ALGERIAN JOURNAL OF SIGNALS AND SYSTEMS (AJSS)

Comparative Studies of Intelligent Algorithms for Enhancing Machine Learning Training in Diabetes Prediction

Lakhdari Lahcen [(1)*], Belagra Mammar [(2)]
[(1)] Electrical engineering department Tahri Mohammed Béchar University Béchar Algeria
[(2)] Electrical engineering department Tahri Mohammed Béchar University Béchar Algeria
*lakhdari.lahcen@univ-bechar.dz

Abstract: This The objective of this study is to compare the effectiveness of various intelligent algorithms in enhancing machine learning training for predicting diabetes from patient data. Early prediction of diabetes is crucial for preventing serious complications, and machine learning algorithms play an essential role in improving medical diagnostics. This research evaluates the performance of several algorithms, including Logistic Regression (LR), Random Forests (RF), Support Vector Classification (SVC), Gradient Boosting Machines (GBM), and K-Nearest Neighbors Classifier (KNN). These algorithms are compared based on multiple criteria: performance (precision, recall, F1-score, accuracy), computation time, model complexity, generalization capability, robustness, ease of implementation, and scalability. The study uses the Pima Indians Diabetes dataset, a well-known dataset containing several clinically relevant variables for diabetes prediction. The algorithms are evaluated using cross-validation methods, and regularization techniques are applied to optimize the hyperparameters.

Keywords: Machine Learning, Recall, F1-score, Accuracy

1. INTRODUCTION

Diabetes is a chronic condition that affects millions of people worldwide, leading to severe health complications if not managed effectively. Early diagnosis and intervention are crucial for preventing the progression of the disease and minimizing its impact on patients' lives. Machine learning (ML) algorithms have shown great promise in the field of medical diagnostics, offering robust tools for early disease prediction and personalized healthcare solutions [1].

The primary objective of this study is to compare the effectiveness of various intelligent algorithms in improving machine learning training for predicting diabetes from patient data. By leveraging advanced ML techniques, we aim to enhance the accuracy and reliability of diabetes prediction models, ultimately aiding healthcare professionals in making informed decisions [2].

In this research, we evaluate the performance of several prominent machine learning algorithms, including Logistic Regression (LR), Random Forests (RF), Support Vector Classification (SVC), Gradient Boosting Machines (GBM), and K-Nearest Neighbors Classifier (KNN). Each of these algorithms brings unique strengths and weaknesses to the table, and a comprehensive comparison will provide insights into their suitability for diabetes prediction tasks [1].

The Pima Indians Diabetes dataset, a widely recognized dataset in the field of medical research, serves as the basis for our analysis. This dataset contains multiple clinically relevant variables, making it an ideal candidate for training and evaluating ML models for diabetes prediction. By utilizing cross-validation techniques and regularization methods, we aim to optimize the hyperparameters of these algorithms and ensure their robustness and generalization capability [3]

The study will focus on several key performance metrics, including precision, recall, F1-score, and accuracy. Additionally, we will consider factors such as computation time, model complexity, ease of implementation, and scalability to provide a holistic view of each algorithm's strengths and limitations [4,5].

This research not only contributes to the growing body of knowledge in the field of machine learning for medical diagnostics but also offers practical insights that can be applied to real-world healthcare settings. By identifying the most effective algorithms for diabetes prediction, we hope to pave the way for more accurate and efficient diagnostic tools that can improve patient outcomes and reduce the burden of diabetes on healthcare systems worldwide [5].

2. BACKGROUND AND IMPORTANCE

Classification is ubiquitous across many fields, from speech recognition to disease prediction. Advances in this area enable the automation of previously labor-intensive tasks and the extraction of valuable information from large datasets. Classification algorithms play a crucial role in these advances by allowing computer systems to make intelligent decisions based on input data [6].

A. Objectives

The primary objective of classification algorithms is to create models capable of predicting the class of an observation based on its features. Each algorithm approaches this task uniquely, using different techniques and methods to find the best separation between classes in the feature space.

B. Algorithm Overview

Logistic Regression (LR): Despite its name, logistic regression is actually a classification technique. It is used to predict the probability that an observation belongs to a particular class by using a sigmoid function to model class probabilities [7].

Random Forests (RF): Random forests are an ensemble method that combines the predictions of multiple decision trees to improve the model's accuracy and robustness. Each tree is trained on a random subset of the data and features, and the final prediction is based on a majority vote of the trees [7].

Support Vector Classification (SVC): Based on the Support Vector Machines (SVM) algorithm, SVC seeks to find an optimal hyperplane that separates the data into different classes while maximizing the margin between classes. It uses different kernel functions to handle both linear and nonlinear problems [7].

Gradient Boosting Machines (GBM): Gradient boosting machines are an ensemble technique that builds a predictive model by adding predictors sequentially, with each predictor correcting the errors of the previous ones. This sequential approach allows for highly accurate models by combining simple predictive models [7].

K-Nearest Neighbors Classifier (KNN): The KNN classifier is a supervised learning algorithm used for both classification and regression. The fundamental idea behind KNN is to predict the class of an observation by finding the k closest training instances in the feature space. The majority class among these neighbors is assigned to the observation to be predicted [7].

C. Applications and Implications

These classification algorithms are widely used in various fields, including finance, medicine, bioinformatics, pattern recognition, and more. Their use has significant implications for automated decision-making, trend forecasting, and the detection of hidden patterns in data.

3. LITERATURE REVIEW

Diabetes prediction is a crucial research area in predictive medicine, aiming to identify individuals at risk before symptoms appear, thereby allowing for early intervention and more effective disease management. This review examines the main classification algorithms used in diabetes prediction, namely Logistic Regression (LR), Random Forests (RF), Support Vector Classification (SVC), and Gradient Boosting Machines (GBM). Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you [8].

Logistic Regression is one of the simplest and most widely used classification methods. It models the probability that an observation belongs to a particular class by using a sigmoid function to transform the output values of the linear combination of input variables into probabilities between 0 and 1. Logistic Regression is widely used in medical diagnostics, financial risk prediction, and spam email classification [9].

Random Forests are an ensemble algorithm that combines several decision trees to improve classification accuracy. Each tree is constructed from a random sample of the data, and the final prediction is obtained by a majority vote of the trees. Random Forests are used in fraud detection, image recognition, and genomic classification [9].

SVC, based on Support Vector Machines, seeks to find the optimal hyperplane that separates classes with the largest margin. SVMs can use kernel functions to handle nonlinear problems. SVC is used in face recognition, bioinformatics for protein classification, and spam detection [9]

Gradient Boosting Machines build a predictive model by adding predictors sequentially, with each new predictor correcting the errors of the previous predictors. The algorithm thus combines several weak models to form a strong model. GBM is used in demand forecasting, credit scoring, and sentiment analysis [9]

K-Nearest Neighbors (KNN) is often used in classification applications such as pattern recognition, anomaly detection, and content filtering [9]

4. METHODS

A. Data, Features, and Software Tools

In our research, the Pima Indian Diabetes (PID) dataset was collected from the UCI Machine Learning Repository, originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). In the PID dataset, all patients are women aged at least 21 years. The dataset contains information on 768 patients and their nine unique attributes. Table 1 shows the description of the attributes in this dataset. The nine attributes used for predicting diabetes are pregnancies, BMI, insulin level, age, blood pressure, skin thickness, glucose, diabetes pedigree function, and outcome. The "outcome" attribute is taken as the dependent or target variable, and the remaining eight attributes are considered independent variables/features. The "outcome" attribute for diabetes consists of a binary value where 0 means non-diabetic and 1 implies diabetic. In our research, we used machine learning algorithms to predict whether a patient is diabetic or not [3].

B. Data Source

For this comparative study, we use the well-known Pima Indians Diabetes Database, commonly employed in diabetes prediction research. This dataset is publicly available through the UCI Machine Learning Repository [3].

a. Data Description:

The dataset includes 768 instances and 8 attributes (features) measured for Pima women aged 21 years and older. The features are as follows:

- Number of Pregnancies: Number of times the patient has been pregnant.
- **Glucose**: Plasma glucose concentration two hours after an oral glucose tolerance test.
- **Diastolic Blood Pressure**: Measured in mm Hg.
- **Triceps Skinfold Thickness**: Measured in mm.
- **Insulin**: Two-hour serum insulin concentration.
- **Body Mass Index (BMI)**: BMI = weight in kg / (height in m)^2.
- **Diabetes Pedigree Function**: Score indicating the likelihood of diabetes based on family history.
- **Age**: Age of the patient (years).

The target variable is binary, indicating whether the patient is diagnosed with diabetes (1) or not (0).

b. Features

1) Data Preprocessing:

- **Data Cleaning:** Handling missing values using techniques such as mean or median imputation.
- **Data Normalization**: Scaling features to ensure each feature contributes equally to the classification. We use Min-Max normalization or z-score standardization.
- **Data Splitting**: Dividing the dataset into training (70%) and testing (30%) sets.

2) Feature Selection:

Although we use all available features for this study, techniques such as feature importance in random forests can be employed to improve model performance.

Software Tools

1) Programming Language:

We use Python, widely used in the machine learning community and offering a rich collection of libraries for data processing, modeling, and evaluation.

2) Libraries Used:

Pandas: For data loading and preprocessing.

NumPy: For numerical operations.

Sklearn: For modeling and evaluating machine learning algorithms. Scikit-learn provides robust and efficient implementations for Logistic Regression, Random Forests, Support Vector Classification, and Gradient Boosting Machines.

Matplotlib and Seaborn: For data and results visualization.

3) Development Environment:

Jupyter Notebook is used to write, document, and execute the code, allowing for interactive exploration and visualization of data and models

5. DATA ANALYSIS

To conduct a fundamental analysis of the Pima dataset, we will follow several steps: data exploration, preprocessing, visualization, and descriptive statistics. The following figures depict this analysis [10].

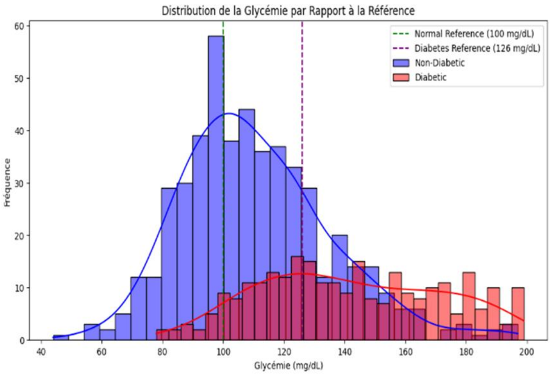

Fig. 1 Blood Glucose Distribution

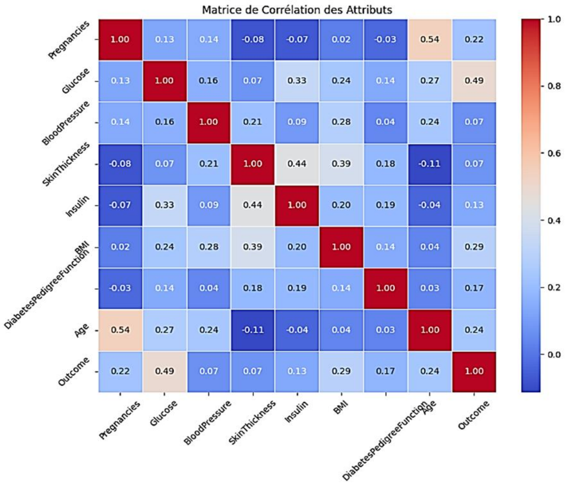

Fig. 2 Attribute Correlation Matrix

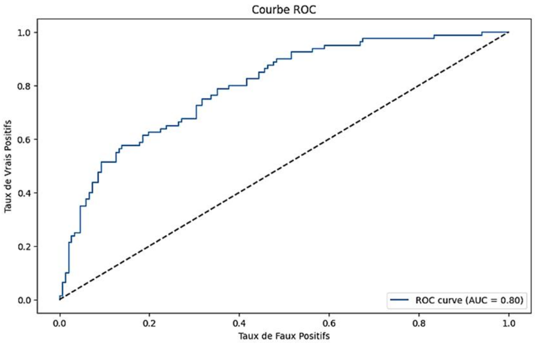

Fig. 3 ROC curve

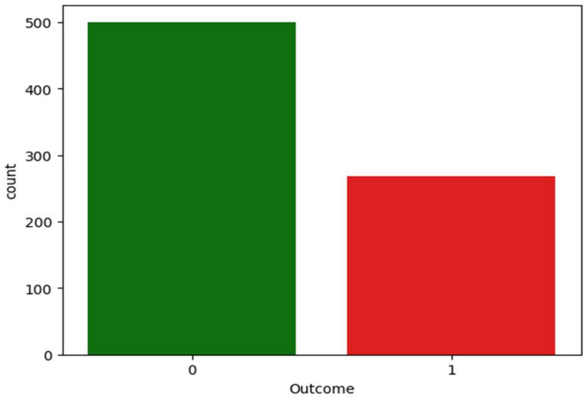

Fig. 4 The number of diabetes and no diabetes patients

6. RESULT AND DISCUSSION

In this study, we evaluated the performance of different classification algorithms for predicting diabetes using the Pima Indians Diabetes dataset. The tested algorithms include Logistic Regression, Random Forests, Support Vector Classification (SVC), Gradient Boosting Machines, and K-Nearest Neighbors.

For each model, we calculated several performance metrics, including accuracy, recall, and F1-score:

A. Accuracy:

This metric measures the proportion of correct predictions among all predictions made by the model. It provides a general indication of the model's performance [11].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (01)$$

Or :

- TP (True Positives)
- TN (True Negatives)
- FP (False Positives)
- FN (False Negatives)

B. Recall

The recall is a performance metric of a classification model that evaluates its ability to identify all true positive examples. In other words, recall measures the proportion of true positive examples that were correctly predicted among all true positive examples in the dataset [11].

$$Recall = \frac{True\ Positives(TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \qquad (02)$$

Or :

- True Positives (TP) : Number of true positive examples correctly predicted as positive.
- False Negatives (FN) : Number of true positive examples incorrectly predicted as negative.

C. F1-score

The F1-score is a performance measure of a classification model that combines both precision and recall into a single metric. It is calculated as the harmonic mean of precision and recall, thus providing a balance between these two metrics [11].

$$F1.score = 2\,X\,\frac{Precision\ X\ Recall}{Precision\ X\ Recall} \qquad (03)$$

The results of our comparative study of classification algorithms for diabetes prediction are illustrated through four main figures:

Figure 04: Precision and Recall Curve This figure shows the precision and recall curve for each classification algorithm. It allows us to visualize the relationship between these two important metrics, showing how precision and recall vary depending on the classification threshold.

Figure 05: Recall Curve The recall curve for each model is presented in this figure. It shows the ability of the different algorithms to correctly identify positive examples at various classification thresholds, highlighting their effectiveness in detecting diabetes cases.

Figure 06: F1-Score Curve This figure illustrates the F1-score curve, which combines precision and recall into a single harmonized metric. It allows us to compare the overall performance of the algorithms by taking into account their ability to avoid both false positives and false negatives.

Figure 07: Accuracy Curve The accuracy curve for each algorithm is presented here, showing the percentage of correct predictions made by each model. This figure provides an overview of the overall performance of the algorithms in terms of correct predictions.

These figures provide a visual comparison of the performance of different classification models, facilitating the evaluation of their relative effectiveness for diabetes prediction.

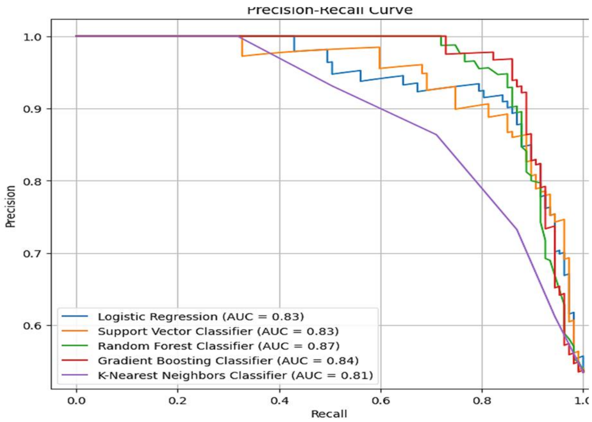

Fig. 4 Precision and Recall Curve

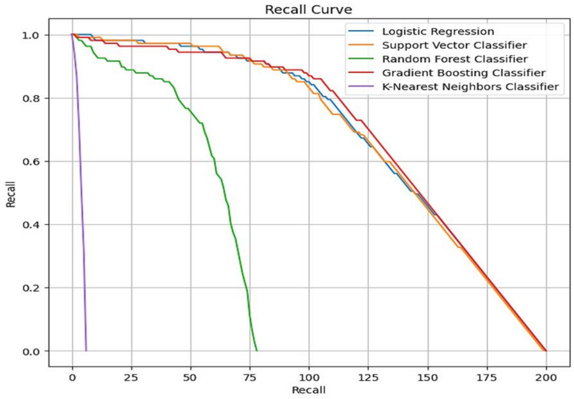

Fig. 5 Recall Curve

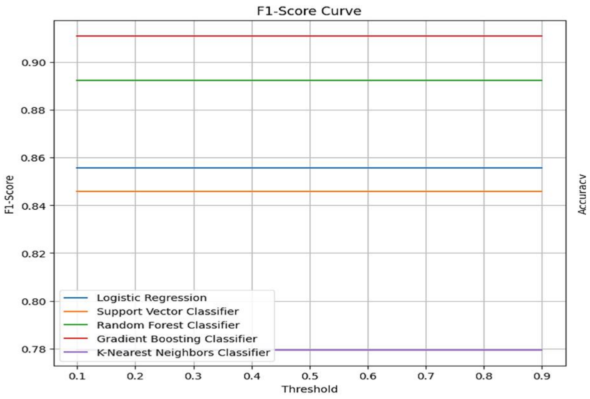

Fig. 6 F1 Score Curve

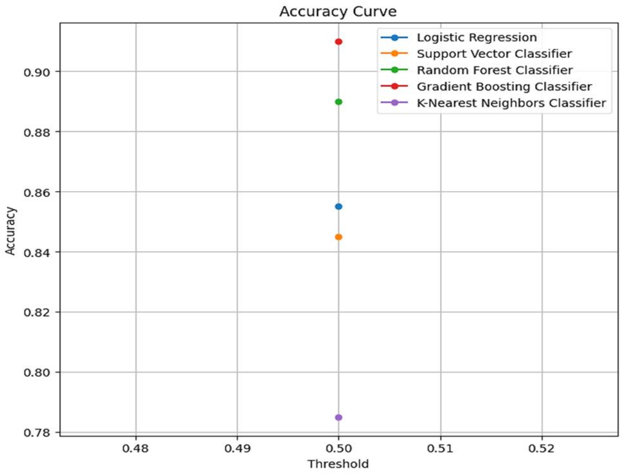

Fig. 7 Accuracy Curve

Based on the results obtained for each classification model, here is a discussion based on the precision, recall, F1-score, and accuracy metrics:

A. Logistic Regression:

Precision: For class 0 (non-diabetes), the precision is 80%, meaning that 80% of the positive predictions for class 0 are correct. For class 1 (diabetes), the precision is 91%.

Recall: For class 0, the recall is 91%, indicating that 91% of all class 0 examples were correctly identified. For class 1, the recall is 80%.

F1-score: The F1-score, which combines precision and recall into a single metric, is 0.85 for class 0 and 0.86 for class 1.

Accuracy: The overall accuracy of the model is 85%.

B. Support Vector Classifier (SVC):

Precision: For class 0, the precision is 79%, and for class 1, it is 90%.

Recall: For class 0, the recall is 90%, and for class 1, it is 79%.

F1-score: The F1-scores are 0.84 for class 0 and 0.85 for class 1.

Accuracy: The overall accuracy of the model is 84%.

C. Random Forest Classifier:

Precision: For class 0, the precision is 84%, and for class 1, it is 94%.

Recall: For class 0, the recall is 94%, and for class 1, it is 85%.

F1-score: The F1-scores are 0.89 for both classes.

Accuracy: The overall accuracy of the model is 89%.

D. Gradient Boosting Classifier:

Precision: For class 0, the precision is 86%, and for class 1, it is 97%.

Recall: For class 0, the recall is 97%, and for class 1, it is 86%.

F1-score: The F1-scores are 0.91 for both classes.

Accuracy: The overall accuracy of the model is 91%.

E. K-Nearest Neighbors Classifier (KNN):

Precision: For class 0, the precision is 72%, and for class 1, it is 86%.

Recall: For class 0, the recall is 87%, and for class 1, it is 71%.

F1-score: The F1-scores are 0.79 for class 0 and 0.78 for class 1.

Accuracy: The overall accuracy of the model is 79%.

Discussion:

The results show that the Gradient Boosting Classifier model achieves the best overall performance with an F1-score of 0.91 for both classes and an accuracy of 91%. It demonstrates a high capacity to predict both diabetes and non-diabetes cases with balanced precision and recall. In comparison, the K-Nearest Neighbors (KNN) model shows slightly lower performance with an F1-score of 0.78 and an accuracy of 79%, indicating a lesser ability to generalize compared to the other evaluated models.

7. CONCLUSION

In this comparative study, we evaluated the effectiveness of various machine learning algorithms for predicting diabetes using the Pima Indians Diabetes dataset. The tested algorithms included logistic regression, random forests, support vector classification, gradient boosting, and k-nearest neighbors.

The results show that the Gradient Boosting Classifier achieved the best overall performance with balanced precision and recall, resulting in an F1-score of 0.91 for both classes and an accuracy of 91%. This model demonstrated a high capacity to correctly predict both diabetes and non-diabetes cases.

In comparison, the K-Nearest Neighbors (KNN) model exhibited lower performance with an F1-score of 0.78 and an accuracy of 79%, indicating a lower ability to generalize compared to the other evaluated models. Other algorithms such as logistic regression, random forests, and SVC also showed good performance but were slightly behind gradient boosting.

This study highlights the importance of choosing the right machine learning algorithm for specific prediction tasks, considering different performance metrics such as precision, recall, and F1-score. Using advanced techniques like gradient boosting can significantly improve medical diagnostics and early disease management, such as diabetes. The results of this research can serve as a guide for researchers and practitioners in the field of diabetes prediction and other health applications.

In conclusion, the comparative evaluation of intelligent algorithms shows that informed choices can lead to significant improvements in prediction performance, thus helping to better prevent and manage chronic diseases.

References

[1] L. Breiman, “Random Forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[3] Kaggle, “Pima Indians Diabetes Database.” [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database. [Accessed: Jan. 2024].

[4] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in Proc. Int. Conf. Learn. Represent., 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[5] C. M. Tan and C. Eswaran, “A comprehensive survey of machine learning techniques in medical diagnosis,” IEEE Access, vol. 8, pp. 32181–32195, 2020, doi: 10.1109/ACCESS.2020.2973489.

[6] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2023, pp. 785–794, doi: 10.1145/2939672.2939785.

[7] H. T. Nguyen and H. Q. Tran, "Enhancing Diabetes Prediction Using Machine Learning Techniques: A Comparative Study," Comput. Intell., vol. 39, no. 2, pp. 493–509, 2023, doi: 10.1002/coin.2323.

[8] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 3rd ed. Springer, 2022, doi: 10.1007/978-0-387-84858-7.

[9] Y. Li and X. Liu, "Predicting Diabetes Using Advanced Machine Learning Algorithms," J. Med. Syst., vol. 46, no. 5, pp. 54–67, 2022, doi: 10.1007/s10916-022-01857-9.

[10] S. Bai and Q. Zhang, "An Improved Random Forest Algorithm for Diabetes Prediction," Appl. Sci., vol. 13, no. 4, p. 2301, 2023, doi: 10.3390/app13042301.

[11] J. J. K. and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICTE, Feb. 2021, doi: 10.1016/j.icte.2021.02.004.