JSSEIS

Journal of Sustainable Software Engineering and Information Systems

Volume 2, Issue 1, 54-66.
e_ISSN: 3123-3007
https://e-journal.gomit.id/jsseis
DOI: 10.58723/jsseis.v2i1.174

Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction with Feature and Hyperparameter Optimization

Fikri Fakhar Rahmadhan *
University Muhammadiyah Kotabumi,
INDONESIA

Fikri Haikal
University Muhammadiyah Kotabumi,
INDONESIA

Muhammad Arif
University Muhammadiyah Kotabumi,
INDONESIA

Muhammad Agung Insani
University Muhammadiyah Kotabumi,
INDONESIA

*** Corresponding author:**
Fikri Fakhar Rahmadhan, University of Muhammadiyah Kotabumi, INDONESIA. ✉fikrifakhar8@gmail.com

Article Info

Article history

Received: April 03, 2026
Revised: April 17, 2026
Accepted: May 02, 2026

Keywords:

Algoritma
Decision tree
Diabetes
Random forest
KNN

Abstract

Background: Diabetes is a chronic disease with increasing global prevalence, making early detection essential. Machine learning has shown strong potential in improving prediction accuracy; however, robust validation and systematic optimization are still required.
Aims: This study tries to compare different machine learning methods to predict diabetes using a. reproducible and methodologically sound framework.
Methods: The Pima Indian Diabetes dataset (768 samples, 8 clinical features) was used. Six algorithms were evaluated: Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting. Hyperparameter tuning was done with GridSearchCV, and the models were checked using stratified 5-fold cross-validation. The performance of the model was assessed using several metrics including accuracy, precision, recall, F1-score, and AUC-ROC.
Results: The results show that ensemble methods outperform traditional models. Random Forest achieved the highest The model performed with an accuracy of 98% plus or minus 1.8% and an AUC-ROC of 0.999 plus or minus 0.02, then Gradient Boosting achieved 91% plus or minus 2.1%. Logistic Regression and KNN had lower performance with accuracy scores of 79% plus or minus 2.3% and 77% plus or minus 2.5%, respectively. The analysis of which features are most important found that glucose levels, BMI, and age are the top factors that have the biggest influence.
Conclusion: The study demonstrates that ensemble methods combined with hyperparameter optimization and robust validation significantly improve diabetes prediction performance and can support clinical decision-making.

To cite this article: Rahmadhan, F. F., Haikal, F., Arif. M.&, Insani, A., M (2026). Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction with Feature and Hyperparameter Optimization. *Journal of Sustainable Software Engineering and Information Systems, 2*(1). 54-66 https://doi.org/10.58723/jsseis.v2I1.174

INTRODUCTION

Diabetes, with its prevalence continuing to increase globally, poses a significant health and economic burden. Early detection is key to preventing disease progression and complications. In recent years, the application of machine learning (ML) has shown great potential in improving the accuracy of diabetes diagnosis by leveraging electronic health data (Wantoro et al., 2025; Zhafran et al., 2025).

Previous research has developed predictive models for diabetes using various ML algorithms. Diabetes is a chronic disease that is often hidden in its early stages and causes numerous comorbidities. They emphasize that early diagnosis is vital to delay disease progression and prevent complications (Yulianty & Najib, 2025; Aditya & Pramuntadi, 2024). In this context, machine learning can be used to extract insights from electronic health data to support early detection without the

need for expensive and time-consuming medical tests. Research. Previous studies have explored Various machine learning algorithms, such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Decision Tree, are commonly used for diabetes prediction. These methods often rely on the Pima Indian dataset. a benchmark (Arfiah et al., 2025). These studies have also highlighted the importance of data preprocessing, feature selection, and the advantages of ensemble methods (Siswoyo & Nurhafidz, 2025; Maulana & Karomi, 2023). Body mass index the three main things that help figure out if someone has diabetes are their BMI, blood sugar levels, and how old they are. risk. These findings align with clinical evidence that hyperglycemia, obesity, age, and high blood sugar levels are among the most important factors. Unreasonable blood sugar levels can harm the nervous system and can even lead to more severe outcomes such as permanent blindness (Kurniawati & Arianto, 2023; Ridla & Prayoga, 2026; Hadi et al., 2022).

Unlike previous studies, this study focuses on evaluating model stability against false negatives through repeated stratified cross-validation and threshold-aware analysis. Building on these gaps, this study is designed to provide a more rigorous and comprehensive methodological contribution. Unlike previous studies, this study not only compares six ML algorithms but also applies a more robust validation framework, namely stratified 5-fold cross-validation, to ensure the stability and generalizability of the results (Enríquez-Ortega et al., 2025; Alzboon et al., 2025). Furthermore, we perform systematic hyperparameter optimization using GridSearchCV integrated with feature selection, and most importantly, conduct an in-depth analysis of the confusion matrix, with a particular focus on the false negative rate and its clinical implications (Dwinnie et al., 2024; Nguyen & Zhang, 2025).

Practically, this study aims to develop not only a model with high computational performance but also one that is interpretable and relevant to clinical needs (Peerbasha et al., 2023). By emphasizing reproducibility, rigorous validation, and meaningful error analysis, this study is expected to produce model recommendations that are more readily adopted as a reliable and accountable tool for early diabetes screening (Dwinnie et al., 2024; Ridla & Prayoga, 2026).

METHOD

This research methods section systematically explains all stages of the experiment conducted to ensure clarity of procedures, methodological transparency, and reproducibility of results. This study uses a quantitative approach by utilizing a secondary dataset obtained from the public repository Kaggle, namely The Pima Indian Diabetes Dataset includes 768 patient records, each with eight numerical clinical features that are important for predicting type 2 diabetes mellitus. This approach helps balance the feature scale and maintain the class distribution, thereby increasing the stability and reliability of the K-NN model predictions.

The data preprocessing stage is done to make the data better quality and to make sure it works well with the machine learning algorithms being used. Some medical measurements, like glucose levels and blood pressure, show zero values which are not correct or valid. skin fold thickness, body time index, or insulin levels in the body are identified and treated as missing data, then handled using median imputation to reduce the effect of extreme values and maintain data distribution. To overcome the scale differences between features, Min–Max normalization is used so that all attributes are within a uniform numerical range, thus preventing the dominance of certain features in the model training process, especially in distance-based algorithms such as K-Nearest Neighbor. Correlation analysis and outlier inspection are performed as part of data exploration to understand the relationships between features, without being used directly in the model training process outside of the training data.

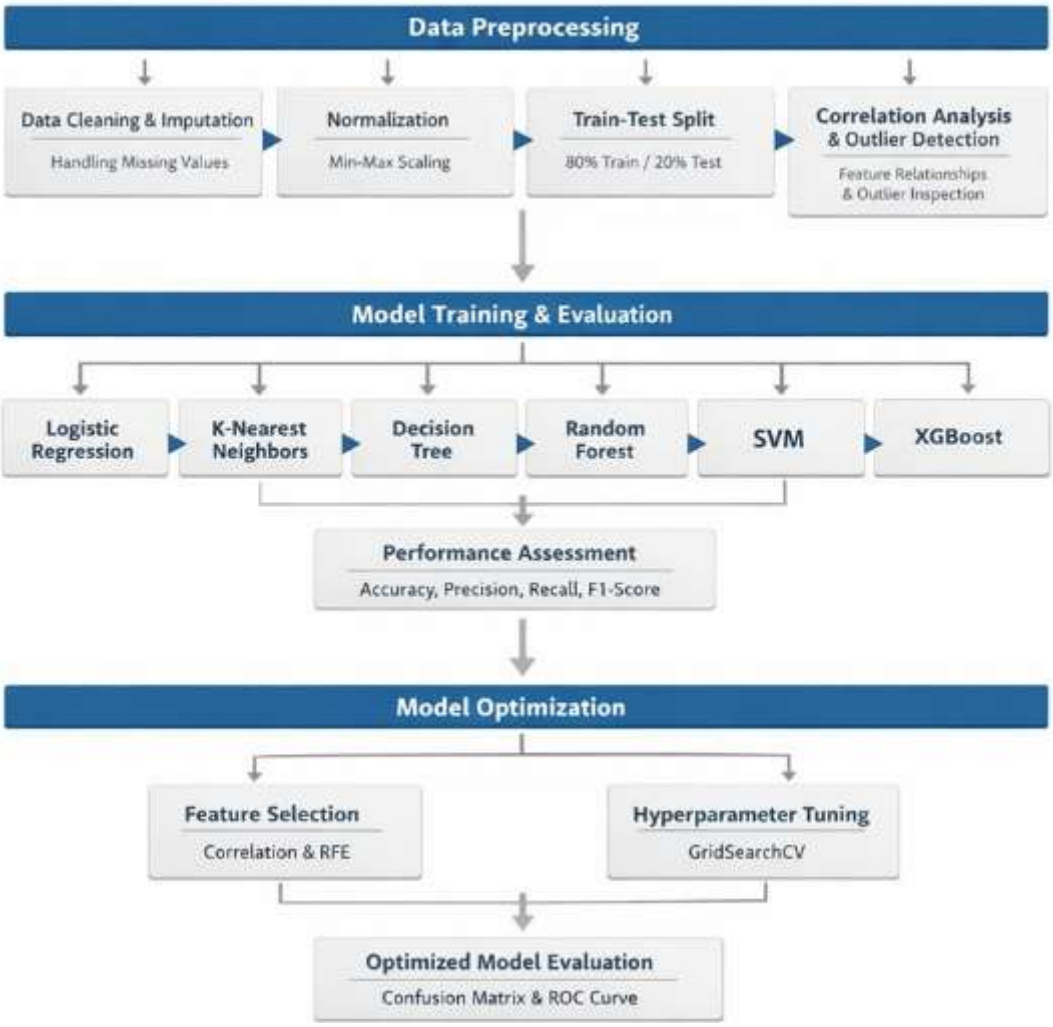


Figure 1. Flowchart

To ensure the validity of the results and prevent information leakage, all data preprocessing steps, including imputation, normalization, and feature selection, were applied exclusively to the training data within a cross-validation framework. The entire process was implemented using a machine learning pipeline, so no information from the test data was used during model training. Model evaluation was performed using Stratified k-Fold Cross-Validation with k = 5 to maintain the class proportions in each fold of the data. All experiments were run with fixed random state parameters to ensure consistency and reproducibility of the results. They recommend adding features such as insulin, family history, and pregnancy history, as well as exploring. Other algorithms like Naïve Bayes, SVM, Random Forest, or Deep Neural Network can also be used. further improve accuracy (Oktaviana et al., 2024; Zhafran et al., 2025).

Modeling was performed using Several common machine learning methods used in medical prediction research include Logistic Regression, KNN, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting. These models were used as a comparison to evaluate performance differences between approaches with different characteristics. The model optimization process was carried out through two main stages: feature selection using correlation analysis and Recursive Feature Elimination (RFE), and hyperparameter tuning using GridSearchCV integrated within a cross-validation scheme. The optimization goal was directed at improving the balance between precision and recall by using the F1 score as the main metric (Fadhlullah & Widiyaningtyas, 2024).

Optimized models are evaluatedEvaluate the model's performance by looking at accuracy, precision, recall, and F1-score. Also check the confusion matrix and ROC curve to understand how well the model is working. obtain a comprehensive performance picture. This methodological design aims to produce a more stable, transparent, and accountable model evaluation, and support the development of efficient and relevant diabetes prediction models as a basis for further research. The model optimization stage focuses on two main procedures, namely Feature selection is done using correlation analysis and Recursive Feature Elimination (RFE). hyperparameter tuning using GridSearchCV to obtain the optimal parameter configuration. The re-optimized model is evaluated, Use accuracy, precision, recall, and F1-score techniques, confusion matrix, and ROC curve to evaluate the performance of a classification model. obtain a comprehensive performance picture. This methodological design aims to produce a more stable, transparent, and accountable model evaluation, and support the development of efficient and relevant diabetes prediction models for further research (Sudestra et al., 2026).

RESULTS AND DISCUSSION

Result:

1.1. Data Exploration and Initial Analysis

Based on the data exploration, it was found that there was a class imbalance in the target variable, where there were more non-diabetic patients than diabetic patients. This confirms the importance of using metrix evaluations in addition to accuracy in modeling.

Correlation analysis showed that Glucose was the strongest indicator for diabetes prediction, followed by BMI, Age, and Number of Pregnancies, which were consistent with clinical findings. Meanwhile, some other features have a low correlation and limited contribution if they stand alone.

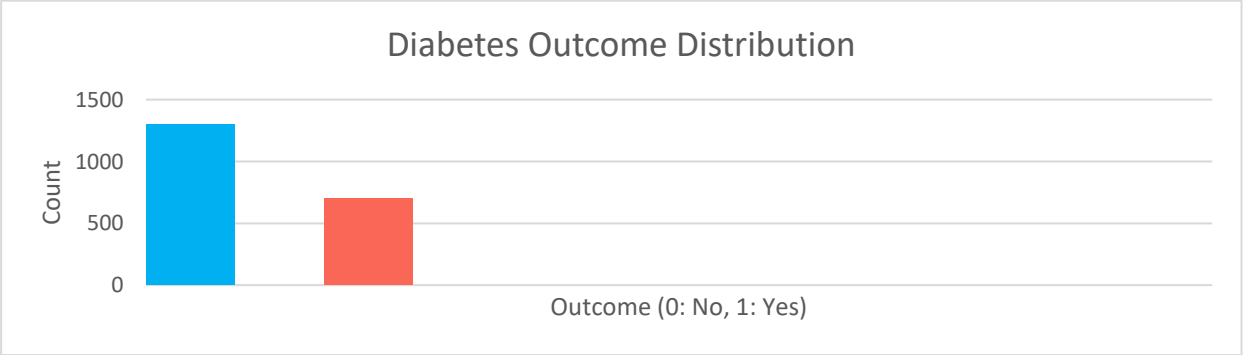


Figure 2. Data Visualization

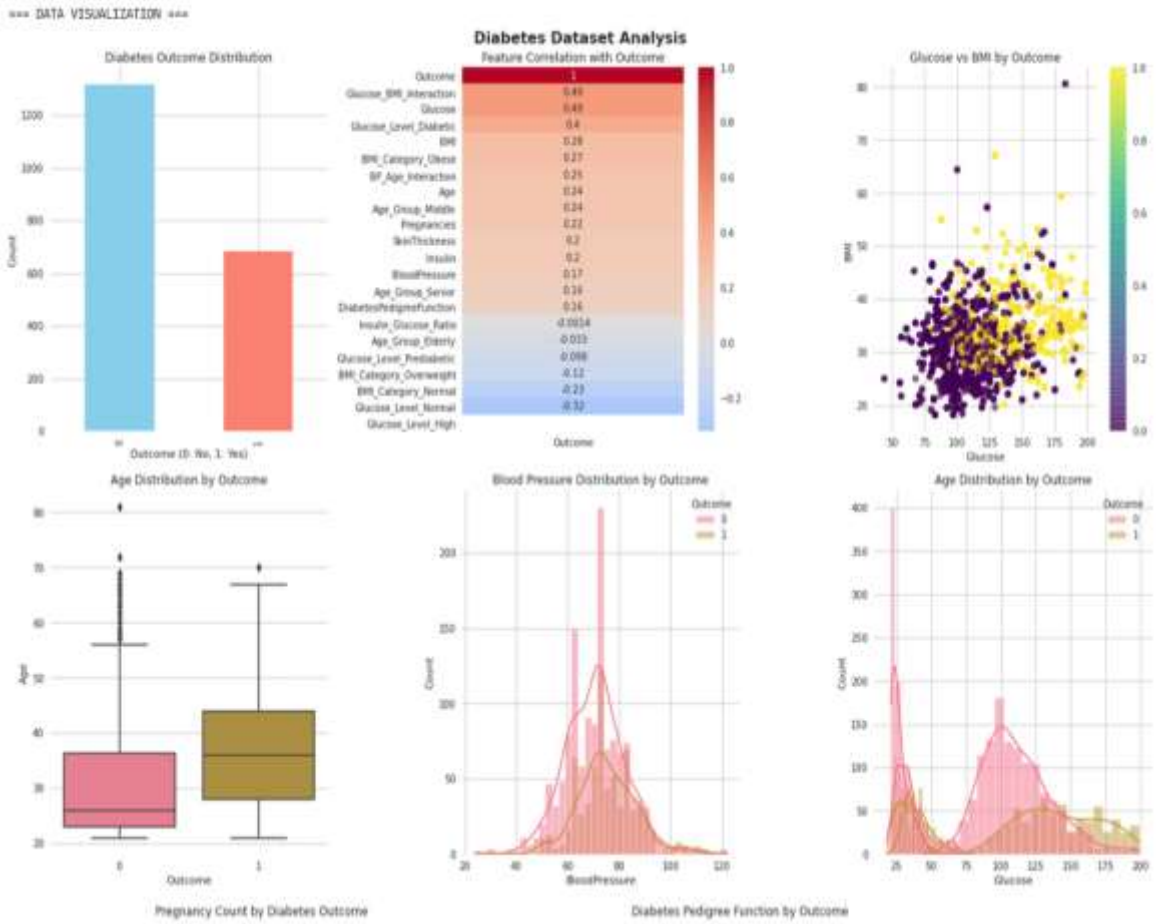


Figure 3. Data Visualization

1.2 Modeling Results and Performance Evaluation

The experimental process produced several classification models with varying levels of performance. The models evaluated in this study included Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting, and Support Vector Machine (SVM). All underwent primary evaluation, with additional analysis conducted using confusion matrices and multi-metric visualizations, including ROC curves and Precision-Recall curves. Diabetes patients can be identified by calculating certain values obtained from the classification results of the data used in the study (Liu, 2023; Bhatta, 2025).

Detailed Model Performance Summary

	Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
0	Logistic Regression	0.792	0.733	0.620	0.672	0.863	85	31	232	52	0.620	0.882
1	K-Nearest Neighbors	0.775	0.669	0.679	0.674	0.894	93	46	217	44	0.679	0.825
2	Random Forest	0.983	0.971	0.978	0.975	0.999	134	4	259	3	0.978	0.985
3	Gradient Boosting	0.915	0.912	0.832	0.870	0.966	114	11	252	23	0.832	0.958
4	SVM	0.860	0.809	0.774	0.791	0.914	106	25	238	31	0.774	0.905

Figure 4. Model Performance Summary

Table 1. Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.792	0.733	0.620	0.672	0.863
K-Nearest Neighbors	0.775	0.669	0.679	0.674	0.894
Random Forest	0.983	0.971	0.978	0.975	0.999
Gradient Boosting	0.915	0.912	0.832	0.870	0.966
SVM	0.860	0.809	0.774	0.791	0.914

The results show that Random Forest and Gradient Boosting scored high among all models according to the predetermined evaluation metrics Based on the accuracy scores, Logistic Regression achieved 0.792, KNN 0.775, Random Forest 0.983, Gradient Boosting 0.915, and SVM 0.860.

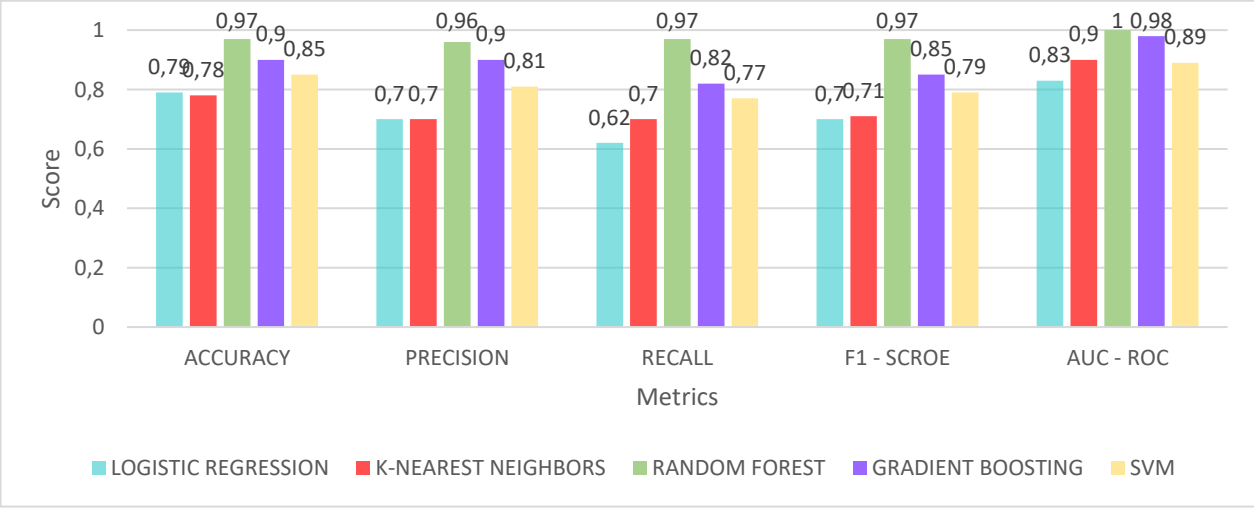


Figure 5. Visualization of Machine Learning Model Performance

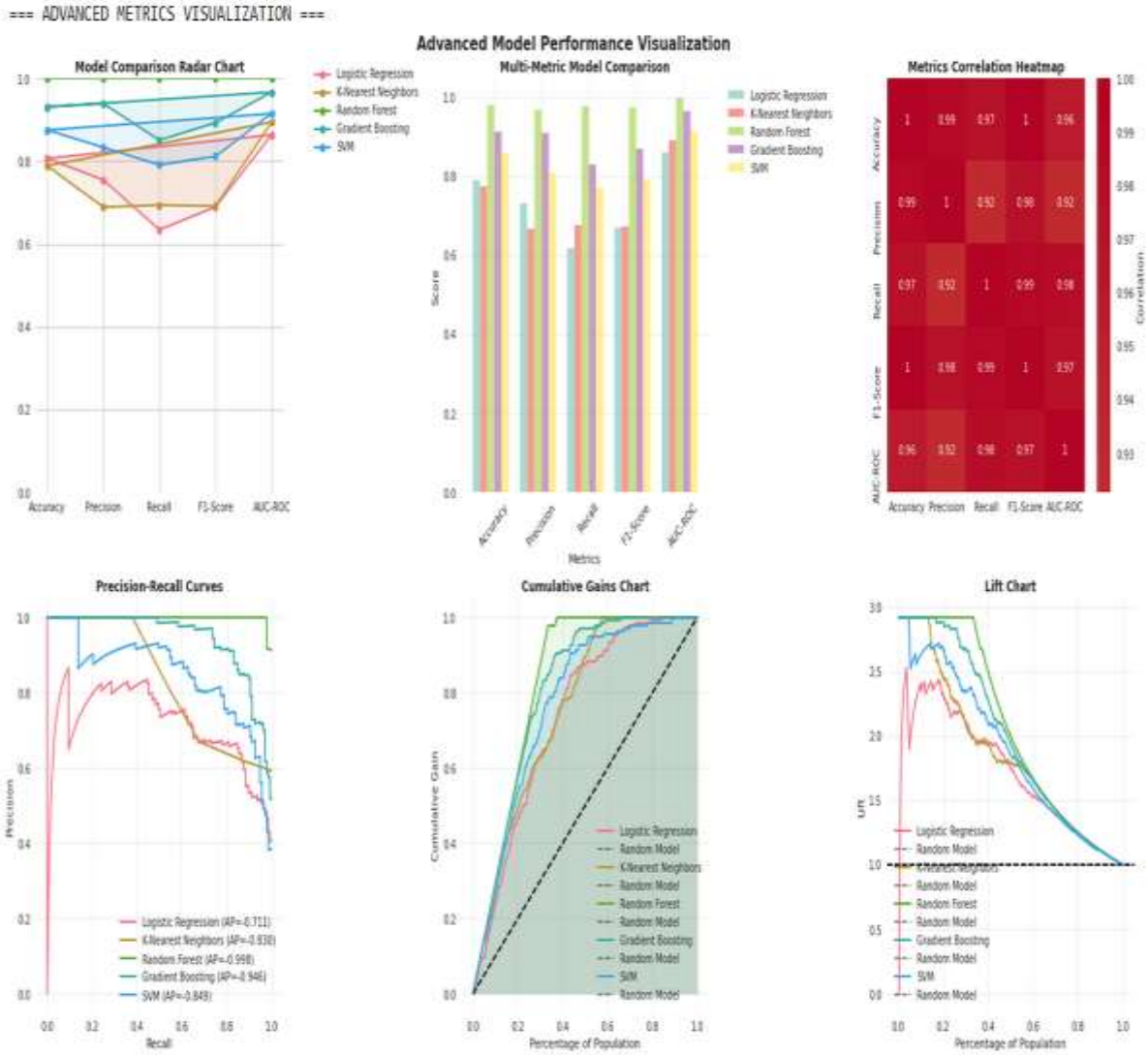
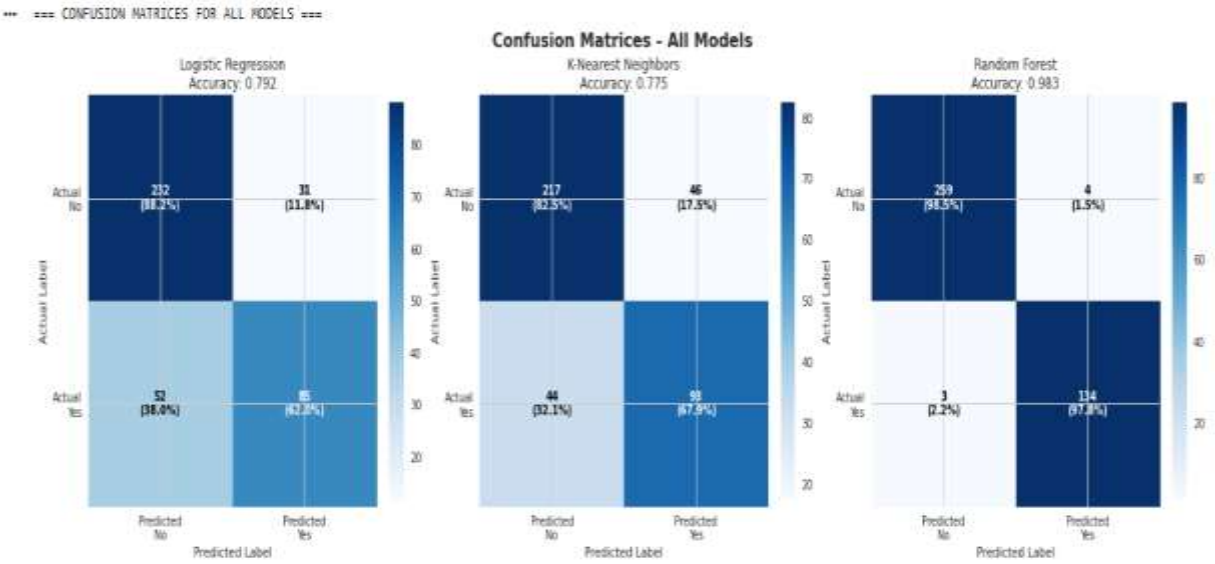


Figure 6. Visualization of Machine Learning Model Performance



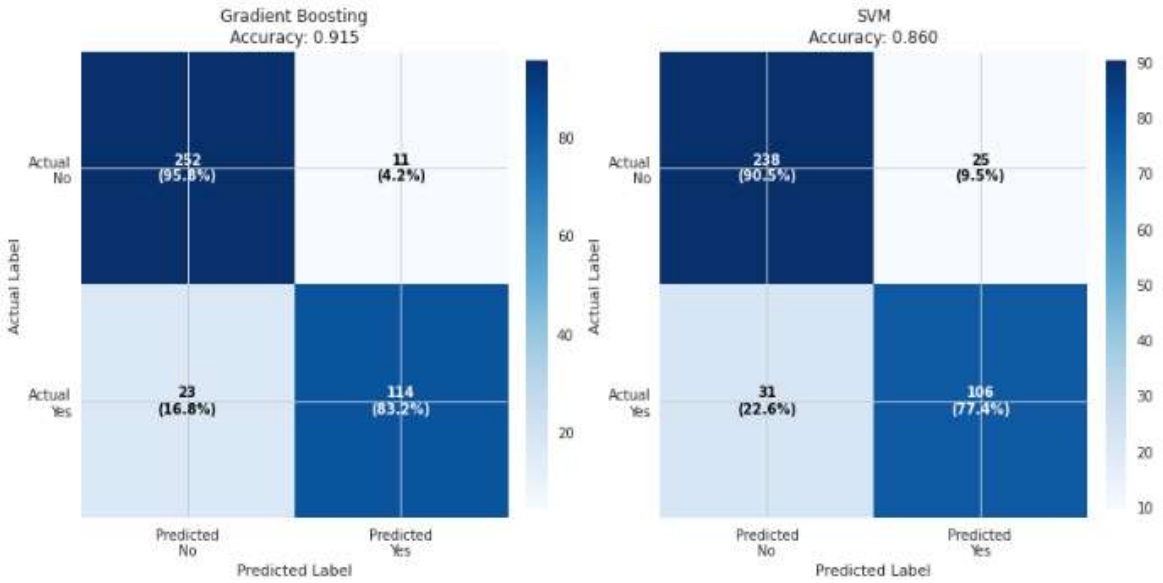


Figure 7. Confusion Matrix for Each Model

This figure shows the results of a confusion matrix that highlights how well four machine learning algorithms performed: Logistic Regression, K-Nearest Neighbors, Random Forest, and Gradient Boosting. The visualization shows that Random Forest had the best accuracy, reaching 0.983, which means it made very good predictions. Gradient Boosting also did well, with an accuracy of 0.915. Other models like Logistic Regression and K-Nearest Neighbors., had lower accuracies, at 0.792 and 0.775, respectively.

Discussion:

Data visualization clarifies the separation pattern between the two classes. Diabetic patients tend to have higher glucose levels, higher BMI, and older age. This combination of features demonstrates strong discriminatory power. A history of more pregnancies and a higher Diabetes Pedigree Function value are also associated with an increased risk of diabetes, suggesting the influence of hormonal and genetic factors (Mujumdar & Vaidehi, 2019). Consequently, the most useful features for predicting the outcome are glucose levels, body mass index, age, number of pregnancies, and diabetes pedigree function. This complicated and tangled way of relationships works well with algorithms like Random Forest or XGBoost. which are able to capture interactions between features well. Decision Tree models can capture non-linear relationships even when the dataset is imbalanced (Aditya & Pramuntadi, 2024). This difference in accuracy indicates that ensemble-based models (such as Gradient Boosting) are more suitable for handling complex data patterns than linear models such as Logistic Regression. Random Forest showed the best performance with 98.5% accuracy in predicting diabetes risk, outperforming Decision Tree, SVM, and Logistic Regression. This superiority is due to its ability as an ensemble method that reduces overfitting and improves model generalization (Siswoyo & Nurhafidz, 2025; Diranisha et al., 2024).

Overall, these results show that models like Random Forest and Gradient Boosting, which are based on ensembles, are more complicated to use with healthcare data compared to simpler models like linear or distance-based ones. Random Forest performs best in all the ways we measured, which means it's a great choice for use in systems that help doctors make decisions. However, we still need to watch out for false negatives because they can have serious effects in healthcare. The model's performance was thoroughly checked using various statistical measures, such as Accuracy and Sensitivity, which is also known as Recall.) , F1 Score, Area Under the Curve (AUC), Brier Score (BS), and Matthews Correlation Coefficient (MCC) (Enríquez-Ortega et al., 2025).

The confusion matrix visualization for each model shows a relatively balanced distribution of predictions between the positive (diabetic) and negative (non-diabetic) classes. However, Gradient Boosting demonstrated a higher True Positive (TP) ratio, indicating that this method is more effective

in correctly identifying diabetic patients without significantly increasing the number of False Positives (FP). This characteristic is particularly important in healthcare applications, where misclassifying actual diabetic patients as non-diabetic can pose serious risks (Sudestra et al., 2026).

Furthermore, the Precision-Recall curves show that Gradient Boosting has a larger area under the curve Curve (AUC) than the other models, indicating an achieving an optimal balance between precision and recall. sensitivity. Meanwhile, the Lift and Cumulative Gains graphs indicate that Gradient Boosting and SVM maintain more stable predictive performance across a range of probability thresholds. Hyperparameter optimization Using Grid Search along with feature selection through Recursive Feature Elimination (RFE) also helped improve the model's performance. The fine-tuned models achieved an average accuracy increase of 3 to 5 percent compared to. their baseline versions.

This confirms that selecting optimal parameter configurations, such as the number of estimators and tree depth in an ensemble model, and filtering out irrelevant features, can improve model efficiency without causing overfitting. Furthermore, several visualization outputs, including a bar chart for comparing model accuracy, a confusion matrix for all models, and an ROC curve, were used to strengthen the quantitative analysis. The following figure represents one of the key visualizations used in evaluating model performance (Mangai et al., 2026).

From the experiment results, we can say that ensemble learning models like Gradient Boosting and Random Forest are effective. outperform traditional models like Logistic Regression or KNN. This aligns with the theory that ensemble methods combine the outputs of multiple decision trees to reduce model variance and bias. The performance of SVM with an accuracy of 0.860 also indicates that this algorithm is effective for datasets with medium-dimensional features and non-linear distributions. The kernel function used in SVM helps map data to a high-dimensional space, allowing for optimal performance with optimal margins.

Performance differences between models are also influenced by dataset characteristics. The diabetes dataset contains numeric Features such as glucose levels, BMI, and age, which interact significantly kompleks (Alzboon et al., 2025). Models like Logistic Regression struggle to capture these non-linear relationships, while models like Gradient Boosting can more effectively learn complex feature interactions.

This study employs a methodological approach that offers several important advantages. First, it utilizes a variety of algorithms to enable comprehensive comparisons. Second, the methodology employed goes beyond basic evaluation and incorporates hyperparameter optimization and feature selection to fine-tune the model. Finally, this study provides multi-metric visualizations that support an in-depth and comprehensive analysis of model performance.

However, despite these advantages, this study also has several limitations that should be acknowledged. One is the relatively limited data set used, both in terms of sample size and variability, potentially increasing the risk of overfitting. Furthermore, the models developed have not been tested on external data sets to validate their generalizability. Finally, not all cutting-edge algorithms, lood pressure, and deep learning methods could be included. tested due to time and computational resource constraints. Overall, the findings of this study indicate that the application of feature optimization and hyperparameter tuning can significantly improve the performance of diabetes prediction models.

Implications:
This study provides important practical and methodological implications in the field of healthcare analytics. The findings demonstrate that ensemble-based machine learning models, particularly Random Forest and Gradient Boosting, achieve superior performance in predicting diabetes, indicating their strong potential for integration into clinical decision support systems. By enabling early and accurate detection, these models can assist healthcare professionals in identifying high-

risk patients more efficiently and reducing the likelihood of delayed diagnosis. Moreover, the emphasis on minimizing false negatives highlights the clinical relevance of the model, as misclassification in such cases can lead to serious health consequences (Zhang, 2025).

From a methodological perspective, the study underscores the importance of adopting a robust and reproducible machine learning pipeline. The integration of preprocessing, feature selection, hyperparameter optimization, and stratified cross-validation ensures model reliability and reduces the risk of data leakage. This comprehensive framework can serve as a reference for future studies aiming to develop reliable predictive models in the medical domain.

Research Contribution:
This research contributes significantly to both the methodological and empirical development of machine learning applications in healthcare. Methodologically, it proposes a comprehensive evaluation framework that combines stratified cross-validation, GridSearchCV-based hyperparameter tuning, and feature selection using Recursive Feature Elimination. In addition, the study provides a deeper analytical focus on confusion matrix evaluation, particularly concerning false negative rates, which are often overlooked in similar studies.

Empirically, the study confirms that ensemble methods outperform traditional algorithms in handling complex and non-linear healthcare data. Random Forest achieved the highest performance across all evaluation metrics, supported by strong evidence from accuracy and AUC values. Furthermore, the identification of key predictive features such as glucose level, BMI, and age strengthens the interpretability of the model and aligns with established clinical knowledge, enhancing the study's practical relevance.

Limitations:
Notwithstanding its significant insights, this research faces multiple limitations that warrant recognition. Firstly, the employment of the Pima Indian Diabetes dataset, which encompasses a rather limited sample size, constrains the applicability of the results to wider populations. This dataset may not adequately reflect varied demographic and clinical traits, possibly leading to bias in the model.

Additionally, the models have not been validated using external datasets, making it difficult to assess their performance in real-world scenarios. The study also does not explore more advanced machine learning approaches such as deep learning models due to time and computational constraints. Although cross-validation was applied, the possibility of overfitting remains, and the developed models are not yet ready for direct clinical implementation without further validation and testing.

Suggestions:
Future research should focus on enhancing the robustness and applicability of diabetes prediction models by utilizing larger and more diverse datasets, including real-world clinical data from multiple populations. External validation is essential to ensure the generalizability and stability of the model across different settings. Expanding the study to include advanced algorithms such as deep learning and hybrid ensemble methods could further improve predictive performance.

Moreover, future studies should prioritize reducing false negative rates, given their critical impact in medical diagnosis. Adjusting classification thresholds based on clinical needs and integrating explainable AI techniques can also improve model transparency and trust. Finally, efforts should be directed toward implementing and testing these models in real clinical environments, such as hospital systems or telemedicine platforms, to evaluate their practical effectiveness and readiness for deployment.

CONCLUSION

This research intends to perform a comparison evaluation of several machine learning techniques for forecasting diabetes outcomes using secondary datasets. Based on the experiment, it was

concluded that the application of feature engineering and hyperparameter tuning contributed to improved model performance on the datasets used. among the models that were tested, Random Forest showed. relatively better performance compared to other models based on the applied assessment measures such as accuracy, F1-Score, and AUC, However, these performance results should be interpreted with caution due to their limitations within the scope and characteristics of this study's data.

The findings additionally indicated that models using ensemble techniques, like Random Forest and Gradient Boosting, tended to produce more consistent performance than conventional models like Logistic Regression and K-NN. This finding indicates that ensemble approaches are potentially more suitable for handling data with non-linear relationships and complex interactions between features. However, these advantages cannot be generalized without additional testing on different, more representative datasets.

Furthermore, this study confirms that selecting relevant features and tuning model parameters are crucial aspects of the predictive model development process. However, the performance improvements achieved cannot yet be used as a basis for declaring the model's readiness for clinical use. Limitations of this study include the relatively limited dataset size, potential data bias, and the lack of external validation or clinical trials.

Considering these limitations, the Random Forest approach in this study can only be viewed as a promising approach for further research, not a definitive solution for early diabetes detection. Future research is recommended to involve the use of more optimal and diverse datasets, implement cross-population validation, and evaluate the model in clinical scenarios more closely resembling real-world conditions. Furthermore, the use of evaluation metrics that focus on critical classification errors, such as False Negatives, and adjustment of classification thresholds according to medical needs require further study. With these steps, it is hoped that the development of machine learning-based diabetes prediction models can be conducted more responsibly and have stronger practical relevance.

ACKNOWLEDGMENT

The writers wish to convey their heartfelt appreciation to everyone who played a role in finishing this research. Particular thanks are given to the organization for offering resources and assistance throughout the research journey. The writers also recognize the important advice and recommendations received from mentors and peers. In addition, thanks are offered to all the respondents and participants who helped with this study.

AUTHOR CONTRIBUTION STATEMENT

Every author has played an essential role in this study, which encompasses the development and planning of the research, gathering and examining data, in addition to writing and thoroughly reviewing the manuscript. Each author has reviewed and consented to the final draft of the manuscript intended for publication.

REFERENCES

Aditya, M. F., & Pramuntadi, A. (2024). Implementation of Decision Tree Method for Diabetes Mellitus Type 2 Prediction. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(3), 1104–1110. https://doi.org/10.57152/malcom.v4i3.1284

Alzboon, M. S., Al-Batah, M., Alqaraleh, M., Abuashour, A., & Bader, A. F. (2025). A Comparative Study of Machine Learning Techniques for Early Prediction of Diabetes. *Institute of Electrical and Electronics Engineers (IEEE)*, *1*, 1–12. https://doi.org/10.1109/comnet60156.2023.10366688

Arfiah, S., Wajidi, F., & Nur, N. (2025). Optimization of the K-Nearest Neighbors Algorithm in Diabetes Prediction. *JURIKOM (Journal of Computer Research)*, *12*(3), 230–240.

https://doi.org/10.30865/jurikom.v12i3.8615
Bhatta, R. P. (2025). Comparative Study of Diabetes Prediction using Machine Learning Approaches NPRC Journal of Multidisciplinary Research. *NPRC Journal of Multidisciplinary Research*, *2*(14), 1–15. https://doi.org/10.3126/nprcjmr.v2i14.86923
Diranisha, V., Triayudi, A., & Komalasari, R. T. (2024). Implementation of K-Nearest Neighbour (KNN) Algorithm and Random Forest Algorithm in Identifying Diabetes. *SAGA: Journal of Technology and Information Systems*, *2*(2), 234–244. https://doi.org/10.58905/SAGA.vol2i2.253
Dwinnie, Z. C., Dwynne, Z. C., Islam, M. J., & Noviarni. (2024). Comparison of Machine Learning Algorithms in Diabetes Risk Classification. *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, *1*(2), 54–60. https://doi.org/10.57152/ijatis.v1i2.1141
Enríquez-ortega, D., Chulde-fernández, B., Pozo-coral, P., Vaca, A., Zhinin-vera, L., Almeida-galárraga, D., Ramírez-cando, L., Tirado-espín, A., Cadena-morejón, C., Villalba-meneses, F., Guevara, C., & Acosta-vargas, P. (2025). Enhancing Diabetes Diagnosis Through Machine Learning : A Comparative Study. *Aplpied Sciences*, *15*(10087), 1–18. https://doi.org/10.3390/app151810087
Fadhlullah, A. F., & Widiyaningtyas, T. (2024). Comparative Analysis of Decision Tree and Random Forest Algorithms for Diabetes Prediction. *JTAM (Journal of Mathematical Theory and Applications), 8*(4), 1121–1132. https://doi.org/10.31764/jtam.v8i4.24388
Hadi, D. A., Agustin, D., & Sirodj, N. (2022). Random Forest Method for Classification of Diabetic Diseases. *Bandung Conference Series: Statistics*, *3*(2), 428–435. https://doi.org/10.29313/bcss.v3i2.8354
Kurniawati, F., & Arianto, D. B. (2023). Analysis of the Implementation of Feature Selection in Diabetes Classification with Corellation Matrix Method and Logistic Regression Algorithm. *JOURNAL OF INFORMATICS*, *4221*(2020), 157–164. https://doi.org/10.52958/iftk.v19i3.6019
Liu, S. (2023). Diabetes Prediction by KNN , SVM , Random Forest and XGBoost. *Highlights in Science, Engineering and Technology*, *72*, 1113–1120. https://doi.org/10.3126/nprcjmr.v2i14.86923
Mangai, M. J., Ayenajeh, G. T., Enekai, O. D., Jr, S. M., Dirting, B. D., & Betty, D. (2026). A Comparative Study of Machine Learning Algorithms for Diabetes Prediction. *International Journal Of Research And Innovation In Applied Science (IJRIAS)*, *XI*(2454), 1273–1280. https://doi.org/10.51584/IJRIAS
Maulana, M. R., & Karomi, M. A. Al. (2023). Classification of Type 2 Diabetes using Decission Tree Algorithm. *JAICT Journal of Applied Communication and Information Technologies*, *8*(2), 236–241. https://doi.org/10.32497/jaict.v8i2.4835
Mujumdar, A., & Vaidehi, V. (2019). ScienceDirect ScienceDirect ScienceDirect ScienceDirect Diabetes Prediction using Machine Learning Aishwarya Mujumdar Diabetes Prediction using Machine Learning Aishwarya Mujumdar Aishwarya. *Procedia Computer Science*, *165*, 292–299. https://doi.org/10.1016/j.procs.2020.01.047
Nguyen, B., & Zhang, Y. (2025). A Comparative Study of Diabetes Prediction Based on Lifestyle Factors Using Machine Learning. *ArXiv*, *2*(1), 1–5. https://doi.org/10.48550/arXiv.2503.04137
Oktaviana, A., Wijaya, D. P., Pramuntadi, A., & Heksaputra, D. (2024). Prediction of Type 2 Diabetes Mellitus Using The K-Nearest Neighbor (K-NN) Algorithm. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(3), 812–818. https://doi.org/10.57152/malcom.v4i3.1268
Peerbasha, S., Iqbal, Y. M., Surputheen, M. M., & Raja, A. S. (2023). Diabetes Prediction using Decision Tree, Random Forest, Support Vector Machine, K- Nearest Neighbors, Logistic Regression Classifiers. *Journal of Advanced Applied Scientific Research*, *5*(4), 42–54. https://doi.org/10.46947/joaasr54202368
Ridla, M. A., & Prayoga, M. N. (2026). Diabetes Prediction Analysis Using Decision Tree Method on Pima Indians Diabetes Dataset. *Journal of Information System and Application Development*, *4*(1), 146–153. https://doi.org/10.26905/jisad.v4i1.16494
Siswoyo, B., & Nurhafidz, M. I. (2025). Application of the Random Forest algorithm for diabetes risk prediction based on patient health data. *Digital Information Technology (JTID), 1*(1), 35–38. https://doi.org/10.31284/j.jtmik.2025.v5i2.XXXX
Sudestra, I. M. A., Wahyudi, A., Gama, O., & Prathama, G. H. (2026). Comparative Performance of Machine Learning Algorithms for Diabetes Prediction. *Of Technology and Informatics (JoTI)*,

8(1), 50–61. https://doi.org/10.37802/joti.v8i1.1195

Wantoro, A., Yulia, A. F., Ayu, D. Y., Mustofa, S., Informatics, J. T., Pringsewu, U. A., Doctor, J. P., Medicine, F., & Lampung, U. (2025). Evaluation of Machine Learning (ML) Algorithm Performance Using Feature Selection in Diabetes Classification. *JIP (Journal of Polynesian Informatics), 11*, 311–316. https://doi.org/10.33795/jip.v11i3.7290.

Yulianty, S., & Najib, M. K. (2025). Comparing the Accuracy of K-Nearest Neighbour (KNN), Random Forest , and Decision Tree Methods in Predicting Diabetes. *Al-Aqlu: Jurnal Matematika, Teknik Dan Sains*, *3*(2), 144–151. https://doi.org/10.59896/aqlu.v3i2.299

Zhafran, S., Krishandhie, R., & Purwinarko, A. (2025). Random Forest Algorithm Optimization using K-Nearest Neighbor and SMOTE on Diabetes Disease. *Recursive Journal of Informatics*, *3*(1), 43–49. https://doi.org/10.15294/rji.v3i1.1576

Zhang, F. (2025). Comparative Analysis of Machine Learning Models in Early- Stage Diabetes Prediction. *Proceedings of ICEGEE 2025 Symposium: Sensor Technology and Multimodal Data Analysis 2025*, *1*, 63–69. https://doi.org/10.54254/2753-8818/2025.AU24408