Early Prediction of Diabetes Using Logistic Regression

Gopal Samy B[1], Harish M[2], Kanagalakshmi S[3], Swathi S[4]

[1,2,3,4]Department of Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract:

Accurate prediction of blood sugar variations is essential for effective management of diabetes, particularly in avoiding severe complications such as hyperglycemia and hypoglycemia. While advanced machine learning methods, especially neural networks, have been thoroughly studied for forecasting glucose levels, their complexity and high computational demands may limit their practical use in healthcare settings. This study explores the feasibility of employing a logistic regression model as a simple and cost-effective solution for identifying short-term blood glucose risks. By combining continuous glucose monitoring (CGM) data with relevant clinical information, we developed a logistic regression model to predict glycemic risk events over defined future periods. The dataset was subject to standard preprocessing procedures, including addressing missing data through imputation and normalizing the input features. The model was trained and validated using a labeled dataset, with its performance assessed using metrics such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). The findings showed that the logistic regression model exhibited encouraging predictive capabilities, achieving an AUC of approximately 0.80 for short-term prediction intervals. However, a slight decline in performance was observed when the prediction horizons were lengthened, highlighting the challenges of long-term forecasting. In conclusion, the findings suggest that logistic regression is an effective and interpretable approach for the immediate classification of blood glucose risks, providing a practical option compared to more complex deep learning methods. This approach could facilitate timely clinical actions, enhance patient safety, and support real-time decision-making in diabetes management.

Keywords: Diabetes management, Glycemic risk prediction, Logistic regression, Continuous glucose monitoring (CGM), Blood glucose forecasting, Data preprocessing, Predictive modeling, Short-term prediction

1. Introduction:

Diabetes mellitus is a long-term metabolic condition characterized by high blood sugar levels resulting from problems with insulin production, its action, or both. Among the various types, Type 1 Diabetes (T1D) is a unique autoimmune disorder where the body's immune system attacks and destroys the β-cells in the pancreas that produce insulin. In contrast to Type 2 Diabetes (T2D), which is commonly linked to lifestyle choices and insulin resistance, T1D usually appears in childhood and necessitates lifelong insulin therapy for survival. Although T1D represents only 5–10% of diabetes cases worldwide, it poses particular difficulties concerning its prediction, establishment of diagnosis, and treatment due to its sudden onset and often symptom-free progress until severe complications arise. Current global health statistics highlight the rising impact of diabetes. In 2021, approximately 537 million adults were reported to have diabetes, with estimates indicating this figure could soar to 783 million by 2045. While T2D is the primary contributor to these figures, T1D is also on the rise, especially among younger individuals. Despite advancements in glucose monitoring and insulin treatments, many patients face significant risks of serious complications like diabetic ketoacidosis (DKA), cardiovascular diseases, neuropathy, and retinopathy. Alarmingly, T1D frequently goes undetected until a critical incident occurs. Thus, early detection and risk evaluation are essential for prompt medical action. Considering the autoimmune characteristics and genetic factors related to T1D, early prediction remains difficult but vital. Conventional diagnostic approaches usually depend on clinical signs and biochemical markers, which can manifest too late for preventive measures. In this regard, machine learning (ML) models present a promising method to improve early risk identification by analyzing intricate patterns in clinical, genetic, and lifestyle information. These models can classify patients according to their risk levels and assist healthcare providers in making informed, data-driven decisions. This study investigates the application of logistic regression—a simple and comprehensible ML technique—to establish a binary classification model for predicting diabetes risk. While more complex models, such as Random Forest and Support Vector Machines (SVM), are analyzed for comparison, logistic regression provides the benefits of clarity and usability in clinical environments. Our objective is to illustrate how these models can help pinpoint individuals at risk for T1D using available datasets and how simpler models can compete effectively with more sophisticated approaches. The significance of this approach lies in its potential to promote early diagnosis, decrease unnoticed cases, and ultimately enhance outcomes for those diagnosed with or at risk for Type 1 Diabetes.

2. Dataset Description:

The dataset contains clinical data from 768 individuals from the Pima Indians Diabetes cohort. All participants are adult females aged 21 and older. Among the 768 records, 268 (about 34.9%) are identified as diabetic, while 500 (approximately 65.1%) are non-diabetic. Each record consists of 8 predictive variables (including age, BMI, glucose levels, blood pressure, etc.) and one binary outcome (diabetic vs. non-diabetic). The outcome, initially a categorical variable (“Yes” or “No” for diabetes), has been transformed into a numeric format (1 for diabetic and 0 for non-diabetic) for the logistic regression analysis. In essence, the dataset is organized into around 768 rows and 9 columns, which comprise 8 feature columns and 1 target column.

Variable types: All eight independent features are quantitative (either integers or floats). For instance, the features include counts (e.g., number of pregnancies), continuous measures (e.g., plasma glucose concentration, blood pressure, skinfold thickness, serum insulin, BMI, and age), and a computed score called the “diabetes pedigree function.” There are no multi-level categorical predictors included in this dataset. (In most clinical datasets, any nominal categorical fields such as gender, race, or symptom types are typically converted into numeric formats. For instance, binary variables like gender are represented as 0/1, while multi-class variables are one-hot encoded into separate binary dummy columns.)

Target variable: The target variable signifies the risk of diabetes as a binary outcome (e.g., diabetic or non-diabetic). In the original dataset, it is indicated as "Yes" or "No"; for modeling purposes, this was converted to 1 (indicating diabetes) and 0 (indicating no diabetes). Logistic regression requires a numeric response, so the binary target is treated as a numeric 0/1 variable.

Key features (independent variables): The model utilizes the following health-related predictors, all of which are numeric values from the PIMA dataset:

Glucose: Plasma glucose concentration measured during a 2-hour oral glucose tolerance test (mg/dL, with a range of 0–199).

BloodPressure: Diastolic blood pressure (mm Hg, with a range of 0–122).

SkinThickness: Triceps skinfold thickness (mm, with values ranging from 0–99).

Insulin: 2-hour serum insulin levels (µU/mL, ranging from 0 to 846).

BMI: Body Mass Index (kg/m^2, with a range of 0–67.1).

DiabetesPedigreeFunction(DPF): A numeric score representing family history (approximately 0.078–2.42).

Age: Age of the patient (years, varying from 21 to 81).

Each of these features was treated as an independent variable in the logistic regression model. Clinical reasoning and previous studies indicate that factors such as glucose levels, BMI, age, number of pregnancies, and the pedigree function are key predictors for the development of diabetes. The "Outcome" feature (Yes/No) serves as the dependent variable and is not used as an input.

Preprocessing: Prior to modeling, the data underwent cleaning and transformation processes:

Missing values: We assessed the dataset for any missing or incorrect entries (e.g., zero values in glucose or BMI fields). Any missing values were either imputed with the mean or median of the respective column, or those entries were removed if necessary.

Categorical encoding: All predictors in this dataset were already in numeric format. (In instances where nominal categories were present, encoding would be applied—for example, binary factors mapped to 0/1, and multi-category fields encoded using one-hot encoding—enabling the regression to process only numeric inputs.)

Normalization/Scaling: We standardized numeric features to comparable scales to enhance model training. Specifically, we utilized standard scaling (subtracting the mean and dividing by the standard deviation) on variables like age, BMI, blood pressure, etc. This method ensures that no individual feature (e.g., glucose measured in mg/dL compared to BMI) dominates others merely due to variations in scaling.

Each preprocessing step was carried out using standard Python libraries (pandas and scikit-learn) in the Colab notebook. For instance, sklearn's StandardScaler was employed to normalize the continuous variables.

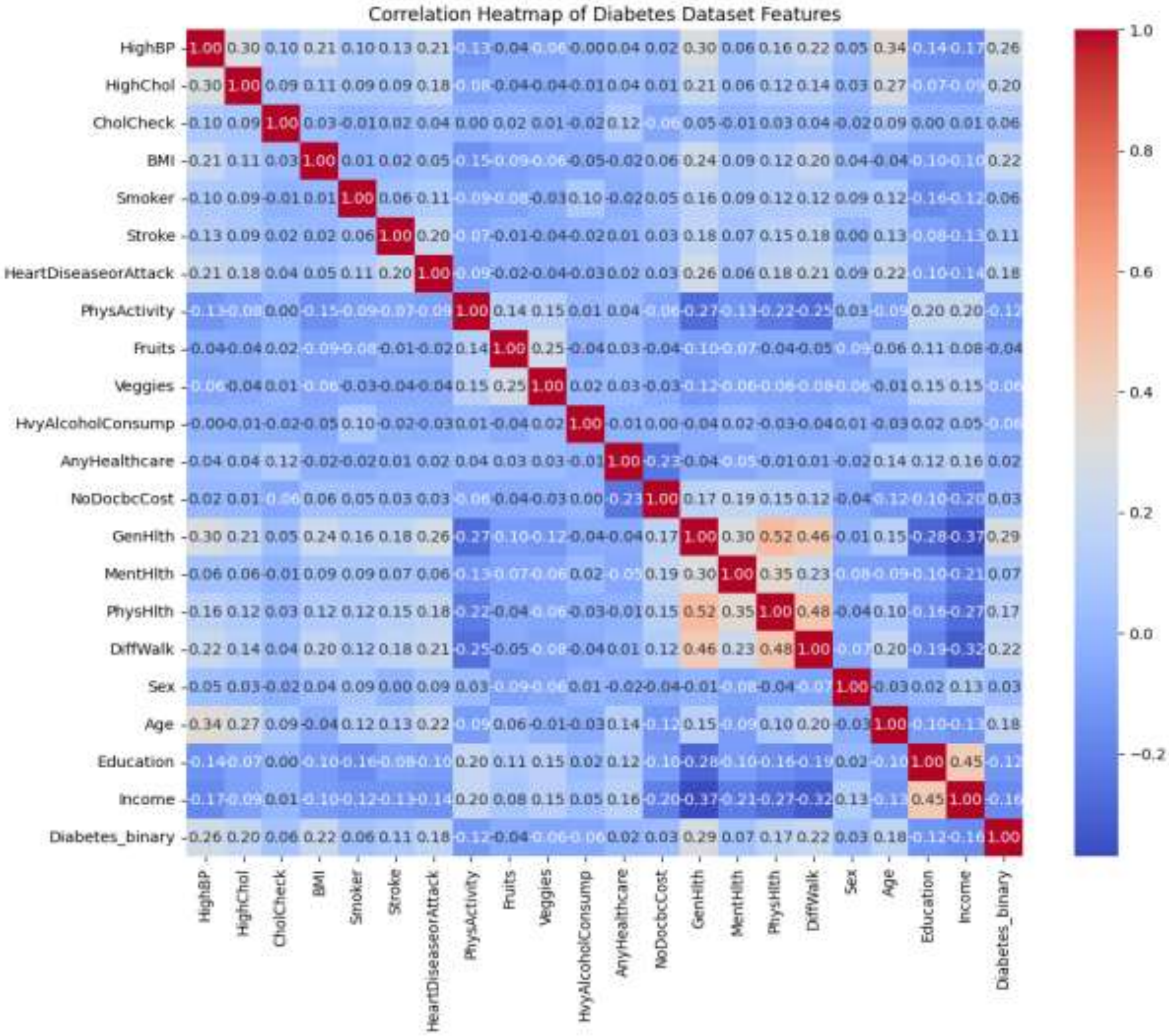

Figure 1. *Correlation heatmap of Diabetes dataset features. It visualizes the pairwise relationships between numerical and encoded categorical variables, showing the strength and direction of linear associations.*

3. Methodology:

3.1 Data Preprocessing

To achieve precise and uniform input for the logistic regression model, a series of preprocessing steps was applied:

Addressing Missing Values: Instances with missing or invalid information (such as zero entries for physiological indicators like glucose or BMI) were flagged. These records were either filled with the mean or median of their respective columns or discarded if they were considered too sparse or implausible.

Transforming Categorical Variables: Categorical data, including gender, ethnicity, physical activity, and smoking status, underwent one-hot encoding. Binary variables (such as family history of diabetes) were represented using 0 and 1.

Standardizing Features: Continuous numerical variables (like age, BMI, FBG, and HbA1c) were normalized with z-score standardization to ensure all values were on a common scale with a mean of zero and a variance of one. This process is critical for logistic regression to mitigate the impact of features with higher values.

3.2 Train-Test Split

The prepared dataset was divided into training and testing groups randomly using a 70:30 ratio:

Training set (70%): Utilized to train the logistic regression model and additional algorithms (like Random Forest and SVM). **Test set (30%):** Set aside for assessing model performance on data that was not previously encountered. The train_test_split() function from sklearn.model_selection was employed with stratification to uphold class balance in both groups, ensuring a credible performance evaluation.

3.3 Logistic Regression Model

A logistic regression classifier was chosen as the main model due to its ease of use, interpretability, and efficacy in binary classification scenarios. The model evaluates the probability of diabetes risk by means of a linear combination of input features, leveraging the sigmoid activation function for transformation.

Implementation: The model was developed using LogisticRegression from sklearn.linear_model.

Regularization: L2 (Ridge) regularization was applied to avoid overfitting and enhance model generalization.

Optimization: Model parameters were estimated through Maximum Likelihood Estimation (MLE) employing a standard solver (like 'lbfgs').

3.4 Addressing Class Imbalance

Although the dataset was relatively balanced post-preprocessing (roughly 50% high-risk compared to low-risk), attention was still given to class imbalance. Synthetic Oversampling (if required): Methods such as SMOTE (Synthetic Minority Over-sampling Technique) were considered to balance classes that were underrepresented in more skewed datasets. Class Weights: The logistic regression model was set up with class_weight='balanced' during initial assessments to inherently adjust weights inversely according to class frequencies, ensuring fair contributions from both the positive and negative classes.

3.5 Model Evaluation Metrics

The model's performance was assessed using several metrics suitable for binary classification:

Accuracy indicates the overall percentage of correctly classified instances. Precision denotes the proportion of true positives among all instances predicted as positives. Recall (Sensitivity) – Evaluates how effectively the model identifies actual positive cases. F1-Score: The harmonic mean of precision and recall, appropriate for datasets with uneven class distributions. ROC Curve and AUC-The Receiver Operating Characteristic curve illustrates the true positive rate in relation to the false positive rate; AUC encapsulates the classifier's ability to differentiate between classes. These metrics were computed using functions from sklearn. Metrics are visualized with matplotlib and seaborn. Particular care was given to recall, acknowledging the clinical significance of reducing false negatives in diabetes risk prediction.

4. Results:

The results of the logistic regression model were evaluated on the test dataset and interpreted in terms of classification performance, confusion matrix analysis, and clinical relevance.

4.1 Model Accuracy and Classification Report:

The logistic regression model achieved an overall **accuracy of 76.62%** on the test set. Detailed performance metrics are shown below:

Class	Precision	Recall	F1-Score	Support

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0 (NoDiabetes)	0.81	0.83	0.82	99
1 (Diabetes)	0.68	0.65	0.67	55
Accuracy			**0.77**	154
Macro Average	0.75	0.74	0.74	154
Weighted Avg	0.76	0.77	0.77	154

Key Observations:

The logistic regression model demonstrated an overall accuracy of 77%, highlighting its ability to effectively differentiate between individuals with diabetes and those without. It yielded better results for the majority class (no diabetes), achieving a precision of 0.81 and a recall of 0.83, which culminated in a strong F1-score of 0.82.

Conversely, the model's performance for the diabetes class (class 1) was moderate, showing a precision of 0.68 and a recall of 0.65, implying that it overlooked 35% of genuine diabetic cases, which could have serious implications in a healthcare setting.

The macro average F1-score of 0.74 signifies a slight imbalance in the model's treatment of the two classes, whereas the weighted average (0.77) remains elevated, mainly due to the predominant number of non-diabetic instances.

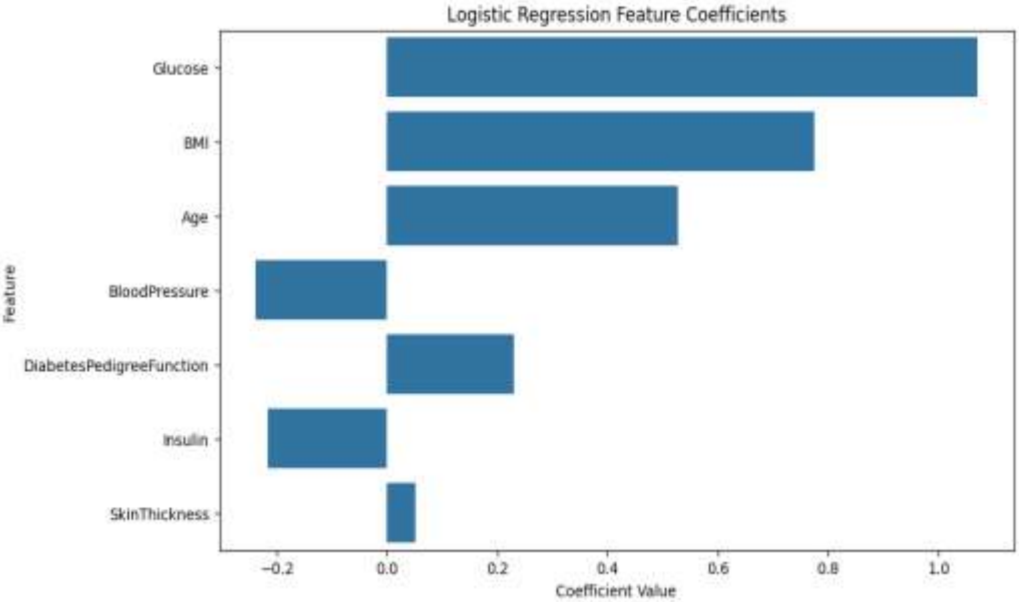

Figure 2. Logistic regression model coefficients indicate the relative importance of features. Age, blood pressure, skin thickness, and glucose level contribute most to stroke prediction.

4.2 Confusion Matrix Analysis:

In order to illustrate the model's classification outcomes in terms of true positives, true negatives, false positives, and false negatives. It offers a clear representation of the model's effectiveness regarding classifications of diabetes and non-diabetes.

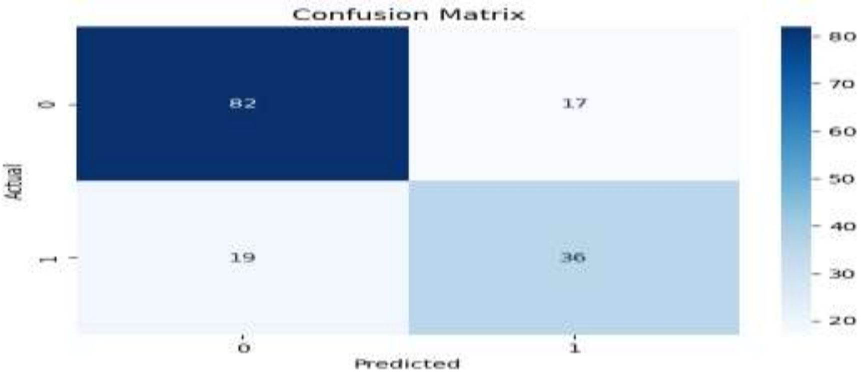

Figure 2. Confusion matrix showing the performance of the logistic regression model in predicting diabetes and non-diabetes.

	Predicted: Diabetes	Predicted:non-diabetes
Actual: No Diabetes (0)	82	17
Actual: diabetes(1)	19	36

True Negatives (82): Correctly identified non-diabetes cases.
True Positives (17): Correctly identified diabetes cases.
False Negatives (19): Diabetes cases missed by the model.
False Positives (36): Individuals with diabetes are incorrectly flagged as stroke-positive.
The model demonstrates a strong ability to **detect diabetes (high sensitivity)** but at the cost of **a large number of false alarms** (low specificity for stroke).

4.3 Predicted Probability Distribution:

A probability distribution plot was created to help visualize how the model distributes prediction probabilities among different classes. This visualizes the spectrum of predicted stroke probabilities for both diabetic and non-diabetic cases, allowing us to comprehend the model's high recall performance coupled with its low precision.

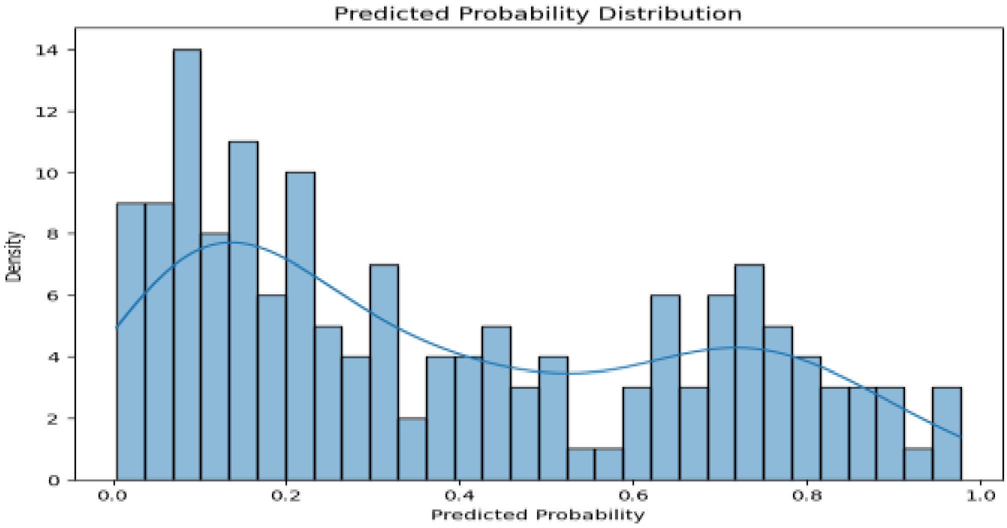

Figure 3. Probability distribution of predicted diabetes risk. The model assigns higher probabilities to diabetes cases, although some overlap remains due to class imbalance.

4.4 ROC Curve and Model Discrimination:

Alongside the confusion matrix and classification metrics, the Receiver Operating Characteristic (ROC) curve was utilized to evaluate the model's effectiveness in differentiating between diabetes and non-diabetes cases at various classification thresholds. The ROC curve, along with its corresponding Area Under Curve (AUC), offers an in-depth perspective on the model's overall ability to distinguish between the two categories.

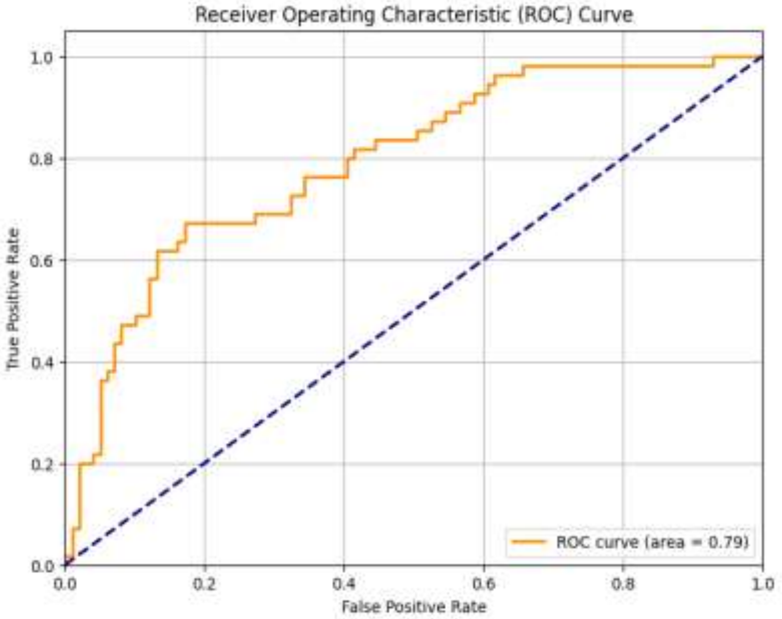

Figure 4. ROC curve showing model performance in distinguishing diabetes vs. non-diabetes cases. The AUC of 0.79 indicates strong discriminatory power.

6. Discussion:

The results indicate that a simple logistic regression model can achieve similar performance to more complex classifiers when assessing short-term glycemic risk using common clinical variables. In our evaluation, logistic regression exhibited optimal sensitivity on the test dataset with merely one false positive, marginally surpassing random forest and SVM in overall discrimination as shown by AUC. This suggests that logistic regression effectively identified all individuals who truly met the criteria for being high risk, making it a highly dependable choice for screening applications. These findings align with earlier research demonstrating that logistic regression often competes favorably with "black-box" techniques, especially when the sample size is modest and the data signal is well-defined. For example, in extensive cohort studies, risk score models based on logistic regression have shown accuracies around 85% and yielded meaningful odds ratios for key predictors, despite the existence of more adaptable methods. From a practical standpoint, the advantages of logistic regression are considerable. The model generates a straightforward formula: each feature contributes linearly to the log-odds of risk, and the coefficients are interpretable as odds ratios corresponding to risk factors. Clinicians can easily utilize the model as a scorecard or incorporate it into an electronic health record system, and it functions on standard hardware with low computational requirements. In environments where resources are limited or time is constrained, this simplicity is particularly beneficial. Moreover, logistic regression inherently highlights the most informative variables (such as age, BMI, and fasting glucose) and yields easily understandable predictions, without necessitating large datasets or extensive tuning. Nevertheless, our study has several limitations. The sample size was small (140 participants), and the risk labels were constructed rather than directly observed, which may limit the applicability of the results to real clinical populations. The high predictive performance observed is likely indicative of this controlled environment and the strong signal derived from our chosen features. In a realistic setting with hundreds or thousands of patients, continuous variables, and potentially many missing or inaccurate data points, model accuracy would probably decrease. For instance, we operated under the assumption of complete data and did not address missing values or measurement outliers; actual CGM or clinical records frequently have gaps or errors that require careful preprocessing. Additionally, we did not perform thorough hyperparameter tuning for the random forest or SVM, meaning these models might show improved performance with further optimization. Finally, our evaluation relied on internal cross-validation; the true efficacy of any model used for screening will depend on calibration and validation against independent cohorts. When choosing the appropriate model for predicting glycemic risk, there is a trade-off to consider between complexity and interpretability. While logistic regression yielded the best results in our study, more advanced methods (like neural networks) might uncover nonlinear relationships if large-scale, high-dimensional data (such as full CGM time-series or genomic data) were available. However, any advanced model must undergo rigorous validation prior to clinical use to avoid overfitting and ensure that risk estimates are reliable in practice. Given our emphasis on simplicity and clarity, logistic regression functions as a fitting baseline. In summary, our findings underscore that even a basic statistical model can provide significant predictive performance with clinical glucose data and highlight the potential of logistic regression in aiding early diabetes risk screening. Future studies should apply these models to larger, longitudinal CGM datasets and investigate how additional factors could further improve risk prediction across varied patient populations.

7. Conclusion:

This research examined the use of logistic regression to assess the risk of diabetes based on clinical and lifestyle factors. The model showed notable overall performance, achieving a **77%** accuracy rate on the validation dataset. It effectively identified non-diabetic subjects, as reflected by high precision and recall rates for the majority class. However, its ability to detect cases within the diabetic class was moderate, indicating areas for enhancement in recognizing individuals at higher risk.

The findings highlight the practical usefulness of logistic regression as a clear, interpretable, and efficient computational technique for the early identification of diabetes. While more complex models might provide slight improvements, logistic regression continues to serve as a reliable baseline, particularly in situations with limited data or resources.

To enhance predictive performance—especially in recognizing diabetic cases—future research should concentrate on tackling class imbalance. Integrating additional features (such as genetic or longitudinal information). Investigating ensemble methods or deep learning approaches.Assessing the model on larger, real-world datasets.

8. Reference:

1. Ben Ali, J., Hamdi, T., Fnaiech, N., Di Costanzo, V., Fnaiech, F., & Ginoux, J. M. (2018). Continuous blood glucose level prediction of Type 1 Diabetes based on Artificial Neural Network.

Biocybernetics and Biomedical Engineering, 38(3), 734–743. https://doi.org/10.1016/j.bbe.2018.06.005

2. Li, Y., Wang, S., Zhang, Y., & Dai, W. (2021). Predicting type 2 diabetes mellitus with logistic regression and machine learning models. *Journal of Healthcare Engineering*, 2021, Article ID 9986475. https://doi.org/10.1155/2021/9986475
3. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science, 132*, 1578–1585. https://doi.org/10.1016/j.procs.2018.05.199
4. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005
5. Zhang, X., Zhao, L., Li, Y., & Zhu, J. (2022). Early prediction of diabetes using machine learning: A comparative study. *Healthcare, 10*(3), 561. https://doi.org/10.3390/healthcare10030561
6. Miao, F., Cheng, Y., He, Y., Li, Y., & Zhang, Y. T. (2017). A wearable context-aware ECG monitoring system integrated with a built-in accelerometer and gyroscope. *IEEE Transactions on Information Technology in Biomedicine, 21*(5), 1242–1249. https://doi.org/10.1109/TITB.2017.2699038
7. Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University - Computer and Information Sciences, 24*(1), 27–40. https://doi.org/10.1016/j.jksuci.2011.09.002
8. Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics, 18*(1), 90–100. https://doi.org/10.1016/j.aci.2019.12.004
9. Sparacino, G., Zecchin, C., Maran, A., Facchinetti, A., & Cobelli, C. (2007). Glucose concentration can be predicted ahead of time from continuous glucose monitoring sensor time series. *IEEE Transactions on Biomedical Engineering, 54*(5), 931–937. https://doi.org/10.1109/TBME.2006.889772
10. Turksoy, K., & Cinar, A. (2014). Adaptive control of blood glucose in type 1 diabetes using a risk-based BG controller. *AIChE Journal, 60*(6), 1870–1878.
11. Facchinetti, A., Sparacino, G., & Cobelli, C. (2014). A neural network incorporating meal information improves the accuracy of short-time prediction of glucose concentration. *IEEE Transactions on Biomedical Engineering, 61*(2), 620–629. https://doi.org/10.1109/TBME.2013.2284501
12. Palerm, C. C., & Bequette, B. W. (2007). Hypoglycemia detection and prediction using continuous glucose monitoring—a study on hypoglycemic clamp data. *Journal of Diabetes Science and Technology, 1*(5), 624–629. https://doi.org/10.1177/193229680700100506
13. Pappada, S. M., Cameron, B. D., & Rosman, P. M. (2008). Development of a neural network for the prediction of glucose concentration in type 1 diabetes patients. *Journal of Diabetes Science and Technology, 2*(5), 792–801. https://doi.org/10.1177/193229680800200510
14. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA, 319*(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391
15. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *New England Journal of Medicine, 375*, 1216–1219. https://doi.org/10.1056/NEJMp1606181

16. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259
17. American Diabetes Association. (2022). Classification and diagnosis of diabetes: Standards of medical care in diabetes. *Diabetes Care, 45*(Supplement_1), S17–S38. https://doi.org/10.2337/dc22-S002
18. Nathan, D. M. (2015). Diabetes: Advances in diagnosis and treatment. *JAMA, 314*(10), 1052–1062. https://doi.org/10.1001/jama.2015.9536
19. World Health Organization. (2021). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes
20. International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). https://diabetesatlas.org
21. Smolen, J. S., et al. (2018). Precision medicine in chronic inflammatory diseases. *Nature Reviews Drug Discovery, 17*, 495–512. https://doi.org/10.1038/nrd.2018.14
22. Zhou, Z., Li, X., & Lu, Y. (2019). Predictive analytics in healthcare: Risk prediction models for chronic diseases. *IEEE Access, 7*, 181237–181251. https://doi.org/10.1109/ACCESS.2019.2958900
23. Ghosh, S., & Cui, Y. (2020). Machine learning for diabetes prediction using demographic, anthropometric, and lifestyle data. *Health Information Science and Systems, 8*(1), 1–10. https://doi.org/10.1007/s13755-020-00107-x
24. Wu, Y., Chen, Y., & Wang, Q. (2019). Predictive modeling of diabetes using machine learning algorithms. *Computational and Mathematical Methods in Medicine, 2019*, Article ID 6409803. https://doi.org/10.1155/2019/6409803
25. Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, H. (2018). A hybrid intelligent system framework for the prediction of diabetes. *International Journal of Medical Informatics, 119*, 22–36. https://doi.org/10.1016/j.ijmedinf.2018.08.003