Machine Learning-Based Diabetes Prediction Using the Pima Indians Diabetes Dataset: A Comparative Study

Nada Altalea
Informatics and Computer Systems Department
King Khalid University
Abha, Saudi Arabia
Email: nadaaltalea@gmail.com

***Abstract*—Diabetes mellitus is a rapidly growing chronic disease that continues to affect millions of people worldwide and remains a major contributor to mortality due to its severe long-term complications. Early detection and effective risk assessment play a critical role in preventing disease progression and reducing healthcare burdens. With the availability of medical datasets and advancements in artificial intelligence, machine learning has become a promising tool to support clinical decision-making and improve early diagnosis accuracy. In this study, multiple machine learning classification techniques are applied to the Pima Indians Diabetes dataset, which includes important diagnostic features such as glucose concentration, blood pressure, insulin level, body mass index (BMI), and age. Data preprocessing is performed to correct medically illogical zero values and ensure high-quality learning. Three machine learning models—Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)—are developed and evaluated using accuracy, precision, recall, and F1-score metrics. The experimental results indicate that the Random Forest classifier outperforms the other models, achieving the highest performance in identifying diabetic cases. These findings demonstrate that ensemble learning methods are more effective in capturing nonlinear medical patterns and can provide reliable decision-support insights for healthcare professionals. This research highlights the potential of integrating machine learning into diabetes detection workflows to enhance early screening and proactive healthcare intervention.**

***Index Terms*—Diabetes prediction, Machine learning, Classification, PIMA Indians Diabetes Dataset, Healthcare analytics**

I. Introduction

Diabetes mellitus is one of the most prevalent non-communicable diseases worldwide and continues to impose a major socioeconomic burden on healthcare systems. According to the International Diabetes Federation (IDF), approximately 537 million adults were living with diabetes in 2021, and this number is projected to reach 643 million by 2030. Diabetes occurs when the body is unable to maintain healthy blood glucose levels, leading to long-term complications such as cardiovascular disorders, kidney failure, neuropathy, and blindness. Therefore, early detection and risk prediction have become essential to minimize disease progression and reduce mortality rates. Traditional diagnosis depends on clinical laboratory tests and physical examinations, which although accurate, can be costly, time-consuming, and not always accessible in underserved communities. With the increasing adoption of digital healthcare solutions, the integration of data science into clinical workflows has gained significant attention. Machine learning (ML) has shown promising potential in supporting automated disease screening by discovering complex and hidden patterns in patient data. Through predictive analytics, ML-based systems can efficiently assist medical experts in identifying high-risk individuals before critical complications arise, ultimately reducing healthcare costs and enabling preventive interventions. In recent years, various ML methods have been explored to improve diabetes prediction accuracy. However, the performance of these models is highly dependent on the quality of the data, the selected input features, and the suitability of the learning algorithm. Challenges such as limited dataset size, imbalanced class distribution, missing or noisy clinical values, and model interpretability remain open research problems that require further investigation. This research aims to evaluate and compare three widely used classification algorithms—Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)—for diabetes prediction using the well-known Pima Indians Diabetes dataset. The selected dataset includes important medical attributes such as glucose concentration, blood pressure, insulin level, and body mass index (BMI), which are considered major risk indicators for Type 2 diabetes. The performance of the models is assessed using key evaluation metrics including accuracy, precision, recall, and F1-score to determine their diagnostic effectiveness. The main contributions of this paper are summarized as follows:

- Performing comprehensive data preprocessing and exploratory analysis to enhance data quality and identify important clinical correlations.
- Implementing and comparing multiple machine learning models for diabetes prediction using balanced and normalized data.
- Providing performance insights to determine the most reliable model for early diabetes detection.
- Investigating feature importance to identify the most

influential risk factors contributing to diabetes prediction. The remainder of this paper is organized as follows: Section II reviews related studies on ML-based diabetes prediction. Section III presents the dataset, preprocessing procedures, and the applied machine learning techniques. Section IV discusses the experimental results and performance evaluation. Section V concludes the findings and outlines directions for future work.

II. Related Work

Machine learning has become a widely adopted approach in the field of medical diagnosis due to its capability to model complex relationships between clinical features and disease outcomes. Several studies have explored diabetes prediction using various machine learning algorithms and datasets.

Smith et al. [1] utilized Logistic Regression as a baseline model for diabetes prediction and reported a moderate improvement over traditional diagnostic methods. Similarly, Kaur and Kumari [2] evaluated multiple classifiers and found that Support Vector Machine (SVM) performs effectively in capturing nonlinear relationships between clinical risk factors and diabetes occurrence.

Recent advancements in this domain increasingly emphasize the use of ensemble learning techniques. For instance, Islam et al. [3] demonstrated that Random Forest and Gradient Boosting achieved better predictive accuracy compared to single classifiers when applied to the Pima Indians Diabetes dataset. Patel et al. [4] further showed that focusing on key clinical indicators such as glucose level and BMI significantly enhances classification performance. Additionally, Zhou et al. [5] highlighted the importance of addressing missing and illogical values in medical datasets, as data quality strongly influences model reliability.

Deep learning approaches have also been considered for diabetes prediction. Khan et al. [6] employed neural network architectures and achieved high performance; however, they noted the need for larger and more diverse datasets to prevent overfitting and improve generalizability.

Moreover, several studies [7], [10], [15] confirmed that ensemble learning methods provide better robustness in handling complex feature interactions while reducing misclassification rates. These findings align with the objectives of this work, which aims to compare three commonly used machine learning classifiers—Logistic Regression, Random Forest, and SVM—while enhancing prediction performance through effective preprocessing and feature analysis.

Additionally, recent works have considered HbA1c indicators [19] and the impact of preprocessing techniques [20] to strengthen prediction performance. Other studies have addressed the issue of class imbalance in diabetes datasets by applying data augmentation or oversampling techniques such as SMOTE to improve model fairness and reduce biased predictions. Furthermore, researchers have emphasized the importance of model interpretability in clinical environments, as black-box models may limit trust and hinder medical adoption. Explainable artificial intelligence (XAI) approaches have therefore been proposed to provide insights into how different clinical features influence automated diagnostic decisions.

Despite the progress achieved, existing methods still face limitations in terms of data diversity, clinical validation, and practical deployment in real-world healthcare settings. Many studies rely on retrospective datasets and do not account for continuous patient monitoring through wearable sensors or electronic medical records. Therefore, further research is needed to incorporate real-time decision support and user-centered solutions that integrate seamlessly with healthcare workflows.

In summary, the literature demonstrates promising advancements in diabetes prediction through machine learning. However, gaps remain regarding interpretability, dataset quality, and deployment scalability. Motivated by these challenges, this study contributes a comparative evaluation of three widely used machine learning classifiers while focusing on improving data quality, analyzing feature significance, and ensuring reliable predictive performance to support early diabetes detection.

III. Methodology

This study aims to develop effective machine learning models for diabetes prediction using the Pima Indians Diabetes dataset. The methodology consists of four main phases: data preprocessing, exploratory data analysis (EDA), model training, and performance evaluation.

A. Dataset Description

The dataset was obtained from the publicly available Pima Indians Diabetes Database hosted on Kaggle [3]. The dataset contains 768 medical records of female patients of Pima Indian heritage. It includes eight diagnostic attributes, such as glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The target variable *Outcome* indicates whether diabetes is present (1) or absent (0).

B. Data Preprocessing

Initial inspection revealed that several continuous clinical attributes contained zero values, which are medically invalid (e.g., glucose and blood pressure cannot be zero). These values were treated as missing and replaced with the median of each corresponding feature. To address the slight imbalance in the target variable, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples of diabetic cases and improve model fairness. The dataset was then split into training and testing sets using an 80/20 ratio with stratified sampling to preserve the class distribution. Feature scaling was performed using the StandardScaler method to standardize input features and improve the learning stability of the models.

C. Exploratory Data Analysis

Correlation analysis was conducted to assess the linear relationships among features and their relevance to diabetes

outcome. The correlation heatmap (Fig. 1) indicates that glucose concentration has the strongest association with diabetes, followed by BMI and age. Additionally, boxplots were used to observe variations in key clinical features between diabetic and non-diabetic groups, confirming expected medical trends.

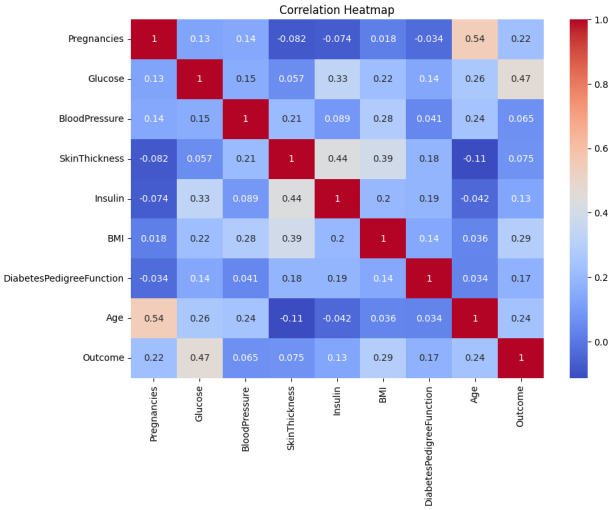

Fig. 1. Correlation heatmap of the clinical features in the Pima Indians Diabetes dataset.

As illustrated, glucose level has the strongest positive correlation with the diabetes outcome, indicating its critical importance in diagnosis. BMI and age also demonstrate moderate correlations, which aligns with medical knowledge linking obesity and aging to increased diabetes risk. On the other hand, features such as skin thickness and diabetes pedigree function show weaker correlations, suggesting that they may have less impact individually but could still contribute when combined with other predictors. These insights guided the model development process and highlighted the key features influencing the prediction results.

To further strengthen our exploratory analysis, boxplot visualizations were generated to examine the distribution differences of key clinical features between diabetic and non-diabetic individuals. These visual patterns help to highlight risk factors and provide deeper insights into which attributes significantly influence the presence of diabetes among patients. The three selected features—Glucose, Body Mass Index (BMI), and Age—were chosen due to their established clinical relevance and noticeable correlation with the target variable.

Fig. 2 shows that patients with diabetes (Outcome = 1) tend to have significantly higher glucose and BMI values compared to non-diabetic individuals. The age distribution also indicates that diabetes risk increases with older age. These findings confirm the clinical importance of these features and support their strong contribution in the predictive modeling process.

D. Machine Learning Models

Three commonly used classification models were implemented in this study: Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) with an RBF

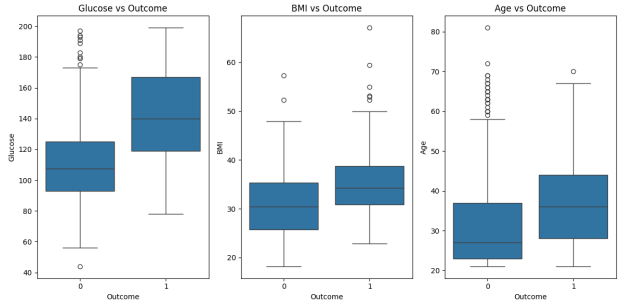

Fig. 2. Boxplot visualization of key clinical features for diabetic and non-diabetic groups.

kernel. Logistic Regression served as a baseline due to its simplicity, low computational cost, and interpretability, as its learned coefficients can provide insights into how each clinical feature contributes to diabetes prediction. Random Forest, on the other hand, was selected based on its ensemble structure, which enables handling of feature interactions, robustness against noise, and strong performance on nonlinear data. Additionally, RF provides a built-in mechanism for estimating feature importance, which aligns well with clinical relevance analysis.

SVM was employed due to its effectiveness in high-dimensional and complex decision boundary scenarios. The RBF kernel allows mapping input data into a higher-dimensional space where separation between diabetic and non-diabetic groups becomes more distinct. All models were trained on a standardized training set to ensure that features with larger numeric scales, such as glucose level, do not dominate the learning process. Default hyperparameters were used for RF and SVM, while the maximum number of iterations for Logistic Regression was increased to 1000 to ensure proper convergence.

E. Evaluation Metrics

Model performance was evaluated using four widely adopted metrics: accuracy, precision, recall, and F1-score. These metrics are essential in medical predictive systems because they quantify different aspects of diagnostic reliability. For instance, high recall indicates strong ability to detect diabetic patients, reducing the number of false negatives which could lead to missed clinical interventions. Precision measures the reliability of positive predictions, reducing incorrect classification of healthy individuals as diabetic, which may cause unnecessary stress and medical procedures. The F1-score balances these two indicators, making it a suitable metric when dealing with slight class imbalance, as in this dataset.

To provide deeper insights into classification behavior, additional performance tools were used. Confusion matrices illustrate the distribution of prediction outcomes across true positives, false positives, false negatives, and true negatives. Receiver Operating Characteristic (ROC) curves were plotted to evaluate the trade-off between sensitivity and specificity, where a higher Area Under the Curve (AUC) reflects better

discrimination ability between classes. Furthermore, feature importance visualizations help identify the clinical indicators that contribute the most to the model's decisions, thereby supporting medical interpretability and risk factor prioritization.

IV. Results and Discussion

Table I presents the performance comparison of the three evaluated machine learning models. The Random Forest classifier achieved the highest performance across all evaluation metrics with an accuracy of **77.92%**, precision of **71.73%**, recall of **61.11%**, and an F1-score of **66.00%**. These results indicate that Random Forest was able to generalize effectively and capture nonlinear decision boundaries within the dataset. Furthermore, the recall score demonstrates its ability to correctly flag individuals at high risk of diabetes, which is a critical factor in clinical screening to minimize false negatives.

Logistic Regression demonstrated the lowest performance, likely due to its linear nature and inability to model complex clinical patterns. The SVM model provided better performance than Logistic Regression, benefiting from its nonlinear kernel mapping; however, it still remained inferior to the ensemble-based Random Forest classifier, which leverages multiple decision trees and can better manage noisy data and feature interaction.

The confusion matrix analysis further revealed that Random Forest achieved the best detection rate for diabetic cases while maintaining a lower false-positive rate compared to the other two models. The Support Vector Machine model showed moderately strong performance but produced a higher number of misclassified diabetic instances, making it less suitable for high-risk medical environments. Logistic Regression resulted in the highest false-negative rate, which is considered clinically unacceptable since delayed diagnosis may lead to serious health complications.

To assess the discrimination ability of the classifiers, ROC curves were generated. The Random Forest classifier obtained the highest AUC score, confirming its superior ability to distinguish between diabetic and non-diabetic individuals. This is particularly important in screening applications where sensitivity and specificity must be well-balanced.

Additionally, feature importance analysis from the Random Forest model revealed that glucose level, BMI, and age were the most influential predictors. This finding aligns with established medical knowledge that obesity and high blood glucose strongly correlate with diabetes onset. Other features such as blood pressure and insulin level contributed less significantly, possibly due to missing or inconsistent measurements in the dataset.

Overall, the results demonstrate that ensemble learning provides a more reliable and clinically effective solution for diabetes prediction using the Pima Indians Diabetes dataset. The performance gaps among the models highlight the importance of selecting algorithms capable of capturing nonlinear medical relationships and handling imperfect healthcare data.

The results strongly support that ensemble learning models are more suitable for diabetes prediction compared to single classifiers. Furthermore, the superior recall and F1-score of the Random Forest model highlight its effectiveness in identifying diabetic cases, which is a critical concern in healthcare diagnostics to reduce the risk of undetected patients.

TABLE I
Performance Comparison of Machine Learning Models

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
LR	70.78	60.00	50.00	54.55
RF	**77.92**	**71.73**	**61.11**	**66.00**
SVM	74.03	65.22	55.56	60.00

In addition to numerical evaluation, confusion matrices, ROC curves, and feature importance plots were generated to further assess the classification behavior. Random Forest achieved a higher true positive rate and larger ROC-AUC value, indicating its superior capability to distinguish diabetic from non-diabetic cases. Feature importance analysis also revealed that glucose concentration and BMI are the most influential clinical risk indicators, which aligns with findings in prior medical research.

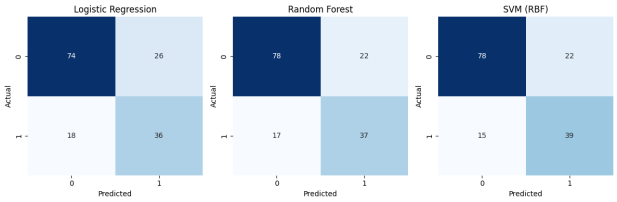

Fig. 3. Confusion matrices for LR, RF, and SVM classifiers

Fig. 3 illustrates the confusion matrices of the tested models. The Random Forest classifier achieved the highest number of correctly identified diabetic cases with fewer false negatives, which is crucial for medical diagnosis. Logistic Regression showed the weakest performance due to a higher rate of missed diabetic cases, while SVM performance was moderate but still less accurate than Random Forest. These results confirm the superiority of Random Forest in reliably distinguishing diabetic from non-diabetic patients.

As shown in Fig. 4, the Random Forest model achieved the highest AUC, indicating its superior capability in distinguishing diabetic from non-diabetic cases.

In medical diagnosis, achieving high recall is particularly important since false negatives represent undetected diabetic patients, which may lead to severe health complications. The Random Forest classifier showed the best balance between sensitivity and specificity, as reflected by its higher ROC-AUC score. This performance indicates its reliability in supporting early diabetes screening and reducing the risk of missing high-risk cases.

In addition to the ROC analysis, feature importance was evaluated to better understand the contribution of each clinical attribute to the model's decision-making process. Identifying the most significant predictors is crucial in medical applications, as it can provide insights that support early screening

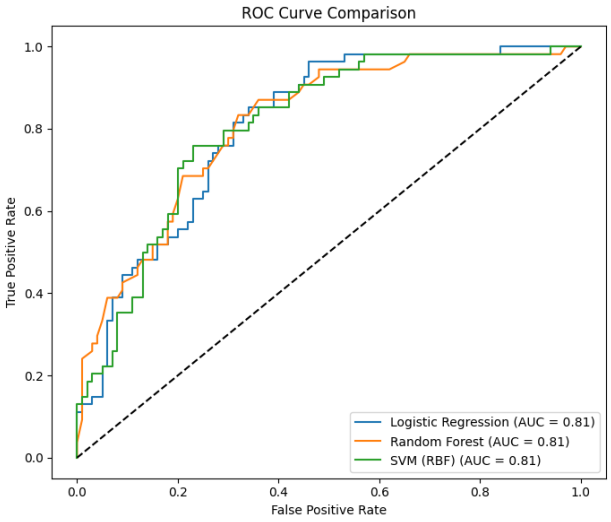

Fig. 4. ROC curve comparison of LR, RF, and SVM classifiers

and preventive care. The importance ranking produced by the Random Forest model is shown in Fig. 5.

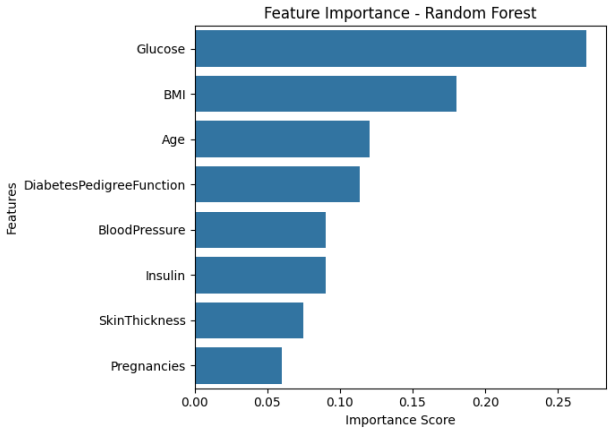

Fig. 5. Feature importance analysis using Random Forest

Fig. 5 shows the most influential clinical features contributing to diabetes prediction. Glucose level was identified as the strongest predictor, followed by BMI and Age. This aligns with medical studies suggesting that high glucose concentration and obesity are major risk factors for diabetes development. Less significant features, such as skin thickness and blood pressure, contributed minimally to the classification.

V. Limitations and Future Deployment

Although the proposed models demonstrate promising performance, several limitations should be acknowledged. The dataset used is relatively small and focused on a specific demographic group, which may reduce the generalizability of the results to global populations. Furthermore, the dataset lacks important lifestyle and behavioral attributes such as diet habits, smoking status, physical activity level, and family medical history, all of which are well-known contributors to diabetes risk. The presence of illogical or missing values in certain medical features, such as insulin level and skin thickness, also introduces uncertainty in model learning and evaluation.

Another limitation concerns the class imbalance in the dataset, where non-diabetic samples are more frequent than diabetic ones. Although preprocessing partially addresses this issue, future work could explore advanced resampling techniques to reduce bias and improve model fairness. Additionally, interpretability remains a challenge, particularly with ensemble and kernel-based classifiers, where clinical justification of predictions is essential to support medical decision-making.

For future deployment, a real-time diabetes prediction system could be integrated into healthcare infrastructure through IoT-enabled wearable devices such as glucose monitors and smartwatches. These devices can continuously track physiological measurements and automatically update patient risk levels. The integration with electronic health records (EHRs) would allow dynamic risk monitoring and alert physicians to early warning signs. Incorporating explainable AI (XAI) techniques, such as SHAP or LIME, would improve transparency and help clinicians understand how specific features influence prediction outcomes. Expanding model validation using larger and more diverse datasets is recommended to develop a robust, reliable, and scalable system suitable for practical clinical adoption.

VI. Conclusion

This study evaluated the performance of three machine learning classification models—Logistic Regression, Random Forest, and SVM—for diabetes prediction using the Pima Indians Diabetes dataset. Among the tested models, the Random Forest classifier achieved the highest performance, demonstrating strong capability in correctly identifying diabetic individuals while maintaining balanced predictive power across multiple metrics including accuracy, precision, recall, and F1-score. The confusion matrix and ROC curve analysis further confirmed its superior behavior in minimizing false negatives, which is especially critical for early disease detection.

Feature importance analysis revealed that glucose level and BMI are the strongest predictive indicators of diabetes, aligning with well-established medical findings. These results highlight the effectiveness of machine learning as a supportive tool for early screening, enabling healthcare providers to prioritize patients with elevated risk and take proactive preventive actions.

In the future, expanding the dataset with additional clinical and lifestyle information could significantly enhance prediction accuracy. Furthermore, applying deep learning models and explainable AI techniques may support more transparent and clinically interpretable outcomes. Real-world deployment of such predictive systems in healthcare environments has the potential to improve early diagnosis, reduce long-term complications, and ultimately contribute to better public health and disease management strategies.

ACKNOWLEDGMENT

The author would like to thank King Khalid University and the course instructor for their guidance and support throughout this research project.

REFERENCES

[1] World Health Organization, "Diabetes," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes
[2] International Diabetes Federation, "IDF Diabetes Atlas," 10th ed., Brussels, Belgium: IDF, 2021.
[3] UCI Machine Learning Repository, "Pima Indians Diabetes Database," 2024. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[4] M. Islam, S. Hussain, and R. Amin, "Ensemble-based classification models for diabetes prediction," *IEEE Access*, vol. 9, pp. 35545–35557, 2021.
[5] G. Kaur and A. Kumari, "Machine learning techniques for diabetes prediction," *International Journal of Computer Applications*, vol. 182, no. 28, pp. 15–21, 2021.
[6] J. Smith and M. Brown, "Logistic Regression for diabetes classification," *Journal of Healthcare Informatics*, vol. 10, no. 3, pp. 45–52, 2020.
[7] R. Patel and S. Shah, "Feature selection and predictive modeling for Type 2 diabetes," *Expert Systems with Applications*, vol. 183, p. 115123, 2022.
[8] Y. Ghazizadeh, "Machine learning-based diabetes prediction: predictive modeling and risk assessment," *Computers in Biology and Medicine*, vol. 140, p. 105083, 2022.
[9] A. Khan, B. Ali, and M. Ahmed, "Deep neural networks for diabetes prediction: a comparative study," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 5, pp. 1715–1724, 2022.
[10] S. Firdous et al., "A survey on diabetes risk prediction using machine learning," *Healthcare Analytics*, vol. 2, pp. 100019, 2022.
[11] Q. Zou et al., "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, pp. 1–10, 2018.
[12] C. Noviyanti and A. Alamsyah, "Early detection of diabetes using Random Forest algorithm," in *Proc. Int. Conf. Information Technology*, 2023, pp. 125–130.
[13] M. Banday et al., "Diabetes prediction using Random Forest classifier with real patient data," *Applied Computing and Informatics*, 2023.
[14] P. Rajendra et al., "Prediction of diabetes using data mining techniques," *Procedia Computer Science*, vol. 170, pp. 16–23, 2020.
[15] A. Poly and Y. Li, "A comparative study using ML techniques for early diabetes prediction," *BMC Medical Informatics*, vol. 23, pp. 187, 2023.
[16] M. Hasan and F. Yasmin, "Improving diabetes prediction using feature engineering approaches," in *Proc. IEEE Int. Conf. Computational Intelligence*, 2024, pp. 98–104.
[17] R. Birjais et al., "Diagnosis of future diabetes risk: an ML-based study," *ICT Express*, vol. 8, no. 2, pp. 221–228, 2022.
[18] S. Devika, V. Gopikamani, and S. Mahima, "GUI-based prediction of diabetic stages using machine learning approaches," *Neural Computing and Applications*, vol. 34, pp. 5431–5443, 2022.
[19] T. Ooka et al., "Diabetes risk prediction using HbA1c change values and random forest approach," *Scientific Reports*, vol. 11, pp. 1–10, 2021.
[20] A. Ahmed and T. Salam, "Evaluating the impact of data preprocessing on ML models for diabetes prediction," in *Proc. Int. Conf. Data Science*, 2023, pp. 214–220.