Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol.10 No. 2

Advanced Predictive Analysis of Diabets Using Axg Boosting Algorithm

K. Emayavaramban[1)] and Thambusamy Velmurugan[2)]*

[1)]LoganathaNarayanasamy Government College, Ponneri, India
[2)]Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Chennai, India

Abstract. the global healthcare landscape, early detection of chronic diseases plays a crucial role in improving patient outcomes and reducing long-term complications. This study focuses on diabetes prediction and risk classification, aiming to identify patients as high, moderate, or low risk based on clinical and demographic features. Traditional machine learning algorithms have been widely used for this task, but they often fall short in prediction accuracy. To address this, we propose a novel ensemble learning method called AXG boosting algorithm, which outperforms conventional algorithms by achieving a 5% improvement in predictive performance and its supports both supervised and unsupervised. The dataset used in this study is sourced from the Kaggle repository and includes key features relevant to diabetes diagnosis. Feature selection techniques are applied to identify the most significant attributes, and the dataset is split into 80% for training and 20% for testing. Experimental results demonstrate the effectiveness of the proposed AXG Boosting algorithm in enhancing risk classification for diabetic patients.

Keywords; Information Boosting algorithm, Machine learning algorithm, AXGB algorithm, diabetes

Cite this paper as: K. Emayavaramban and Thambusamy Velmuruga (2026) "Advanced Predictive Analysis Of Diabets Using Axg Boosting Algorithm", Journal of Industrial Information Technology and Application, Vol. 10. No. 2, pp. 1296-1307

[*]Corresponding author : emayammsc2002@gmail.com
Received: Apr. 9. 2026 Accepted: May. 23. 2026 Published: Jun. 30. 2026

1. Introduction

Diabetes mellitus is a chronic metabolic disorder impacting millions of individuals worldwide and is increasingly recognized as a significant public health issue. Diabetes is marked by high blood glucose levels and can result in serious complications, including kidney failure, cardiovascular disease, nerve damage (neuropathy), and vision impairment if not diagnosed and managed promptly. The increasing prevalence of diabetes, particularly in low- and middle-income nations, underscores the critical importance of early detection and ongoing monitoring to enhance patient outcomes and minimize healthcare expenses. In contemporary clinical settings, extensive patient data is gathered through wearable devices, laboratory tests, and electronic health records (EHRs). When analyzed effectively, this extensive data can uncover essential insights for predicting and preventing diabetes. Traditional diagnostic methods frequently face limitations due to time constraints, resource availability, and a significant dependence on physician expertise. To address these challenges, machine learning (ML) algorithms have proven to be effective instruments in the healthcare sector. These techniques are capable of automatically identifying and classifying diseases, including diabetes, by analyzing patterns and trends within complex patient datasets. In Machine learning models such as Naïve Bayes, Random Forest, Support Vector Machines (SVM), Logistic Regression, along with clustering algorithms like K-Means and Fuzzy C-Means, have been effectively utilized to predict the likelihood of diabetes by learning from historical patient datasets. These models offer significant advantages, including high accuracy, flexibility, and the ability to handle complex and non-linear relationships commonly found in medical data. Building on this foundation, the proposed AXG Boosting algorithm enhances classification by categorizing patients into high, moderate, andlow-risk groups, enabling personalized treatment strategies and targeted medical interventions. Moreover, AXG Boosting proves valuable even when working with unlabeled medicaldatasets, as it can uncover hidden patterns essential for exploratory analysis and early detection efforts in diabetes care.

2. Related Works

The paper [1] discuss about the prediction of diabetes mellitus using the (IDMPF) intelligent diabetes mellitus prediction frame work. The article [2] deals with predictive value of machine learning models for progression of gestational diabetes

mellitus (GDM) to type2 diabets (T2DM). The research paper [3] discuss about the diabetes prediction across four data sets (PDD, EDD, MDD, BDD) and achieved the good performance in gradient boosting. This article [4] discusses the challenges associated with incomplete diabetes data and presents a knowledge graph dataset that has been validated using link prediction methods. The research article [5] presents the development and assessment of an innovative RFE-GRU model for the classification of early-stage diabetes utilizing the PIMA Indian Diabetes Dataset, showcasing enhanced performance compared to conventional machine learning models. The paper [6]. presents a data-driven methodology for forecasting diabetes-related complications through the analysis of blood parameters and KNN classification, with the objective of facilitating early diagnosis and enhancing disease management.

The paper [7] presents a new data-driven methodology for the early prediction of diabetes-related complications by utilizing patient blood parameters, aimed at enhancing disease management and alleviating the healthcare burden. The article [8] presents the advancement of an innovative deep learning framework utilizing a Convolutional Gated Recurrent Unit (CGRU) model aimed at precise early detection and severity classification of diabetes, showcasing enhanced performance compared to current methodologies. The paper [9] shows the traditional medicine, ophthalmologists and retinal imaging are few, making early screening and identification of diabetic retinal problems like DR and DME difficult. It tests whether ChatGPT-4, a large language model, can be used to create accurate, web-based risk calculators for DR and DME using tabular health checkup data (e.g., medical history and blood tests), allowing non-experts to develop and deploy predictive tools without coding or medical imaging.

The article [10] discusses the complexities involved in effectively predicting diabetes through the application of machine learning models. The analysis assesses the performance of multiple models on an extensive medical dataset, highlighting XGBoost as the most effective, with glycated hemoglobin identified as a significant predictor. The article [11] discuss the development of an attention-based multi-modal deep learning system that predicts Type 2 Diabetes (T2D) complications up to 72 hours in advance, demonstrating high accuracy, low false positives, and strong generalizability across patient populations. The paper [12] addresses the challenge of accurately predicting hospital readmissions for diabetes patients to improve healthcare resource management. The study [13] focus on Healthcare Big Data Analytics and Machine Learning utilizing Apache Spark's MLlib to examine a CDC’s Behavioral Risk Factor Surveillance System(CDC BRFSS) diabetes dataset. Logistic Regression outperforms Naive Bayes in medical decision-making predicted accuracy, showing how ML can provide insights.

The article [14] compares nine clinical indices (including VAI, TyG, HOMA-IR, BMI, LAP, WHtR, TyG-BMI, TyG-WC, and TyG-WHtR) to predict MASLD in Chinese T2DM patients. Logistic regression and ROC curve research show that the lipid accumulation product (LAP) is the best index predictor. The paper [15] deals in validating an immune related diagnostic model for diabetic foot ulcer by identifying key genes and examining their functional enrichment. The article [16] discuss about the development and validation of a deep learning-based model, REDAPM, that predicts depression and anxiety in type 2 diabetes mellitus (T2DM) patients by utilizing regional electronic health record(EHR) data—both structured and unstructured. It emphasizes the model's superior performance in comparison to conventional methods, the significance of temporal clinical data, and the identification of critical predictive features through interpretability analyses.

The research paper [17] shows the comparative analysis of Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) algorithms for the classification of diabetes mellitus patient data. The performance is assessed through standard metrics, including accuracy, precision, recall, and F1-score, leading to the conclusion that SVM demonstrates superior performance compared to KNN and is more effective for the early detection of diabetes within the analyzed dataset. The research paper [18] discuss about Integrating AI-based predictive models (logistic regression, deep learning, gradient boosting, and decision trees) with causal discovery approaches (like LiNGAM) to predict diabetes and find risk factor causal linkages.

The research paper [19] deals two-phase machine learning approach employing ensemble methods and the Pima Indian Diabetes dataset to improve early diabetes prediction using SMOTE and SMO. Bagging decision trees had the highest accuracy, showing that the model addressed class imbalance and improved categorization. The article [20] discuss about the creation of a diabetes prediction model utilizing a range of machine learning and ensemble methods, such as Logistic Regression, SVM, Naïve Bayes, Random Forest, XGBoost, LightGBM, CatBoost, Adaboost, and Bagging. This highlights CatBoost as the most effective method for early detection, supported by performance evaluations using real-world data.

3. Research Methodology

The input taken as a medical dataset from kaggle repository with having different metadata. The meta data parameters like patient name, blood pressure, bmi, height, weight, etc, first we need to convert the data in to lower case, preprocess the data, remove the punctuation, assign the label encoding for the categories low, high,

moderate. Implement the proposed AXGB algorithm with training 60% and testing 40% to improve the performance.

A. Data preprocessing

The process involves transforming raw data into an appropriate format in different techniques. Feature values are typically assigned as either numerical or textual data, each with varying ranges. The dataset is likely to include a range of features, including plasma glucose levels and BMI ranges, among others. Identifying missing or null values in the dataset is a crucial step. Filter those null values in the dataset to clean the data according to the model.

- Converting the data in to lower case

After loading the input data, verify that all textual attribute values are transformed to lower case prior to before carrying out punctuation removal.

- Removing the punctuation

The process of removing punctuation involves unwanted punctuation marks like commas, points, etc. This step is commonly employed in text preprocessing for natural language processing (NLP), data cleaning, or text mining tasks to streamline the text and minimize noise.

- Removing the stop words

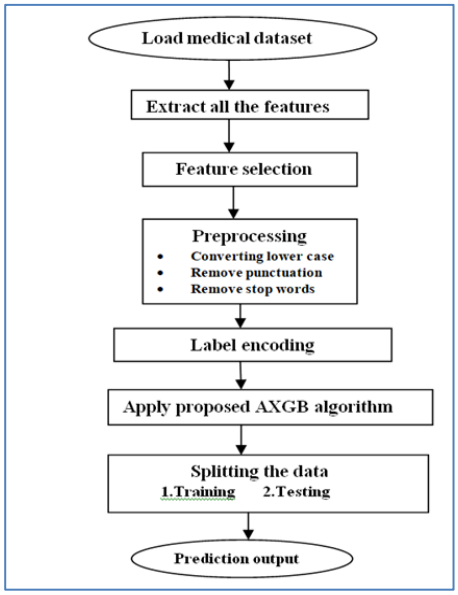
The removal of stop words involves the elimination of common, non-informative words from text data during the preprocessing phase in Natural Language Processing (NLP).

B. Label encoding

It is a technique used to convert the categorical data in to numerical data and assign Label for each category to predict the model. where here to classify the patient category by assigning label low -1, high -2 and moderate -3.

C. Proposed method- AXG Boosting

This proposed method AXG Boosting, Where it support both supervised and unsupervised model to predict the diabetes categories. For supervised data set we assign a Label encoding and For unsupervised data consider the null values to predict the categories.



Algorithm 1 AXGBoost$(I, L, K, \eta, \lambda, \gamma, t_n)$

Input:

- I: a training set $\{(x_i, y_i)\}_{i=1}^{n}$ with feature dimension m
- $L(y, \hat{y})$: a differentiable loss function
- K: number of boosting iterations
- η: the learning rate
- λ: regularization coefficient
- γ: minimum loss reduction required for a split
- t_n: Sum of the tree levels

Initialize the model with a constant value:
$F_0(x) = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$
for $k = 1$ to K **do**
  /* Compute the sum of gradients and Hessians of all the samples */
  $G \leftarrow \sum_{i \in I} \frac{\partial L(y_i, F_{k-1}(x_i))}{\partial F_{k-1}(x_i)}$
  $H \leftarrow \sum_{i \in I} \frac{\partial^2 L(y_i, F_{k-1}(x_i))}{\partial F_{k-1}(x_i)^2}$
  $T \leftarrow$ Build-Tree(I, G, H)
  Let the terminal regions of T be R_j for $j = 1, \ldots, J_k$
  for $j = 1$ to J_k **do**
    $w_j = -\frac{G_j}{H_j + \lambda + 2t_n}$
    where G_j and H_j are the sum of gradients and Hessians at leaf node j
  end for
  $F_k(x) = F_{k-1}(x) + \eta \sum_{j=1}^{J_k} w_j \mathbf{1}\{x \in R_j\}$
end for
Output the final ensemble $F_K(x)$

Figure 1. proposed architecture model

The AXGBoost algorithm is an enhanced version of the XGBoost framework, designed to improve the regularization and control of decision tree boosting. It takes as input a training dataset I = {(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)}, where each x_i is a feature vector and y_i is the corresponding label. The algorithm relies on a differentiable loss function L (y, ŷ), and proceeds over K boosting iterations. Several hyperparameters guide its operation: the learning rate (η) controls how much each tree contributes to the final model; the regularization parameter (λ) penalizes overly complex trees; γ is the minimum loss reduction required for a node to split; and t_n denotes the total depth of trees (sum of levels), which is used for additional structural regularization. The model starts with an initial prediction $F_0(x)$ that minimizes the total loss across the training data. At each iteration k, the algorithm computes the gradients (G) and second-order derivatives (Hessians, H) of the loss with respect to the current prediction $F_{k-1}(x)$. A regression tree is then constructed using these gradients, aiming to reduce the residual errors. Each leaf node of the tree represents a region R_j, and a score w_j is assigned using the formula:

$$w_j = -G_j / (H_j + \lambda + 2t_n) \quad (1)$$

where G_j and H_j are the sums of gradients and Hessians for samples in region R_j. This equation introduces an extra regularization factor ($2t_n$), penalizing deeper trees and helping to control overfitting.

$$F_k(x) = F_{k-1}(x) + \eta \times \Sigma \left(w_j \times I\{x \in R_j\}\right) \quad (2)$$

where $I\{x \in R_j\}$ is an indicator function that equals 1 if x belongs to region R_j, and 0 otherwise. After completing all K iterations, the final prediction function $F_k(x)$ is produced. By incorporating tree depth regularization through t_n, AXGBoost offers better control over model complexity and generalization compared to standard boosting methods.

4. Results and Discussion

The proposed AXG Boosting strategy significantly improves predictive accuracy across six diabetes-related datasets when compared to the current XGBoost method. The accuracy of the suggested AXG Boosting approach for the Diabetic Patient Dataset was 100%, while XGBoost accuracy was 94% proposed AXG Boosting. obtained 99% accuracy in the Pima Indians Diabetes dataset, marginally outperforming XGBoost 95% accuracy.

In the Diabetes Health Indicators Dataset, AXG Boosting achieved 100% accuracy, a significant improvement than XGBoost 87% accuracy. AXG Boosting once more achieved perfect accuracy 100% in the Diabetes Database, surpassing XGBoost 90% accuracy. The dataset Predict Diabetes from Medical Records shows a similar pattern, with AXG Boosting surpassing XGBoost with 100% accuracy compared to 94%. Finally, the most notable improvement was seen in the dataset named "Diabetes" with AXG Boosting, which raised accuracy from 75% with XGBoost to 100%. Overall, our findings unequivocally show that the suggested AXG Boosting approach performs better and is more reliable in diabetes prediction tasks than the current XG Boost algorithm.

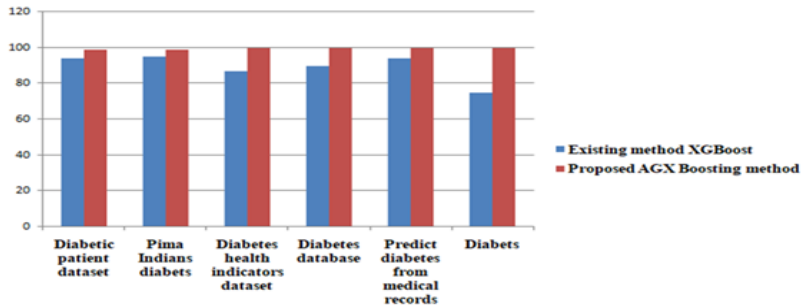


Figure 2. Comparing accuracy Existing vs proposed method

Table 1. Applying dataset in both XGBoost and proposed method AXG Boosting

Sno	Dataset sources	Dataset name	Existing method Xg boost	Proposed AXG boosting %accuracy
1	https://www.kaggle.com/datasets/shantanudhakadd/diabetes-dataset-for-beginners	Diabetic dataset for beginners	94	100
2	https://www.kaggle.com/code/chanchal24/diabetes-dataset-eda-prediction	Pima Indians Diabetes Data	95	99
3	https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook	Diabetes Health Indicators Dataset	87	100
4	https://www.kaggle.com/code/nancyalaswad90/diabetes-database	Diabetes Database	90	100
5	https://www.kaggle.com/code/paultimothymooney/predict-diabetes-from-medical-records	Predict Diabetes From Medical Records	94	100
6	https://www.kaggle.com/code/saramah/diabets-decisiontrees-randomforest	Diabetes	75	100

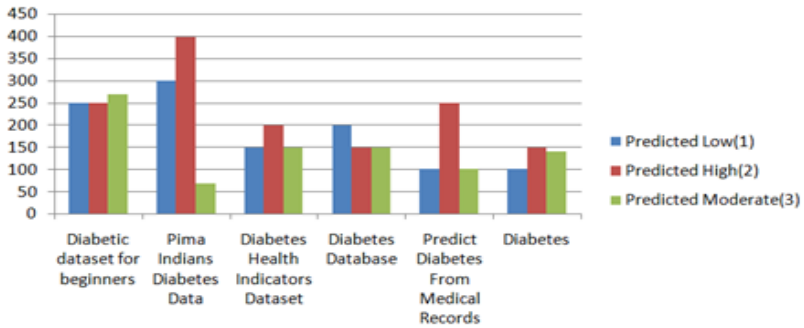


Figure3. Diabetes in Prediction dataset

Table 2. Prediction results for different diabetes-related datasets

Sno	Dataset Name	Records	Predicated		
			Low(1)	High(2)	Moderate(3)
1	Diabetic dataset for beginners	769	250	250	269
2	Pima Indians Diabetes Data	768	300	400	68
3	Diabetes Health Indicators Dataset	500	150	200	150
4	Diabetes Database	500	200	150	150
5	Predict Diabetes From Medical Records	450	100	250	100
6.	Diabetes	390	100	150	140

In this study, multiple diabetes-related datasets were used to evaluate the performance of the proposed prediction technique in classifying the risk levels of diabetes. The datasets varied in size, with the number of records ranging from 390 to 769. Each dataset was classified into three categories: Low Risk (1), High Risk (2), and Moderate Risk (3). For example, in the "Diabetic Dataset for Beginners" containing 769 records, the model predicted 250 records as Low Risk, 250 as High Risk, and 269 as Moderate Risk. Similarly, the "Pima Indians Diabetes Data" with 768 records resulted in 300 Low Risk, 400 High Risk, and 68 Moderate Risk predictions. Other datasets showed different distributions; the "Diabetes Health Indicators Dataset" (500 records) had an almost balanced prediction across the three categories, while the "Predict Diabetes From Medical Records" dataset (450 records) showed a higher number of High Risk predictions (250) compared to Low and Moderate Risk (each 100). These results demonstrate the variability in risk prediction outcomes across diverse datasets, highlighting the adaptability of the proposed technique in classifying diabetes risk levels in various medical data scenarios.

4.1 **Evaluation metrics**

The AXG(Advanced Extreme Gradient) Boosting Performance Metrics Dataset provides the evaluation results of the AXG Boosting algorithm as applied to a range of diabetes-related datasets. This analysis encompasses three essential performance metrics: Precision, Recall, and F1 Score.

Table 3. Performance metrics for AXG Boosting

Sno	Dataset name	Precision	Recall	F1 score
1	Diabetic patient dataset	100	100	100
2	Pima Indians Diabetes Data	99.6	99.4	98.8
3	Diabetes Health Indicators Dataset	100	100	100
4	Diabetes Database	100	100	100
5	Predict Diabetes From Medical Records	100	100	100
6	Diabetes	100	100	100

These metrics are widely utilized to evaluate classification models, especially within the medical field, where the consequences of false positives and false negatives can be significant. The datasets assessed comprise the Diabetic Patient Dataset, Pima Indians Diabetes Data, Diabetes Health Indicators Dataset, Diabetes Database, Predict Diabetes from Medical Records, and a general Diabetes dataset. The results demonstrate that AXG Boosting attained exceptional performance (100%) across the majority of datasets in all three metrics, reflecting a high level of accuracy in predictions. The sole exception is the Pima Indians Diabetes dataset, where the algorithm attained marginally lower values—99.6% precision, 99.4% recall, and 98.8% F1 score—indicating it presented a somewhat greater challenge for the model, potentially due to its more balanced class distribution or heightened variability. The results indicate that AXG Boosting is effective in identifying diabetic conditions across various datasets; however, it is essential to ensure that this high performance is not due to overfitting, particularly when metrics achieve 100%.

Table 4. Micro and Macro average metrics

Metrics	Micro average	Macro average
Precision	99.93	99.93
Recall	99.90	99.90
F1 score	99.80	99.80

The micro average consolidates performance across all predictions, treating each instance with equal importance, whereas the macro average considers each dataset equally, irrespective of its size. The closely aligned values of micro and macro averages, all exceeding 99%, demonstrate a consistent level of performance across all datasets. The AXG Boosting algorithm exhibits remarkable accuracy in distinguishing between diabetic and non-diabetic cases, showing minimal performance variation across different datasets. The metrics validate the strength and dependability of AXG Boosting in managing a variety of diabetes-related medical data.

5. Conclusion

The Collection of dataset is taken from the kaggle repository implemented in XGboost and proposed AXG boosting, The proposed AXG Boosting method demonstrates superior performance compared to the existing XGBoost algorithm across all assessed diabetes-related datasets. Consistent improvements in accuracy have been observed, with the proposed method attaining near-perfect or perfect accuracy levels ranging from 99% to 100%, in contrast to the existing method lower accuracy range of 75% to 95%. This illustrates the strength and efficiency of AXG Boosting in improving predictive performance for diabetes diagnosis and easy to classify the patients categorize for predicting a diabetes mellitus.

Reference

[1] Ismail, Leila, and HunedMaterwala. "IDMPF: intelligent diabetes mellitus prediction framework using machine learning." Applied Computing and Informatics 21.1/2 (2025): 78-89.

[2] Zhao, Meng, et al. "Predictive value of machine learning for the progression of gestational diabetes mellitus to type 2 diabetes: a systematic review and meta-analysis." BMC Medical Informatics and Decision Making 25.1 (2025): 18.

[3] GA, Yashaswini. "Assessing the predictive power of boosting techniques for diabetes." Multimedia Tools and Applications (2025): 1-25.

[4] Singh, Sushmita, and ManviSiwach. "Evaluating diabetes dataset for knowledge graph embedding based link prediction." Data & Knowledge Engineering (2025): 102414.

[5] Shams, Mahmoud Y., Zahraa Tarek, and Ahmed M. Elshewey. "A novel RFE-GRU model for diabetes classification using PIMA Indian dataset." Scientific Reports 15.1 (2025): 982.

[6] Prabakar, T. N., B. Shadaksharappa, and P. Ramkumar. "Prediction of Diabetic Diseases Using Data Analysis." International journal of health sciences 6.S1 (2030): 10013-10021.

[7] Khurshid, Muhammad Rizwan, et al. "Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction." PloS one 20.1 (2025): e0310218.

[8] Alsayed AO, Ismail NA, Hasan L, Binsawad M, Embarak F. 2025. Leveraging a hybrid convolutional gated recursive diabetes prediction and severity grading model through a mobile app. PeerJ Computer Science 11:e2642 https://doi.org/10.7717/peerj-cs.2642

[9] Choi, Eun Young, Joon Yul Choi, and Tae KeunYoo. "Automated and code-free development of a risk calculator using ChatGPT-4 for predicting diabetic retinopathy and macular edema without retinal imaging." International Journal of Retina and Vitreous 11.1 (2025): 11.

[10] Sun, Qi, et al. "Machine learning-based assessment of diabetes risk." Applied Intelligence 55.2 (2025): 1-13.

[11] Xiong, Ke, et al. "A Multi-modal Deep Learning Approach for Predicting Type 2 Diabetes Complications: Early Warning System Design and Implementation." Journal of Theory and Practice of Engineering Science 5.1 (2025): 33-45.

[12] García-Mosquera, Jorge, et al. "Transformer-Based Prediction of Hospital Readmissions for Diabetes Patients." Electronics 14.1 (2025): 174.

[13] Nauman, Muhammad, et al. "The Role of Big Data Analytics in Revolutionizing Diabetes Management and Healthcare Decision-Making." IEEE Access (2025).

[14] Hu, Mengmeng, et al. "Prediction of MASLD using different screening indexes in Chinese type 2 diabetes mellitus." Diabetology& Metabolic Syndrome 17.1 (2025): 10.

[15] Gao, Hengkun, et al. "Elaboration and verification of immune-based diagnostic biomarker panel for diabetic foot ulcer." Journal of Diabetes and its Complications (2025): 108957.

[16] Feng, Wei, et al. "Deep Learning Based Prediction of Depression and Anxiety in Patients with Type 2 Diabetes Mellitus Using Regional Electronic Health Records." International Journal of Medical Informatics (2025): 105801.

[17] Habibi, Ahmad Rizky Nusantara, et al. "Diabetes Mellitus Disease Analysis using Support Vector Machines and K-Nearest Neighbor Methods." Indonesian Journal of Modern Science and Technology 1.1 (2025): 22-27.

[18] Noh, Mi Jin, and Yang Sok Kim. "Diabetes Prediction Through Linkage of Causal Discovery and Inference Model with Machine Learning Models." Biomedicines 13.1 (2025): 124.

[19] Talari, Praveen, et al. "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2." Plos one 19.1 (2024): e0292100.

[20] Modak, Sandip Kumar Singh, and Vijay Kumar Jha. "Diabetes prediction model using machine learning techniques." Multimedia Tools and Applications 83.13 (2024): 38523