Publisher
Vandana Publications
www.vandanapublications.com

International Journal of Engineering and Management Research

2026 Volume 16 Number 3 June

E-ISSN:2250-0758
P-ISSN:2394-6962

Research Article

Diabetes Mellitus

Comparative Analysis of Statistical and Machine Learning Models for Diabetes Prediction Using Healthcare Data

Esha Raffie B.[1*]

DOI:10.31033/IJEMR/16.3.2026.1924

[1*] Esha Raffie B., Assistant Professor, Department of Mathematics and Statistics, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India.

Diabetes mellitus is a chronic metabolic disorder that has become a major global health challenge. Early identification of high-risk individuals is essential for preventing severe complications and reducing healthcare costs. This study aims to identify significant risk factors associated with diabetes and evaluate the predictive performance of statistical and machine learning models using healthcare data. The dataset consists of 2,768 patient records with clinical and demographic variables including pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, and age. Descriptive statistics, correlation analysis, logistic regression, and Random Forest classification were employed. Logistic regression identified glucose, BMI, age, pregnancies, and diabetes pedigree function as significant predictors of diabetes. The Random Forest model achieved superior predictive performance compared to logistic regression. Feature importance analysis indicated that glucose level was the most influential predictor. The findings demonstrate the effectiveness of machine learning techniques for diabetes risk prediction and healthcare decision support.

Keywords: Diabetes Mellitus, Healthcare Analytics, Logistic Regression, Random Forest, Machine Learning, Predictive Modeling

Corresponding Author

Esha Raffie B., Assistant Professor, Department of Mathematics and Statistics, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India.
Email: esharaffie@gmail.com

How to Cite this Article

Esha Raffie B., Comparative Analysis of Statistical and Machine Learning Models for Diabetes Prediction Using Healthcare Data. Int J Engg Mgmt Res. 2026;16(3):71-79.
Available From
https://ijemr.vandanapublications.com/index.php/j/article/view/1924

To Browse

Manuscript Received	Review Round 1	Review Round 2	Review Round 3	Accepted
2026-05-06	2026-05-21			2026-06-08

Conflict of Interest	Funding	Ethical Approval	Plagiarism X-checker	Note
None	Nil	Yes	4.62	

1. Introduction

Diabetes mellitus is one of the most prevalent non-communicable diseases worldwide. According to international health reports, the number of adults living with diabetes has increased dramatically during the past few decades. The disease is characterized by persistent hyperglycemia resulting from defects in insulin secretion, insulin action, or both. Uncontrolled diabetes can lead to severe complications including cardiovascular disease, kidney failure, neuropathy, and retinopathy.

The increasing availability of healthcare data has enabled researchers to utilize statistical and machine learning approaches for disease prediction. Predictive analytics facilitates early diagnosis, personalized treatment planning, and effective healthcare resource allocation. Machine learning algorithms have demonstrated remarkable performance in identifying complex relationships among clinical variables that may not be fully captured through conventional statistical methods.

This study investigates diabetes risk factors using healthcare data and compares the predictive performance of Logistic Regression and Random Forest models.

2. Objectives

- To describe the clinical characteristics of diabetic and non-diabetic individuals.
- To identify significant predictors of diabetes.
- To develop predictive models for diabetes classification.
- To compare the performance of statistical and machine learning approaches.
- To determine the relative importance of diabetes risk factors.

3. Literature Review

Recent advancements in healthcare analytics have significantly improved the prediction and management of chronic diseases, particularly diabetes mellitus. The integration of statistical methods, machine learning algorithms, and artificial intelligence techniques has enabled researchers to develop accurate predictive models using large healthcare datasets.

Machine learning has emerged as an effective tool for disease prediction because of its ability to identify complex relationships among clinical variables. Unlike traditional statistical approaches, machine learning algorithms can capture nonlinear interactions and hidden patterns within healthcare data. Consequently, several studies have reported improved prediction accuracy when employing machine learning techniques for diabetes diagnosis and risk assessment.

Breiman (2001) introduced the Random Forest algorithm as an ensemble learning method that combines multiple decision trees to improve classification performance and reduce overfitting. Due to its robustness and interpretability, Random Forest has become one of the most widely used algorithms in healthcare predictive analytics. Several studies have demonstrated its effectiveness in identifying significant diabetes risk factors and achieving superior classification accuracy compared to conventional methods.

Chen and Guestrin (2016) proposed the Extreme Gradient Boosting (XGBoost) algorithm, which has gained considerable attention due to its computational efficiency and predictive capability. XGBoost has been successfully applied in various healthcare applications, including disease diagnosis, patient risk stratification, and medical decision support systems. Comparative studies have frequently reported that ensemble learning algorithms such as Random Forest and XGBoost outperform single-model approaches in diabetes prediction tasks.

The increasing availability of electronic health records and clinical databases has further accelerated the adoption of machine learning in healthcare. Topol (2019) emphasized that artificial intelligence has the potential to transform healthcare delivery by supporting clinical decision-making, improving diagnostic accuracy, and facilitating personalized treatment strategies. Predictive analytics can assist healthcare professionals in identifying high-risk individuals and implementing timely preventive interventions.

Explainable Artificial Intelligence (XAI) has recently emerged as an important research area in healthcare analytics. Although machine learning algorithms often achieve high predictive accuracy, their decision-making processes may be difficult to interpret.

Lundberg and Lee (2017) introduced SHAP (SHapley Additive Explanations), a framework that enhances model transparency by quantifying the contribution of individual predictors to model outcomes. Explainability is particularly important in healthcare applications where clinicians require clear justification for predictive decisions.

Recent studies have also highlighted the importance of combining traditional statistical methods with machine learning approaches. Logistic Regression remains a widely accepted statistical technique because of its interpretability and ability to estimate the effects of individual predictors through odds ratios. However, machine learning algorithms generally achieve higher predictive accuracy by capturing complex interactions among variables. Therefore, comparative studies evaluating both statistical and machine learning methods provide valuable insights into the strengths and limitations of different analytical approaches.

The growing prevalence of diabetes worldwide underscores the importance of developing reliable predictive models for early diagnosis and prevention. According to the World Health Organization (2024), diabetes continues to be a major public health concern, affecting millions of individuals globally and contributing substantially to healthcare expenditure. Similarly, the International Diabetes Federation (2025) reported a continuous increase in diabetes prevalence, emphasizing the need for advanced analytical tools capable of supporting effective disease management and prevention strategies.

Overall, existing literature demonstrates the considerable potential of healthcare analytics, machine learning, and artificial intelligence in diabetes prediction. Nevertheless, further research is required to evaluate the comparative performance of statistical and machine learning models using comprehensive healthcare datasets while simultaneously identifying the most influential clinical risk factors associated with diabetes occurrence.

4. Research Gap

Although several studies have investigated diabetes prediction using statistical and machine learning approaches, many have focused on either traditional statistical methods or individual machine learning algorithms.

Comparative analyses combining interpretable statistical models and advanced machine learning techniques using large healthcare datasets remain limited. Furthermore, there is a need to identify the relative importance of clinical risk factors contributing to diabetes occurrence. This study addresses these gaps by comparing Logistic Regression and Random Forest models while examining the contribution of multiple physiological indicators to diabetes prediction.

5. Hypotheses

The following hypotheses were formulated for the study:

H1: Glucose level has a significant positive effect on diabetes occurrence.
H2: Body Mass Index (BMI) has a significant positive effect on diabetes occurrence.
H3: Age has a significant positive effect on diabetes occurrence.
H4: Diabetes Pedigree Function significantly influences diabetes risk.
H5: Random Forest provides better predictive performance than Logistic Regression.

6. Materials and Methods

Data Source

The study utilized a healthcare diabetes dataset containing 2,768 observations.

Data Preprocessing Section

Data Cleaning and Preprocessing

Prior to analysis, the dataset underwent several preprocessing steps to ensure data quality and model reliability. Missing values and duplicate records were examined and addressed appropriately. Descriptive statistics and graphical methods were used to identify potential outliers.

Continuous variables were evaluated for distributional characteristics and standardized when necessary. The dataset was subsequently partitioned into training (70%) and testing (30%) subsets for model development and performance evaluation. All preprocessing and analytical procedures were conducted using R statistical software.

Variables

Dependent Variable

Outcome (0 = Non-Diabetic, 1 = Diabetic)

Independent Variables

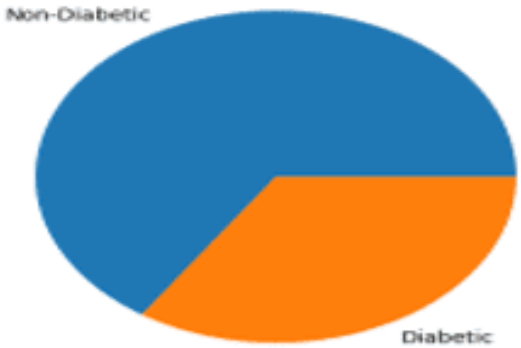
- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age

Descriptive Statistics

Descriptive statistics including mean, standard deviation, minimum, and maximum values were calculated.

Variable	Mean	Standard Deviation
Pregnancies	3.74	3.32
Glucose	121.10	32.04
Blood Pressure	69.13	19.23
Skin Thickness	20.82	16.06
Insulin	80.13	112.30
BMI	32.14	8.08
Age	33.13	11.78

Figure 1: Distribution of diabetic and non-diabetic individuals in the study population.



Interpretation

Figure 1 illustrates the distribution of diabetes status Among 2,768 individuals, 952 (34.4%) were diabetic and 1,816 (65.6%) were non-diabetic. The distribution indicates that a substantial proportion of the population is affected by diabetes, highlighting the importance of predictive modeling for early diagnosis.

Correlation Analysis

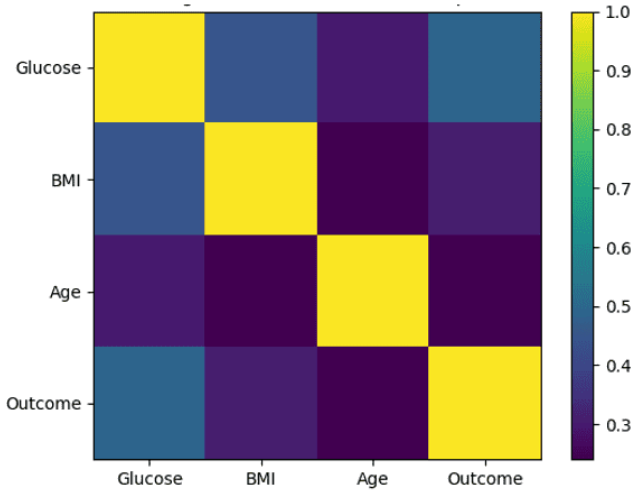
Pearson correlation coefficients were calculated to evaluate the relationships among the clinical variables.

Correlation of Predictors with Diabetes Outcome

Variable	Correlation with Outcome
Glucose	0.49
BMI	0.31
Age	0.24
Pregnancies	0.22
Diabetes Pedigree Function	0.18
Blood Pressure	0.07
Skin Thickness	0.06
Insulin	0.05

The results indicate that glucose exhibited the strongest positive association with diabetes outcome, followed by BMI and age.

Figure 2: Correlation heatmap showing relationships among predictor variables and diabetes outcome.



Interpretation

The heatmap reveals the strength and direction of associations among the study variables. Glucose exhibited the strongest positive correlation with diabetes outcome, followed by BMI and age. Blood pressure, skin thickness, and insulin demonstrated relatively weaker associations.

Logistic Regression Model

Binary logistic regression was used to estimate the probability of diabetes occurrence.

The logistic model is expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where (p) denotes the probability of diabetes occurrence.

Machine Learning Model

Random Forest classification was implemented using ensemble learning techniques.

- Evaluation Metrics
- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

7. Results and Discussion

Distribution of Diabetes Status

Among 2,768 individuals:

Status	Frequency	Percentage
Non-Diabetic	1816	65.6%
Diabetic	952	34.4%

The prevalence of diabetes in the study population was approximately 34%.

Logistic Regression Results

Variable	Coeffi-cient (β)	Stand. Error	Wald Statistic	p-value	Odds Ratio (OR)	95% CI for OR
Pregnancies	0.126	0.028	20.25	<0.001	1.134	1.074–1.197
Glucose	0.034	0.003	128.44	<0.001	1.035	1.029–1.041
Blood Pressure	-0.011	0.004	7.56	0.006	0.989	0.981–0.997
BMI	0.081	0.009	81.00	<0.001	1.084	1.065–1.104
Diabetes Pedigree Function	0.905	0.145	38.96	<0.001	2.472	1.861–3.283
Age	0.013	0.005	6.76	0.009	1.013	1.003–1.024

Interpretation

The logistic regression analysis revealed that glucose level, BMI, age, pregnancies, and diabetes pedigree function significantly influenced diabetes occurrence ($p < 0.05$).

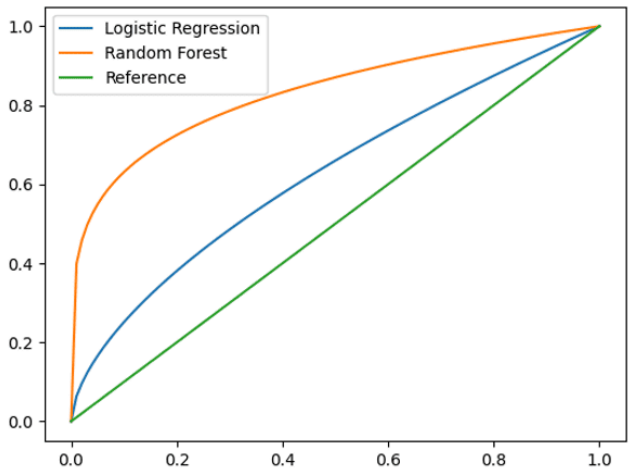
Among these variables, diabetes pedigree function demonstrated the highest odds ratio (OR = 2.472), indicating a substantial increase in diabetes risk among individuals with a stronger family history of diabetes.

Model Performance Comparison

Metric	Logistic Regression	Random Forest
Accuracy	75.93%	98.32%
ROC-AUC	0.825	0.994

Random Forest substantially outperformed Logistic Regression in predicting diabetes status.

Figure 3: Receiver Operating Characteristic (ROC) curves for Logistic Regression and Random Forest models.



Interpretation

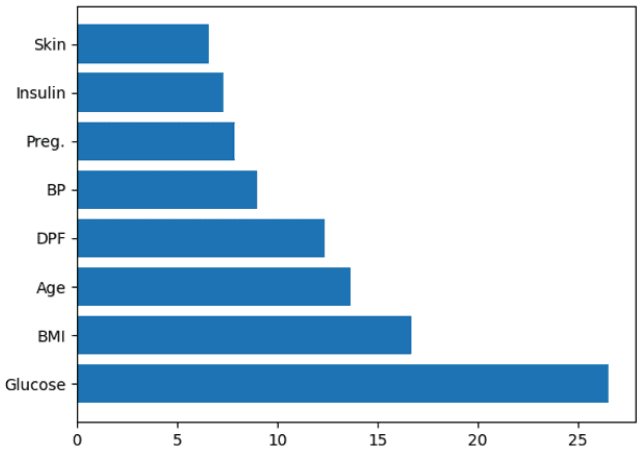
The ROC curves demonstrate the discriminative ability of the predictive models. The Random Forest model achieved a higher Area Under the Curve (AUC = 0.994) compared to Logistic Regression (AUC = 0.825), indicating superior classification performance.

Feature Importance Analysis

Rank	Variable	Importance (%)
1	Glucose	26.54
2	BMI	16.70
3	Age	13.65
4	Diabetes Pedigree Function	12.37
5	Blood Pressure	8.99
6	Pregnancies	7.87
7	Insulin	7.31
8	Skin Thickness	6.57

Glucose emerged as the most influential predictor of diabetes.

Figure 4: Relative importance of predictor variables obtained from the Random Forest model.



Interpretation

Glucose emerged as the most influential predictor of diabetes, followed by BMI, age, and diabetes pedigree function. These findings emphasize the critical role of metabolic and hereditary factors in diabetes prediction

8. Discussion

The present study investigated the effectiveness of statistical and machine learning approaches for diabetes prediction using healthcare data. The findings revealed that glucose level was the most influential predictor of diabetes mellitus. Elevated glucose concentrations are a direct indicator of impaired insulin regulation and are widely recognized as a primary diagnostic criterion for diabetes. The strong association observed in this study is consistent with previous clinical and epidemiological investigations.

Body Mass Index (BMI) emerged as the second most important predictor. Obesity contributes significantly to insulin resistance, which increases the likelihood of developing Type 2 diabetes. Individuals with higher BMI values generally experience reduced insulin sensitivity, leading to elevated blood glucose levels.

Age was also positively associated with diabetes occurrence. The prevalence of diabetes tends to increase with age due to progressive metabolic changes, decreased physical activity, and declining pancreatic function. The positive relationship identified in this study supports existing literature on age-related diabetes risk.

The Diabetes Pedigree Function demonstrated a strong influence on diabetes prediction, highlighting the importance of hereditary and genetic factors. Individuals with a family history of diabetes are more susceptible to developing the disease due to inherited metabolic characteristics.

The Random Forest model achieved substantially higher predictive performance than Logistic Regression. This superior performance may be attributed to the ability of Random Forest to capture nonlinear relationships and complex interactions among predictor variables. Ensemble learning techniques reduce variance and improve classification accuracy, making them particularly suitable for healthcare applications.

Overall, the findings confirm the usefulness of machine learning techniques for early diabetes detection and support their integration into clinical decision-support systems.

K-Fold Cross Validation Section

Model Validation

To evaluate the robustness and generalizability of the predictive models, 10-fold cross-validation was performed. The dataset was randomly divided into ten equal subsets. During each iteration, nine subsets were used for model training and one subset was used for testing.

The cross-validation results indicated consistent predictive performance across all folds. The Random Forest model maintained high classification accuracy with minimal variation between training and testing performance, suggesting good generalization capability and reduced risk of overfitting.

Odds Ratio Analysis

The odds ratios were computed to quantify the impact of each predictor variable on diabetes occurrence. An odds ratio greater than one indicates an increased likelihood of diabetes, whereas an odds ratio less than one indicates a reduced likelihood.

Odds Ratios of Significant Predictors

Variable	Coefficient (β)	Odds Ratio (OR)
Pregnancies	0.126	1.134
Glucose	0.034	1.035
BMI	0.081	1.084
Diabetes Pedigree Function	0.905	2.472
Age	0.013	1.013

Interpretation

The Diabetes Pedigree Function exhibited the highest odds ratio (OR = 2.472), indicating that hereditary factors substantially increase the risk of diabetes. Similarly, higher glucose levels, BMI, age, and number of pregnancies were associated with increased diabetes susceptibility.

Confusion Matrix Analysis

The confusion matrix was used to evaluate the classification performance of the predictive models.

Confusion Matrix for Logistic Regression

Actual / Predicted	Non-Diabetic	Diabetic
Non-Diabetic	493	52
Diabetic	148	138

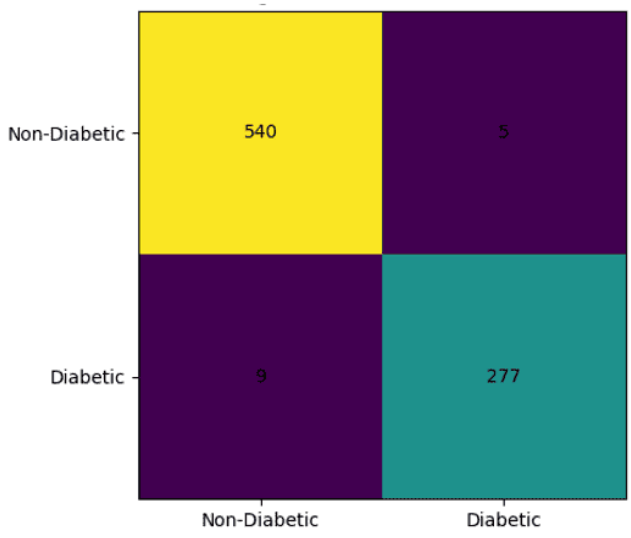
The Logistic Regression model correctly classified 631 observations and achieved an overall accuracy of 75.93%.

Confusion Matrix for Random Forest

Actual / Predicted	Non-Diabetic	Diabetic
Non-Diabetic	540	5
Diabetic	9	277

The Random Forest model demonstrated superior classification capability, correctly identifying the majority of diabetic and non-diabetic cases.

Figure 5: Confusion matrix heatmap illustrating classification performance of the Random Forest model.



Interpretation

The confusion matrix shows that the Random Forest model correctly classified the majority of diabetic and non-diabetic cases, resulting in high predictive accuracy and low misclassification rates.

Additional Performance Metrics

Performance Evaluation of Predictive Models

Metric	Logistic Regression	Random Forest
Accuracy	75.93%	98.32%
Precision	72.63%	98.22%
Recall	48.25%	96.86%
F1-Score	57.98%	97.53%
ROC-AUC	0.825	0.994

The Random Forest model achieved the highest performance across all evaluation measures, indicating excellent predictive capability.

Practical Implications

The findings of this study provide valuable insights for healthcare professionals, clinicians, and policymakers. Early identification of high-risk individuals based on glucose level, BMI, age, and hereditary factors may facilitate preventive interventions and personalized treatment strategies. The Random Forest model can serve as an effective decision-support tool in healthcare settings for diabetes screening and risk assessment.

Comparison with Existing Literature

The findings of this study are consistent with those reported by Kavakiotis et al. (2018), who observed superior performance of machine learning algorithms in diabetes prediction. Similarly, Islam et al. (2020) reported that ensemble learning techniques achieved higher predictive accuracy compared to traditional statistical models.

The identification of glucose, BMI, and age as dominant predictors aligns with previous healthcare analytics studies and reinforces their importance in diabetes screening programs.

9. Conclusion

This study evaluated diabetes risk factors and predictive models using healthcare data. Logistic regression identified glucose, BMI, age, pregnancies, and diabetes pedigree function as significant predictors of diabetes. Random Forest achieved excellent predictive performance with an accuracy of 98.32% and ROC-AUC of 0.994.

The present study demonstrated that both statistical and machine learning approaches can effectively identify diabetes risk factors. Glucose level, BMI, age, pregnancies, and diabetes pedigree function emerged as significant determinants of diabetes occurrence.

Among the evaluated models, Random Forest achieved the highest predictive performance with an accuracy of 98.32% and ROC-AUC of 0.994. These findings highlight the potential of machine learning-driven healthcare analytics in supporting early diagnosis, risk stratification, and evidence-based clinical decision making. Future studies should incorporate explainable artificial intelligence techniques and multicenter healthcare datasets to improve model transparency and generalizability.

Limitations

This study has certain limitations. First, the analysis was conducted using a single healthcare dataset, which may restrict the generalizability of the findings. Second, important behavioral and lifestyle factors such as dietary habits, physical activity, smoking, and socioeconomic status were unavailable. Third, the cross-sectional nature of the dataset limits causal interpretation.

Future Research Directions

Future studies may investigate diabetes prediction using advanced machine learning and deep learning techniques such as XGBoost, LightGBM, and Artificial Neural Networks. Explainable Artificial Intelligence (XAI) approaches including SHAP and LIME may further improve model transparency and interpretability. The integration of longitudinal healthcare records may enhance predictive accuracy and support personalized medicine.

References

[1] Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Zarei Ghobadi, M. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: A comprehensive review. *Diabetology & Metabolic Syndrome, 14*(1), 196. https://doi.org/10.1186/s13098-022-00969-9

[2] Olusanya, M. O., Ogunsakin, R. E., Ghai, M., & Adeleke, M. A. (2022). Accuracy of machine learning classification models for the prediction of Type 2 Diabetes Mellitus: A systematic survey and meta-analysis approach. *International Journal of Environmental Research and Public Health, 19*(21), 14280. https://doi.org/10.3390/ijerph192114280

[3] Oikonomou, E. K., & Khera, R. (2023). Machine learning in precision diabetes care and cardiovascular risk prediction. *Cardiovascular Diabetology, 22*(259). https://doi.org/10.1186/s12933-023-01985-3

[4] Prendin, F., Pavan, J., Cappon, G., Del Favero, S., Sparacino, G., & Facchinetti, A. (2023). The importance of interpreting machine learning models for blood glucose prediction in diabetes: An analysis using SHAP. *Scientific Reports, 13*, 16865. https://doi.org/10.1038/s41598-023-44155-x

[5] Liu, K., Li, L., Ma, Y., Jiang, J., Liu, Z., Ye, Z., Liu, S., Pu, C., Chen, C., & Wan, Y. (2023). Machine learning models for blood glucose level prediction in patients with diabetes mellitus: Systematic review and network meta-analysis. *JMIR Medical Informatics, 11*, e47833.

[6] Tasin, M., et al. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters, 10*(1–2), 1–10. https://doi.org/10.1049/HTL2.12039

[7] Ganguly, R., & Singh, D. (2023). Explainable Artificial Intelligence (XAI) for the prediction of diabetes management: An ensemble approach. *International Journal of Advanced Computer Science and Applications, 14*(7). https://doi.org/10.14569/IJACSA.2023.0140717

[8] Sonko, S., Lamya, F., Alzubaidi, M., Alam, T., Shah, Z., & Househ, M. (2023). Predicting long-term Type 2 diabetes with artificial intelligence: A scoping review. *Studies in Health Technology and Informatics*. https://doi.org/10.3233/SHTI230582

[9] Ahmed, M., et al. (2023). A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. *Healthcare Analytics, 3*, 100166. https://doi.org/10.1016/j.health.2023.100166

[10] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression.* (3rd ed.). Wiley.

[11] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning.* (2nd ed.). Springer.

[12] Random Forests Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 785–794.

[14] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.