pandas and NumPy for data manipulation, Matplotlib and Seaborn for visualization, and Imbalanced-learn for handling class imbalance using SMOTE. The training process was performed on a MacBook Air M1 (2020), which has an 8-core ARM architecture processor and 8GB of RAM. This device includes built-in CPU optimizations for efficient Python execution.

Model training used three machine learning algorithms: k-NN, SVM with an RBF kernel, and Random Forest Classifier. Each model was trained with SMOTE-balanced data and tested on normalized data using StandardScaler. Evaluation included accuracy, precision, recall, F1-score, and confusion matrix analysis. This process was repeated for each model to ensure objective performance comparison.

Evaluation of Classification Model Performance

This study evaluated the performance of three widely used machine learning algorithms, k-NN, SVM, and RF, in classifying diabetes outcomes. The evaluation was based on four key performance metrics: accuracy, precision, recall, and F1-score. These metrics were chosen for their ability to reflect not only the overall correctness of the model but also its effectiveness in handling class imbalance, which is common in medical datasets. The results are summarized in Table 1. From the results, the Random Forest model achieved the highest accuracy at 89.7%, indicating superior overall performance in correctly classifying both diabetic and non-diabetic cases. Although the k-NN model attained the highest recall (0.750), it suffered from low precision (0.375), suggesting a high rate of false positives. This outcome highlights a common trade-off in medical classification tasks, where maximizing recall can come at the expense of precision. On the other hand, the SVM model showed more balanced results, with moderate precision and recall, though still lower in overall performance compared to Random Forest.

The F1-score, which balances precision and recall, provides a more holistic view of model performance, especially in imbalanced datasets. Random Forest outperformed the others with the highest F1-score (0.636), making it the most reliable model in this comparative analysis for identifying diabetes cases accurately and consistently. The superior performance of Random Forest in this specific clinical context can be attributed to several factors inherent to the dataset. Medical datasets, such as the Pima Indians Diabetes database, often contain non-linear relationships and complex interactions between physiological features (the interaction between BMI, age, and glucose levels). Unlike k-NN, which relies on distance metrics and is highly sensitive to noisy data, or SVM, which seeks a single optimal geometric margin, RF builds multiple uncorrelated decision trees through feature bagging (Random Subspace Method). This ensemble mechanism allows the algorithm to inherently filter out 'noise' from less relevant features, such as gender or minor blood pressure, variations and place stronger emphasis on critical diagnostic signals like blood glucose, as corroborated by the feature importance analysis. Furthermore, tree-based models are naturally robust to the varied numerical distributions often found in clinical markers, allowing the Random Forest to construct a highly accurate and generalizable decision boundary without overfitting the training data.

Confusion Matrix Analysis

To gain deeper insight into the behavior of each model, confusion matrices were analyzed. These matrices provide detailed information on the distribution of true positives, true negatives, false positives, and false negatives, which are critical for evaluating classification models in healthcare applications.

Table 1. Performance comparison of classification models

Model	Accuracy	Precision	Recall	F1-Score
k-Nearest Neighbors	0.769	0.375	0.750	0.500
Support Vector Machine	0.846	0.500	0.583	0.538
Random Forest	**0.897**	**0.700**	**0.583**	**0.636**

k-Nearest Neighbors

The k-Nearest Neighbors confusion matrix (Figure 2) showed 51 non-diabetic and 9 diabetic cases correctly predicted. However, the model missed 15 diabetic cases (false negatives). This is a significant medical concern, as undiagnosed diabetes can cause serious complications. Compared to k-NN, the SVM model (Figure 3) correctly identified 59 non-diabetic cases, misclassified 7, and detected 7 of 12 diabetic cases. This shows improved precision, though false negatives remain notable. Building on these results, the Random Forest model (Figure 4) demonstrated the most robust performance, with 63 true negatives and 7 true positives, and only 3 false positives and 5 false negatives. This indicates not only high accuracy but also a balanced ability to identify both classes, minimizing the risks associated with misclassification.

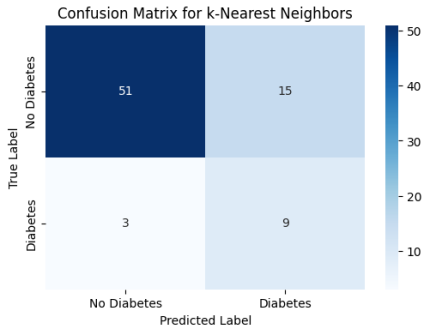

Figure 2. Confusion Matrix of k-Nearest Neighbors

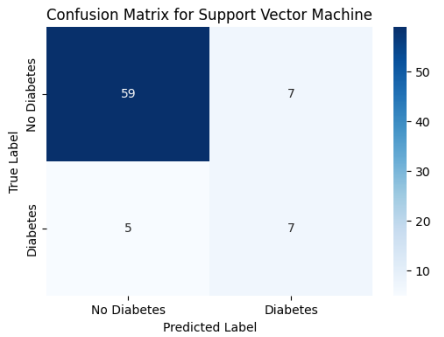

Figure 3. Confusion Matrix of SVM

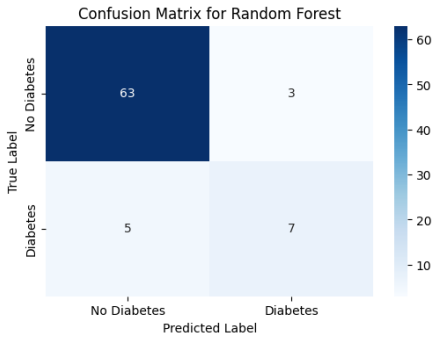

Figure 4. Confusion Matrix Matrix of Random Forest

The confusion matrices confirm earlier performance metrics. They reinforce that Random Forest best balances sensitivity and specificity among the models.

ROC Curve and AUC Analysis

A ROC curve analysis was conducted to comprehensively evaluate model discrimination at different thresholds. The ROC curve plots True Positive Rate (Sensitivity) against False Positive Rate (1 - Specificity), showing the trade-off between correctly identifying diabetic patients and misclassifying non-diabetic ones. The Area Under the Curve (AUC) was calculated to measure overall model performance. An AUC closer to 1.0 shows a better ability to distinguish between classes. As illustrated in Figure 5, the Random Forest model exhibited the highest predictive performance, achieving an outstanding AUC score of 0.934. This indicates a robust and highly reliable capacity for class separation. The Support Vector Machine (SVM) model also displayed strong discriminatory power, yielding an AUC of 0.890. In contrast, the k-Nearest Neighbors (k-NN) algorithm recorded the lowest performance among the three, with an AUC of 0.793. As shown in Figure 5, the Random Forest model had the highest predictive performance with an AUC score of 0.934. This score indicates robust and reliable class separation. The Support Vector Machine (SVM) model also performed well, achieving an AUC of 0.890. In comparison, the k-Nearest Neighbors (k-NN) algorithm had the lowest performance, recording an AUC of 0.793. These AUC results match the earlier discussed evaluation metrics. They provide further evidence that Random Forest is the most effective model for diabetes prediction in this study. Random Forest's superior performance is mainly due to its ensemble approach [25]., which combines predictions from many independent decision trees. This allows it to capture complex, non-linear relationships among clinical features and reduces overfitting, a common problem with high-dimensional medical data. The SVM model performs well by using the RBF kernel to map data and find optimal separation boundaries. However, it lacks the adaptability of Random Forest's feature sub-sampling. In contrast, k-NN's lower AUC and precision reflect its sensitivity as a distance-based classifier. This makes it more affected by noisy data and overlapping classes in complex clinical settings.

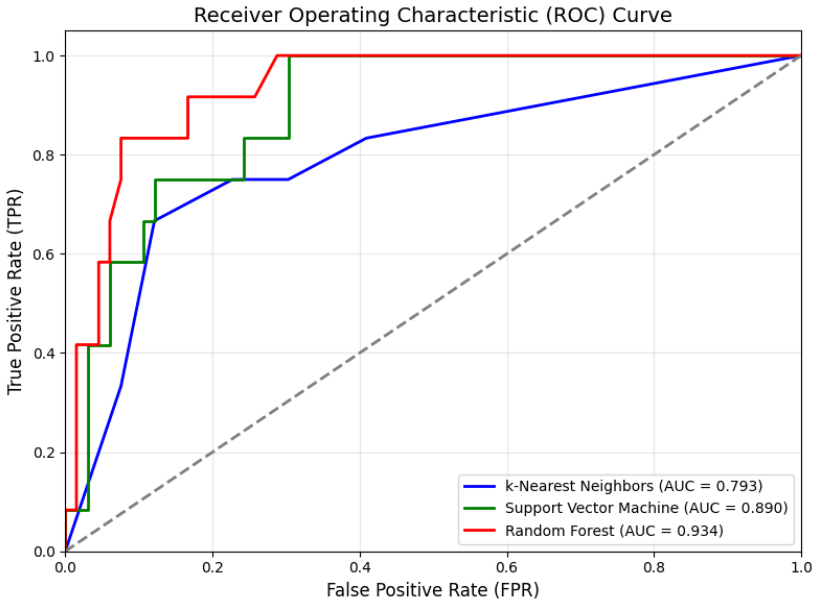

Figure 5. Receiver Operating Characteristic (ROC) curves comparing the predictive performance of k-Nearest Neighbors, Support Vector Machine, and Random Forest models.

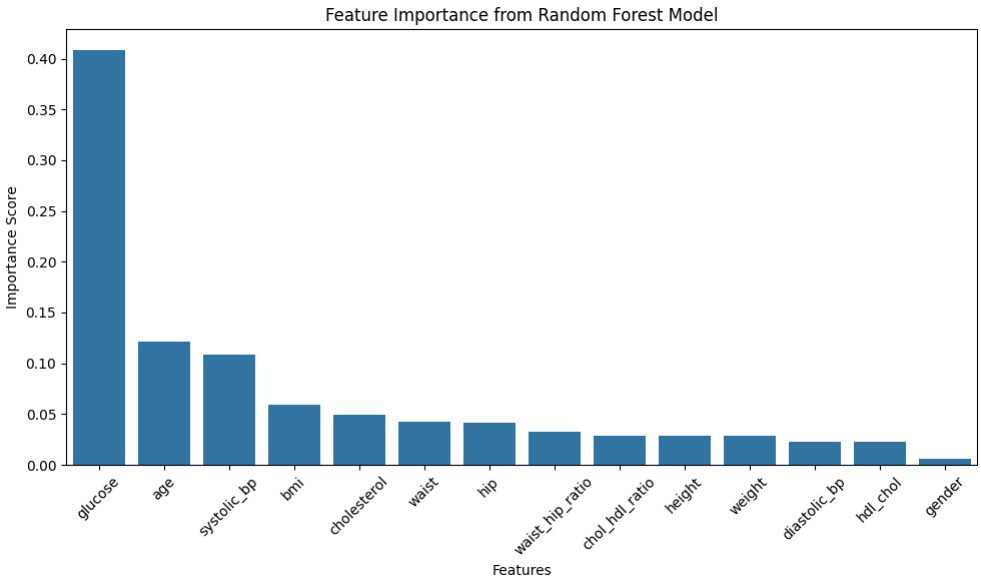

Figure 6. Feature importance rankings based on the Random Forest model

Feature Importance Analysis

To identify key features influencing classification decisions, a feature importance analysis was performed using the Random Forest model. This approach quantifies each input variable's contribution to the predictive outcome, clarifying which factors are most instrumental in diabetes detection. Figure 6 indicates that blood glucose level is the decisive feature, accounting for over 40% of the model's decisions. This result aligns with clinical knowledge, as high glucose levels typically signal diabetes. Other influential features include age and systolic blood pressure, both established risk factors. Conversely, gender exerted little influence on predictions, suggesting its limited role in diabetes classification for this dataset. This information can guide future feature selection, enabling model simplification without sacrificing accuracy.

CONCLUSIONS AND SUGGESTIONS

This study successfully evaluated and compared the performance of machine learning algorithms for diabetes prediction by following a structured methodological pipeline. Initially, data preprocessing was conducted, which included categorical encoding and feature standardization. To address the inherent class imbalance within the dataset, SMOTE was applied. Subsequently, three classification models, k-NN, SVM, and RF, were trained and evaluated using standard classification metrics.

The Random Forest model achieved the highest accuracy at 89.7%. This result indicates superior overall performance in correctly classifying both diabetic and non-diabetic cases. Although the k-NN model attained the highest recall (0.750), it suffered from low precision (0.375). This suggests a high rate of false positives. The outcome highlights a common trade-off in medical classification tasks, where maximizing recall can come at the expense of precision. In contrast, the SVM model showed more balanced results, with moderate precision and recall. However, it still had lower overall performance compared to Random Forest.

The F1-score, which balances precision and recall, provides a more holistic view of model performance, especially in imbalanced datasets. Random Forest outperformed the others with the highest F1-score (0.636), making it the most reliable model in this comparative analysis. As discussed, the superior performance of Random Forest in this specific clinical context is attributed to its ensemble mechanism (feature bagging), which effectively filters out noise from less relevant features and handles the non-linear relationships and complex interactions inherent in medical datasets like the Pima Indians Diabetes database without overfitting. Future studies could explore systematic hyperparameter tuning to further optimize these predictive models.

REFERENCES

[1] R. D. Joshi and C. K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *Int. J. Environ. Res. Public Health*, vol. 18, no. 14, p. 7346, Jul. 2021, doi: 10.3390/ijerph18147346.

[2] M. Guhdar, A. Ismail Melhum, and A. Luqman Ibrahim, "Optimizing Accuracy of Stroke Prediction Using Logistic Regression," *Journal of Technology and Informatics (JoTI)*, vol. 4, no. 2, pp. 41–47, Jan. 2023, doi: 10.37802/joti.v4i2.278.

[3] F. Yunita Sari, M. S. Kuntari, H. Khaulasari, and W. Ari Yati, "Comparison of Support Vector Machine Performance with Oversampling and Outlier Handling in Diabetic Disease Detection Classification," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 3, pp. 539–552, Jul. 2023, doi: 10.30812/matrik.v22i3.2979.

[4] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, "Predicting the Onset of Diabetes with Machine Learning Methods," *J. Pers. Med.*, vol. 13, no. 3, p. 406, Feb. 2023, doi: 10.3390/jpm13030406.

[5] M. KIVRAK, "Early Diagnosis of Diabetes Mellitus by Machine Learning Methods According to Plasma Glucose Concentration, Serum Insulin Resistance and Diastolic Blood Pressure Indicators," *Medical Records*, vol. 4, no. 2, pp. 191–5, May 2022, doi: 10.37990/medr.1021148.

[6] I. G. A. Gunadi and D. O. Rachmawati, "A Comparative Study on the Impact of Feature Selection and Dataset Resampling on the Performance of the K-Nearest Neighbors (KNN) Classification Algorithm," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 13, no. 2, pp. 419–427, Jul. 2024, doi: 10.23887/janapati.v13i2.82174.

[7] P. V. S. Kumar and N. S. Kumar, "Analysis and comparison for prediction of Diabetic Pregnant women using Innovative Principal Component Analysis algorithm over Support Vector Machine Algorithm with Improved Accuracy," ., no. 25, pp. 942–948, Feb. 2023, doi: 10.18137/cardiometry.2022.25.942948.

[8] L. P. Nguyen *et al.*, "The Utilization of Machine Learning Algorithms for Assisting Physicians in the Diagnosis of Diabetes," *Diagnostics*, vol. 13, no. 12, p. 2087, Jun. 2023, doi: 10.3390/diagnostics13122087.

[9] S. Qin, "Apply multiple machine learning models to diabetes prediction," *Applied and Computational Engineering*, vol. 86, no. 1, pp. 240–249, Jul. 2024, doi: 10.54254/2755-2721/86/20241610.

[10] R. K. Dewi and S. K. Wardhani, "Prediction of Women's Potential Type 2 Diabetes with Similarity Classifier Based on P-Probabilistic Extension," *Journal of Information Technology and Cyber Security*, vol. 1, no. 2, pp. 76–84, Dec. 2023, doi: 10.30996/jitcs.9945.

[11] A. Syahri, U. Fariha, R. Afandi, and I. Nurliyana, "Comparison of Logistic Regression, Random Forest and Adaboost Algorithms for Diabetes Mellitus Classification," *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, vol. 1, no. 1, pp. 41–46, May 2024, doi: 10.57152/ijatis.v1i1.1116.

[12] Suresh Reddya M and Ramakrishnan V, "Diabetes Prediction Using Blood Sample Data with Novel Voting Classifier over Random Forest," 2022. doi: 10.3233/APC220045.

[13] M. Suda, T. Ooka, and Z. Yamagata, "Prediction and predictor elucidation of metabolic syndrome onset among young workers using machine learning techniques: A nationwide study in Japan," *Environmental and Occupational Health Practice*, vol. 4, no. 1, pp. 2021-0023-OA, 2022, doi: 10.1539/eohp.2021-0023-OA.

[14] S. Wang, "Diabetes Prediction Using Random Forest in Healthcare," *Highlights in Science, Engineering and Technology*, vol. 92, pp. 210–217, Apr. 2024, doi: 10.54097/5ndh9a05.

[15] P. Saha *et al.*, "Predicting Time to Diabetes Diagnosis Using Random Survival Forests," Feb. 07, 2024. doi: 10.1101/2024.02.03.24302304.

[16] A. Edet, S. Inyang, I. Umoren, and U. E. Etuk, "Machine Learning Approach for Classification of Cyber Threats Actors in Web Region," *Journal of Technology and Informatics (JoTI)*, vol. 6, no. 1, pp. 70–77, Oct. 2024, doi: 10.37802/joti.v6i1.679.

[17] W. A. Arifin, I. Ariawan, A. A. Rosalia, L. Lukman, and N. Tufailah, "Data scaling performance on various machine learning algorithms to identify abalone sex," *Jurnal Teknologi dan Sistem Komputer*, vol. 10, no. 1, pp. 26–31, Jan. 2022, doi: 10.14710/jtsiskom.2021.14105.

[18] R. Sumiati, Moh. Chamim, D. Leni, Y. Rosa, and H. Hanif, "Modeling Mechanical Component Classification Using Support Vector Machine with A Radial Basis Function Kernel," *Jurnal Teknik Mesin*, vol. 16, no. 2, pp. 165–174, Dec. 2023, doi: 10.30630/jtm.16.2.1250.

[19] I. Rehan, S. Khan, and R. Ullah, "Raman spectroscopy assisted support vector machine: a steadfast tool for noninvasive classification of urinary glucose of diabetes mellitus," *Phys. Scr.*, vol. 99, no. 2, p. 026004, Feb. 2024, doi: 10.1088/1402-4896/ad1da8.

[20] F. R. Lumbanraja, F. Lufiana, Y. Heningtyas, and K. Muludi, "IMPLEMENTASI SUPPORT VECTOR MACHINE (SVM) UNTUK KLASIFIKASI PEDERITA DIABETES MELLITUS," *Jurnal Komputasi*, vol. 10, no. 1, pp. 75–83, Apr. 2022, doi: 10.23960/komputasi.v10i1.2940.

[21] G. Abdurrahman, "Klasifikasi Kanker Payudara Menggunakan Algoritma SVM dengan Kernel RBF, Linier, dan Sigmoid," *JUSTIFY : Jurnal Sistem Informasi Ibrahimy*, vol. 2, no. 1, pp. 74–80, Jul. 2023, doi: 10.35316/justify.v2i1.3370.

[22] S. Wang, "Diabetes Prediction Using Random Forest in Healthcare," *Highlights in Science, Engineering and Technology*, vol. 92, pp. 210–217, Apr. 2024, doi: 10.54097/5ndh9a05.

[23] Alfi Indah Nurrizqi, Erfiani, and Agus Mohamad Soleh, "Comparison of Ensemble Method Performance in Classifying Blood Sugar Levels Output from Non-Invasive Device," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 11, no. 3, pp. 330–336, Jun. 2024, doi: 10.32628/IJSRSET2411322.

[24] X. Fu, Y. Chen, J. Yan, Y. Chen, and F. Xu, "BGRF: A broad granular random forest algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 5, pp. 8103–8117, May 2023, doi: 10.3233/JIFS-223960.

[25] Lukman Arif Sanjani, R. Bimo Mandala Putra, and U. Laili Yuhana, "Exploring the Application of Machine Learning for Automatic Inbound Email Classification in CRM System at XYZ Company," *Journal of Technology and Informatics (JoTI)*, vol. 6, no. 1, pp. 1–7, Oct. 2024, doi: 10.37802/joti.v6i1.715.