Journal of Computer and Communications, 2026, 14(9), 71-92
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

# A Sequential Logistic Regression-Gradient Boosting Framework for Diabetes Classification

**Moyang Li[1*], Bowen Cai[2]**

[1]Shady Side Academy, Pittsburgh, PA, USA
[2]School of Environment, Tsinghua University, Beijing, China
Email: *27lim@shadysideacademy.org, cbw@tsinghua.edu.cn

**How to cite this paper:** Li, M.Y. and Cai, B.W. (2026) A Sequential Logistic Regression-Gradient Boosting Framework for Diabetes Classification. *Journal of Computer and Communications*, **14**, 71-92.
https://doi.org/10.4236/jcc.2026.149006

**Received:** July 21, 2026
**Accepted:** September 20, 2026
**Published:** September 23, 2026



## Abstract

Diabetes mellitus, particularly type 2 diabetes, is a major global public health problem, making early detection and intervention essential for reducing disease incidence. This study used the Pima Indians Diabetes Dataset as the source of the training and internal validation sets and the Frankfurt Diabetes Dataset as the external test set. Since the outcome is imbalanced, with non-diabetic cases (Outcome = 0) as the majority class, a two-stage integrated model (a sequential two-stage classification framework) was proposed to reduce the false negative rate. Logistic regression first identifies high-confidence positive cases and immediately classifies them as diabetic; remaining patients are passed to GBM as a second stage. The final classification combines positive predictions from both stages, providing an additional opportunity to identify Outcome = 1 cases while reducing false-negative predictions. The two-stage model achieved an accuracy of 77.27% with a false negative rate of 38.89% on the internal validation set and an accuracy of 89.35% with a false negative rate of 20.47% on the external test set, demonstrating strong robustness and transferability. Additionally, SHAP was applied to enhance model interpretability, identifying Glucose and Insulin as the most important features, followed by BMI, DiabetesPedigreeFunction, and Age.

## Keywords

Diabetes, Classification, Sequential Framework, Two-Stage Model, SHAP

## 1. Introduction

Type 2 diabetes mellitus (T2DM) is a growing global public health challenge that impacts over 800 million people worldwide, causing serious complications including blindness, kidney failure, heart attacks, and strokes, and causing the death of

millions every year [1]. In fact, the global number of adults with diabetes has increased by 630 million from 1990 to 2022, showing an unprecedented surge in the risk of various complications caused by diabetes [2]. Before developing diabetes, most people go through a pre-diabetic stage where they have high blood glucose either resulting from insulin resistance or insulin deficiency [3]. Because timely identification can support clinical management, accurate detection and classification of diabetes are important for patients and healthcare systems.

Early detection requires a comprehensive evaluation of various biochemical and physical risk variables. Obesity, defined as having a high Body Mass Index (BMI), is a significant risk factor [4]. High blood pressure and blood glucose levels are also key features for determining a person's prediabetic state [4]. Furthermore, non-modifiable characteristics such as age, family history, and reproductive history are also important traits that need to be monitored [5]. Besides, since decreased insulin production is essential to the development of chronic diabetes conditions, insulin level is also a crucial risk variable [6].

Machine learning (ML) techniques have emerged as a useful tool in making classifications for diabetes: it succeeds by applying to datasets such as Pima Indians Diabetes Dataset collected by the National Institute of Diabetes and Digestive and Kidney Diseases [7]. For example, Zou *et al.* [8] demonstrated the effectiveness of ML methods like Random Forest and Neural Networks: it successfully found intricate patterns such as non-linear connections in biochemical indicators, which are used to be challenging with conventional statistical techniques. Moreover, Kiran *et al.* [9] show a trend from traditional methods toward advanced ensemble models, such as Random Forest or Gradient Boosting. It suggests the effectiveness of ensemble-based machine learning methods in the prediction and interpretation of T2DM. However, class imbalance remains a major challenge in diabetes prediction: dominance of non-diabetic cases causes models to favor the majority class and reduce their ability to identify diabetic patients [10].

Building on these challenges, this study proposes a sequential two-stage classification framework in **Figure 1**, aiming to improve diabetes classifications while reducing missed diagnoses: using the Pima Indians Diabetes Dataset to prioritize lowering false negatives, which is crucial for medical safety because missed diagnosis can result in serious problems. The first step is data cleaning and preprocessing, including RF imputing missing data and SMOTE addressing data imbalance. Then, four main algorithms (RF, SVM, ANN, and GBM) were evaluated, with GBM standing out with the highest accuracy of 76.62%. In addition, to reduce false negative rate, a two-stage model was created: this approach uses Logistic Regression (LR) as a high-confidence positive identification stage. Gradient Boosting Machines (GBM) are then used to identify overlooked instances in the remaining subset.

This two-stage model was trained on the 80% training set drawn from the Pima Indians Diabetes Dataset. The remaining 20% was used as the internal validation set to evaluate the model's performance; additionally, the Frankfurt Diabetes Da-

taset was used as the external test set to assess model's generalizability. On the internal validation set, the model achieved an overall accuracy of 77.27% with a false negative rate of 38.89%; on the external test set, the model achieved a higher accuracy of 89.35% and a lower false negative rate of 20.47%. The results demonstrate model's strong robustness and transferability.

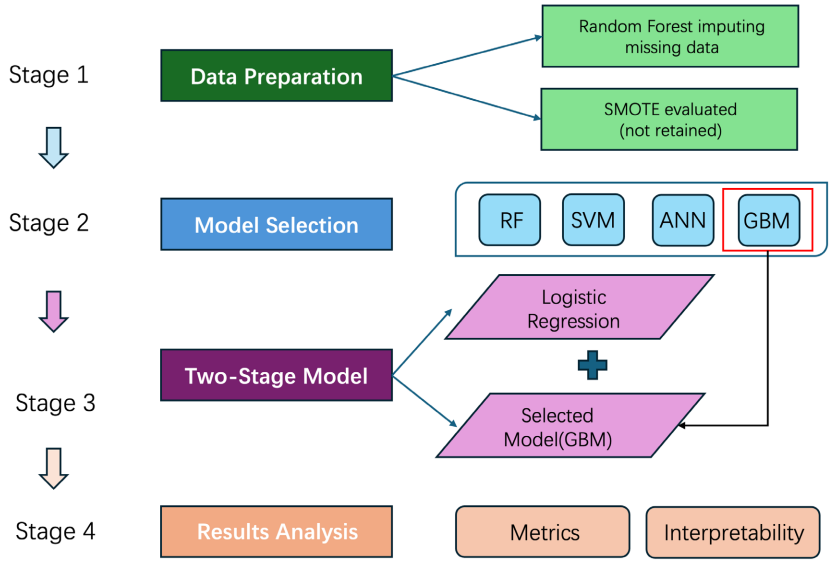



**Figure 1.** Overview of the sequential two-stage Logistic Regression-Gradient Boosting classification framework.

Result analysis further revealed critical thresholds, including sharp risk increases at Glucose ≈ 127 - 130 mg/dL and BMI ≥ 30 kg/m$^2$. SHAP feature importance plot showed Glucose as the strongest predictor, followed by BMI, Age, and Insulin; Interaction plot illustrated strong relationships between Glucose and Age and Glucose and BMI.

Therefore, the sequential two-stage framework lowers false negative rates while maintaining overall prediction accuracy, effectively preventing missed diagnoses; it simultaneously holds high interpretability, providing useful information for clinical practice. Model's approach is scalable to different clinical settings, facilitating early detection of diabetes.

## 2. Literature Review

### 2.1. Background: T2DM Etiology and Research Value

Data reveal that approximately 463 million adults worldwide were living with diabetes mellitus in 2019, and it will surge by 51% to reach 700 million by 2045 [11]. Despite its growing prevalence, many cases of T2DM could be prevented with lifestyle changes, including weight control, maintaining healthy diet, staying physically active, and so on [12]. Therefore, accurate classifications and clinical management of diabetes are crucial to prevent disease progression and severe complications in the future.

In addition, the pathogenesis of T2DM is highly complex, it involves interaction of multiple biological and environmental factors, which influence one another through complicated pathways rather than simple linear relationships [13]. As a result, diabetes prediction requires advanced methods that are capable of capturing complex nonlinear relationships. Choi *et al.* [14] demonstrate that machine learning models consistently outperform traditional statistical methods in identifying undiagnosed diabetes, particularly when data is complex and non-linear. This highlights the necessity of training machine learning methods in modern diabetes prediction research to boost prediction performance.

### 2.2. Existing ML Approaches for T2DM Classification

Ijaz *et al.* [15] addressed imbalance and noisy dataset by using DBSCAN for outlier detection and SMOTE for class balance, improving the performance of RF classifiers.

Hasan *et al.* [16] provide a framework for ensemble machine learning methods: they used an ensemble of classifiers, including AdaBoost and XGBoost, weighing different models by AUC. Similarly, Ganie *et al.* [17] used boosting ensembles such as CatBoost and XGBoost with cross-validation and splitting data into multiple folds. These two studies show the effectiveness of ensemble models in early diabetes classifications.

Bala Manoj Kumar *et al.* [18] used Deep Neural Networks to achieve 98.16% classification accuracy on the Pima Indians Diabetes Dataset. However, their extremely high accuracy often causes concerns regarding overfitting and low generalizability, proving the important balance between prediction accuracy and clinical applicability.

In contrast, Edlitz and Segal [19] used a complex GBDT (Gradient Boosting Decision Tree) model but still holds high interpretability. Vakil *et al.* [20] combine multiple machine learning models with explainable AI methods and SHAP (Shapley Additive Explanations) for enhancing the interpretability of early detection of diabetes. These studies highlight the importance of balancing prediction performance and clinical interpretability.

### 2.3. Motivation for a Sequential Two-Stage Classification Framework

Although machine learning methods have made progress in diabetes classifications, a critical challenge still remains in clinical datasets: class imbalance. In most screening populations, non-diabetic individuals constitute the majority of cases. For example, the global diabetes prevalence in 2019 is estimated to be 9.3% (463 million people) [11]. Talebi Moghaddam *et al.* [21] demonstrated that this class imbalance can significantly affect model performance and often leads to reduced sensitivity. In particular, classifiers tend to favor the majority class and incorrectly classify diabetic individuals as non-diabetic, resulting in a high false negative rate. In clinical screening, these missed diagnoses are especially problematic because

patients may lose the opportunity for timely intervention and treatment, increasing the risk of disease progression and long-term complications [22].

Furthermore, different prediction models exhibit distinct strengths and limitations. Seto *et al.* [23] showed that Gradient Boosting Decision Trees is better in capturing complex non-linear relationships and achieving higher predictive accuracy; Logistic Regression remains valuable because of its simplicity, interpretability, and well-calibrated probability estimates. Similarly, Abousaber *et al.* [10] evaluated multiple machine learning algorithms for diabetes classification, including Random Forest (RF), Gradient Boosting Machines (GBM), Support Vector Machines (SVM), Artificial Neural Networks (ANN). Their results showed that no single algorithm consistently outperformed all others across different datasets and preprocessing strategies, suggesting that different models possess complementary strengths.

Beyond model selection, the choice of interpretability method is equally important. Lundberg *et al.* [24] showed that the SHAP method can improve the interpretability of tree-based models, including measuring local feature interaction effects and combining many local explanations of each prediction.

## 3. Data Preparation

### 3.1. Data Preliminary Analysis

The Pima Indians Diabetes Dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases [25]. Original purpose of the dataset is to diagnostically predict the presence of diabetes based on certain measurements. Notably, the dataset has some constraints: all instances in the dataset are females at least 21 years old of Pima Indian heritage.

The dataset captures 768 samples of patients to analyze the biochemical and physical factors [7]. It consists of eight explanatory variables: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The target variable, Outcome, is binary-coded, where 1 represents the presence of diabetes and 0 indicates its absence.

On the second row of **Table 1**, one important finding is that there are zero values in a number of categories including blood pressure, skin thickness, insulin, BMI, and glucose. These zeros are medically impossible: Glucose and Insulin are primary biochemical indicators of glycemic control and pancreatic function, while BMI, Blood Pressure, and Skin Thickness serve as physical markers for obesity and vascular health associated with insulin resistance. Collectively, these five variables define the metabolic profile necessary for accurately assessing diabetic risk and identifying latent physiological abnormalities. Features like Skin Thickness (227 zeros) and Insulin (374 zeros) have a significant amount of missingness, as the table illustrates, indicating that these data points were either randomly absent or not recorded. These zeros are regarded as missing observations and will be handled using sophisticated imputation techniques to preserve the integrity of the ensuing predictive modeling.

Table 1. Descriptive statistics for explanatory variables in the Pima Indians Diabetes Dataset.

| Variable | Missing value | Mean | Std. | Min | Max |
|---|---|---|---|---|---|
| Pregnancies | 0 | 3.85 | 3.37 | 0 | 17 |
| Glucose | 5 | 121.69 | 30.54 | 44 | 199 |
| BloodPressure | 35 | 72.41 | 12.38 | 24 | 122 |
| SkinThickness | 227 | 29.15 | 10.48 | 7 | 99 |
| Insulin | 374 | 155.55 | 118.78 | 14 | 846 |
| BMI | 11 | 32.46 | 6.92 | 18.2 | 67.1 |
| DiabetesPedigreeFunction | 0 | 0.47 | 0.33 | 0.08 | 2.42 |
| Age | 0 | 33.24 | 11.76 | 21 | 81 |
| Outcome | 0 | 0.35 | 0.48 | 0 | 1 |

### 3.2. External Test Dataset

To further evaluate the sequential two-stage framework, the Frankfurt Diabetes Dataset, collected from a hospital in Frankfurt, Germany, was introduced as the external test set [26]. The 80% training set drawn from the Pima Indians Diabetes Dataset was used to train the model, and the 20% internal validation set was used to validate the model. The Frankfurt Diabetes Dataset comprises 2000 female patients and shares the same eight explanatory variables and the same binary Outcome label. Thus, the trained model can be directly applied to the external test set.

Similarly, the external test set in Table 2 contains physiologically implausible zero values in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI. These zero values were regarded as missing values. Therefore, missing data issue exists in both two datasets and, hence, a method is proposed to address this problem in 3.3.

Table 2. Descriptive statistics for explanatory variables in the Frankfurt Diabetes Dataset.

| Variable | Missing value | Mean | Std. | Min | Max |
|---|---|---|---|---|---|
| Pregnancies | 0 | 3.70 | 3.31 | 0.00 | 17.00 |
| Glucose | 13 | 121.98 | 30.63 | 44.00 | 199.00 |
| BloodPressure | 90 | 72.40 | 12.23 | 24.00 | 122.00 |
| SkinThickness | 573 | 29.34 | 10.80 | 7.00 | 110.00 |
| Insulin | 956 | 153.74 | 111.27 | 14.00 | 744.00 |
| BMI | 28 | 32.65 | 7.24 | 18.20 | 80.60 |
| DiabetesPedigreeFunction | 0 | 0.47 | 0.32 | 0.08 | 2.42 |
| Age | 0 | 33.09 | 11.79 | 21.00 | 81.00 |
| Outcome | 0 | 0.34 | 0.47 | 0.00 | 1.00 |

### 3.3. Handling of Missing Values

To address the missing values in features in 3.2, a Random Forest (RF) imputation method, specifically the MissForest algorithm, was employed. Unlike simple mean or median imputation, which can distort the data distribution and ignore inter-variable correlations, RF imputation is a non-parametric approach that predicts missing values by leveraging the complex, non-linear relationships between all observed features.

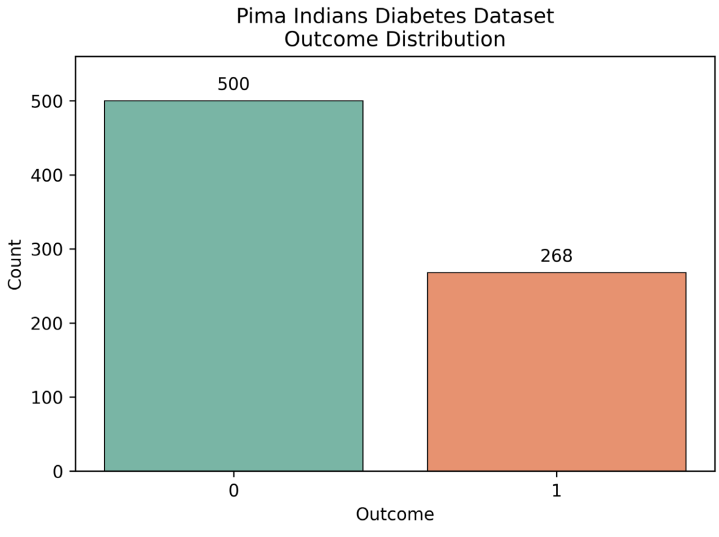
The imputation process treats each variable with missing data as a target (Y) and the remaining variables as predictors (X) [27]. For a missing value in variable j, an ensemble of decision trees learns a functional mapping. The imputation accuracy is evaluated using NRMSE (Normalized Root Mean Square Error):

$$\mathrm{NRMSE} = \sqrt{\frac{\mathrm{mean}\left(\left(X_{\mathrm{true}} - X_{\mathrm{imp}}\right)^2\right)}{\mathrm{var}\left(X_{\mathrm{true}}\right)}} \tag{1}$$

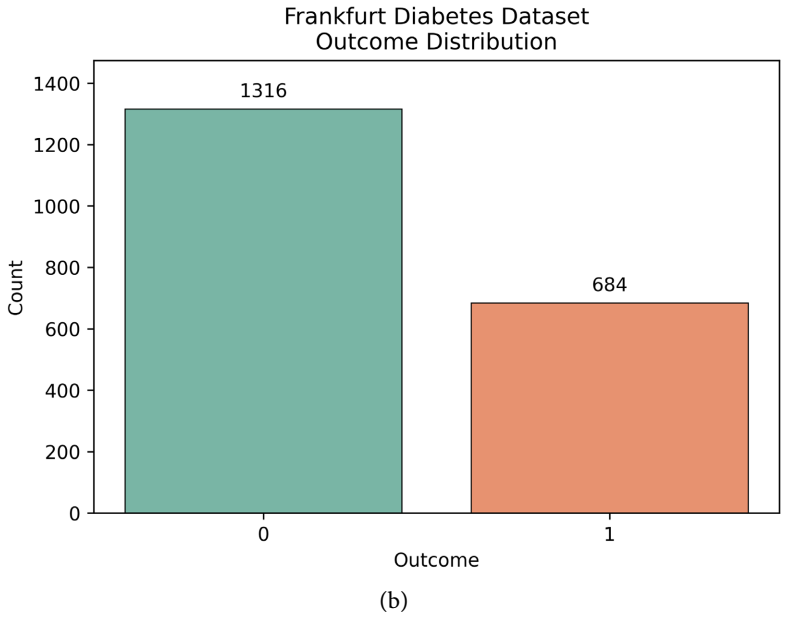
By leveraging the covariance structure of complete variables like Age and BMI, the model effectively reconstructs high-missingness features such as Skin Thickness and Insulin. This method provides a solid basis for the prediction model by reconstructing high-missing features like skin thickness and insulin from the covariance structure of complete variables like age and BMI using an Iterative Imputer framework.

### 3.4. Addressing Data Imbalance

In **Figure 2**, the Outcome distribution in two datasets both indicates data imbalance: in the Pima Indians Diabetes Dataset, 500 instances (65.1%) labeled as 0 and 268 instances (34.9%) labeled as 1; in the Frankfurt Diabetes Dataset, 1316 instances (65.8%) labeled as 0 and 684 instances (34.2%) labeled as 1. This data inequality, characterized by a larger proportion of non-diabetic cases, may bias model performance toward the majority class.




(a)




(b)

**Figure 2.** Distribution of the outcome (0 = no diabetes, 1 = diabetes) in the Pima Indians Diabetes Dataset and the Frankfurt Diabetes Dataset. (a) Pima Indians Diabetes Dataset; (b) Frankfurt Diabetes Dataset.

Hence, to address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was employed to generate synthetic samples of the minority class (Outcome = 1) [28]. However, experimental findings showed that SMOTE did not improve the performance of the best model in this investigation. The top-performing single-stage classifier, the Gradient Boosting Machine (GBM), maintained the same accuracy on the original and SMOTE-balanced datasets. The original imbalanced dataset was thus kept for all ensuing modeling and interpretability analyses in order to preserve the real-world data.

## 4. Methodology

### 4.1. Model Selection

The Pima Indians Diabetes Dataset was stratified into an 80% training set and a 20% internal validation set to ensure consistent class distribution across splits. Four widely used classification algorithms were implemented as baseline comparators: Random Forest (RF) [29], Support Vector Machine (SVM) [30], Artificial Neural Network (ANN) [31], and Gradient Boosting Machine (GBM) [32]. Model performance was assessed using confusion matrix-derived metrics, including accuracy, false positive rate (actual non-diabetic cases incorrectly classified as diabetic), and false negative rate (actual diabetic cases incorrectly classified as non-diabetic).

**Figure 3** shows Gradient Boosting Machine (GBM) had the highest accuracy of 76.62%. Therefore, GBM was chosen as the secondary model for the later two-stage model.

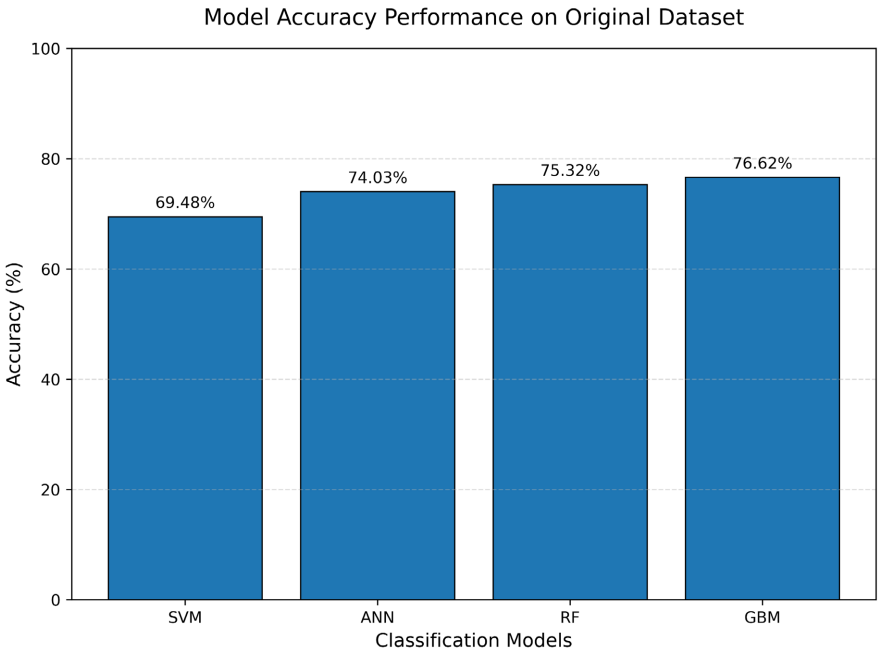



**Figure 3.** Comparison of four classification model accuracy on the internal validation set using the original imbalanced dataset.

## 4.2. A Sequential Logistic Regression-Gradient Boosting Framework for Diabetes Classification

In the standalone GBM evaluation, Gradient Boosting Machine (GBM) achieved an accuracy of 76.62%. However, it still has a high false negative rate (presence of diabetes is overlooked) of 40.74%, which is unsuitable for clinical screening. A missed diagnosis can cause severe consequences. Therefore, to ensure patients receive necessary early care and reduce the false negative rate, a two-stage model is proposed.

In the first stage, Logistic Regression serves as a high-confidence positive identification stage that identifies positive cases (Outcome = 1); the remaining cases are then passed to GBM, which provides an additional opportunity to identify positive cases that were not classified as positive in the first stage.

To evaluate the robustness and transferability of the two-stage model, two evaluation procedures were conducted: first, the Pima Indians Diabetes Dataset was split into an 80% training set and a 20% internal validation set, where the internal validation set is used for testing the model's robustness; second, the Frankfurt Diabetes Dataset was used as the external test set to assess transferability using the established model.

Stage 1: Logistic Regression

A logistic regression model is first trained on the 80% training set drawn from the Pima Indians Diabetes Dataset. Logistic Regression (LR) is a foundational statistical model used for binary classification. It utilizes the logistic function to map input variables into a probability space between 0 and 1. As shown in the following formula [33]:

$$p(x)=\frac{1}{1+e^{-(x-\mu)/s}} \tag{2}$$

It employs the sigmoid function to estimate the probability $p$ of diabetes (Outcome = 1). In Stage 1, observations with logistic-regression probabilities greater than 0.9 were classified as positive. All remaining observations were passed to the GBM classifier for the final prediction.

Stage 2: Secondary Classifier (Gradient Boosting Model)

Gradient Boosting Machine is an advanced ensemble technique that builds models sequentially, with each new model attempting to correct the residual errors of the previous ones. Following the iterative optimization framework proposed by Friedman, the algorithm minimizes a loss function $L\left(y,F\left(x\right)\right)$ by adding weak learners (typically decision trees) using a gradient descent-like procedure. The model $F(x)$ is updated at each step m according to the formula:

$$F_m\left(x\right)=F_{m-1}\left(x\right)+\gamma_m h_m\left(x\right) \tag{3}$$

where $h_m\left(x\right)$ is the weak learner trained to predict the negative gradient (residuals) of the loss function, and $\gamma_m$ is the step size or learning rate [32]. GBM was selected in this study because of its exceptional prediction capability, as well as its capacity to manage complicated feature relationships and missing data. In addition to Random Forest's stability and ANN's structural depth, GBM offers a highly accurate boundary for diabetic risk assessment by concentrating on the "hard-to-classify" cases in the training set.

In the two-stage model, GBM acts as the second-stage classifier: Samples predicted as Outcome = 0 (absence of diabetes) in logistic regression are then passed to a secondary GBM classifier. If the secondary model predicts 1, the final output is Outcome = 1; otherwise, Outcome = 0.

The final diagnosis of diabetes is the union of positive predictions from both stages (A + B in **Figure 4**), thereby reducing the false negative rate (the rate of missed diagnoses). Instead of aiming only to achieve higher accuracy, this two-stage model combines a high-confidence positive identification stage with GBM classification for the remaining cases.

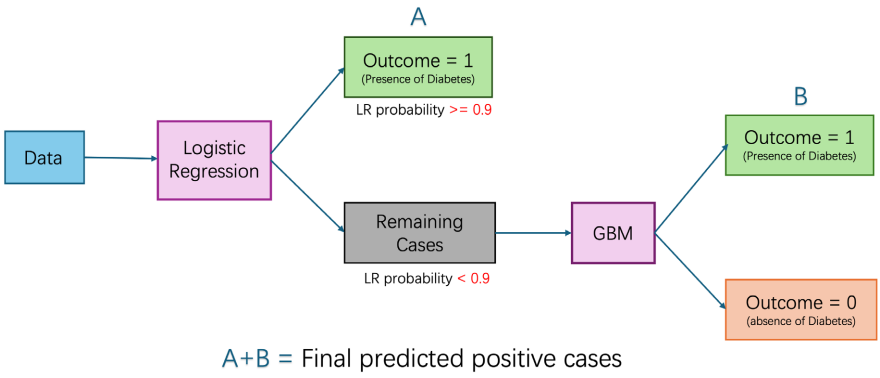



**Figure 4.** Procedure of the two-stage combined model.

Detailed preprocessing and model specifications are summarized in Table 3.

Table 3. Model specifications.

| Item | Specification |
|---|---|
| Missing-value recoding | Glucose, BloodPressure, SkinThickness, Insulin, BMI: 0 → NaN |
| Imputer settings | IterativeImputer; RandomForestRegressor (n_estimators = 10); max_iter = 10; initial_strategy = mean; sample_posterior = False |
| Scaling procedures | StandardScaler for ANN only; no scaling for RF, SVM, LR, or GBM |
| GBM and LR hyperparameters | **GBM**: learning_rate = 0.0216, max_depth = 4, max_features = None, min_samples_leaf = 1, min_samples_split = 9, n_estimators = 177, subsample = 0.8071. **LR**: penalty = L2, C = 1.0, solver = liblinear, max_iter = 1000. |
| Random seed | 42 |
| Software versions | Python 3.12.13; NumPy 2.0.2; pandas 2.2.2; scikit-learn 1.6.1; TensorFlow 2.20.0. |

The training-only Stage 1 threshold comparison is presented in Table 4.

Table 4. Training-only stage 1 threshold comparison.

| Threshold | Accuracy (%) | False Negative Rate (%) | False Positive Rate (%) |
|---|---|---|---|
| 0.50 | 76.38 | 34.58 | 17.75 |
| 0.60 | 76.71 | 37.85 | 15.50 |
| 0.70 | 77.04 | 38.32 | 14.75 |
| 0.75 | 77.04 | 38.32 | 14.75 |
| 0.80 | 76.87 | 38.79 | 14.75 |
| **0.90** | **76.87** | **38.79** | **14.75** |

0.90 was selected as the threshold as a conservative high-confidence rule-in cut-off, allowing only the most confident positive cases to be classified directly in Stage 1 while passing all remaining cases to the GBM.

### 4.3. SHAP

Having established the two-stage model to reduce the false negative rate, model interpretability is also equally important in order to identify which variables most strongly drive the model's risk predictions. SHAP (SHapley Additive exPlana-

tions) is a game-theoretic approach used to explain the outputs of any machine learning model by decomposing each prediction into additive contributions from individual features [34]. Specifically, SHAP treats each feature as a player in a cooperative game. The core idea of SHAP is to decompose a model's prediction into the sum of contributions from individual features. For a given prediction, SHAP calculates each feature's contribution by considering all possible combinations of feature values. Unlike conventional importance measures, this method captures both the magnitude and the direction of each feature's effect.

To efficiently compute SHAP values for tree-based ensembles such as GBM, this study employs TreeExplainer, an algorithm that provides tools for visualizing both global feature importance and the distribution of feature-level effects across individual samples [24]. In this study, SHAP is applied on the GBM model on the training set. Then, TreeExplainer computes the SHAP value of each feature for every sample. The global importance of each feature is calculated using the mean absolute SHAP value across all samples: larger SHAP values indicate greater overall influence on the model predictions.

## 5. Results Analysis

Results analysis includes two perspectives: first, the results from the 80% training set and 20% internal validation set drawn from the Pima Indians Diabetes Dataset and the external test set drawn from the Frankfurt Diabetes Dataset are used to examine the robustness and transferability of the two-stage model; second, SHapley Additive exPlanations (SHAP) and Gradient Boosting Machine (GBM)-based feature attribution analysis are used to enhance the interpretability of the two-stage model.

### 5.1. Model Evaluation

The model was first trained on the 80% training set drawn from the Pima Indians Diabetes Dataset and evaluated on the remaining 20% internal validation set to ensure model robustness. In addition, to further guarantee the model's transferability, the Frankfurt Diabetes Dataset, an independent dataset of 2000 female patients collected from a hospital in Frankfurt, Germany, was used as the external test set.

In Table 5, the internal validation set shows that the two-stage model achieved an overall accuracy of 77.27% with a false negative rate of 38.89% (21 out of 54 actual positive cases). In Table 6, the external test set shows that the model achieved an overall accuracy of 89.35% and a lower false negative rate of 20.47% (140 out of 684 actual positive cases). Moreover, the missed positive cases accounted for only 13.64% of the internal validation set (21/154) and 7.00% of the external test set (140/2000), further demonstrating the effectiveness of the model. Thus, high accuracy and low false negative rates in both the internal validation set and the external test set together support the robustness and transferability of the two-stage model.

Table 5. Confusion matrix for the two-stage model on the internal validation set drawn from the Pima Indians Diabetes Dataset (n = 154).

| | Predicted Negative | Predicted Positive |
|---|---|---|
| Actual Negative | 86 (TN) | 14 (FP) |
| Actual Positive | 21 (FN) | 33 (TP) |

Table 6. Confusion matrix for the two-stage model on the external test set drawn from the Frankfurt Diabetes Dataset (n = 2000).

| | Predicted Negative | Predicted Positive |
|---|---|---|
| Actual Negative | 1243 (TN) | 73 (FP) |
| Actual Positive | 140 (FN) | 544 (TP) |

## 5.2. SHAP Analysis

To improve upon the limitations of the impurity-based GBM importance measure, SHAP (SHapley Additive exPlanations) values were computed using TreeExplainer, which decomposes each individual prediction into additive feature contributions grounded in cooperative game theory [34]. This approach not only ranks feature importance but also reveals the direction and distribution of each feature's effect across all samples [24].

Figure 5 ranks Glucose first, followed by Insulin, BMI, DiabetesPedigreeFunction, and Age. These variables are all important features in clinical interpretation, which is broadly consistent with established medical knowledge.

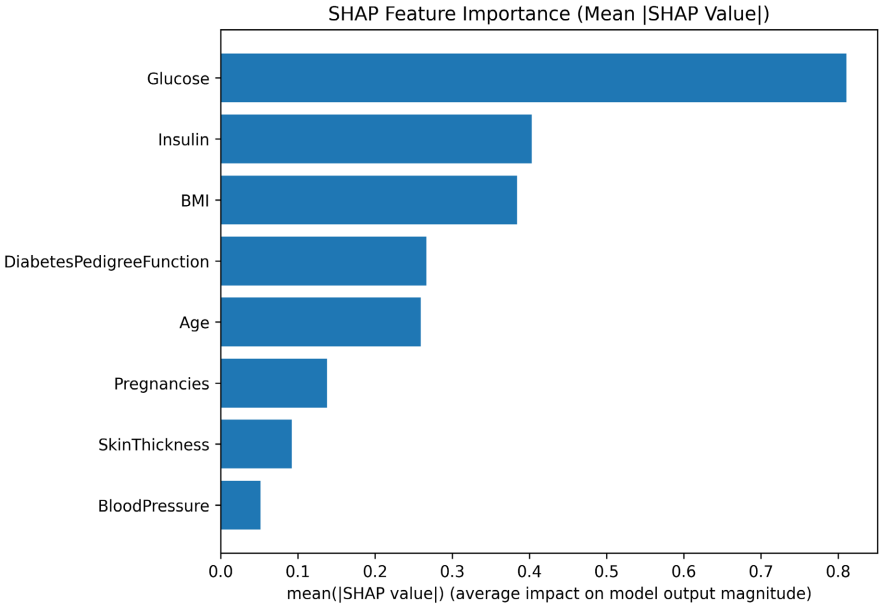



Figure 5. SHAP feature importance based on mean absolute SHAP values for the GBM model.

Pregnancies, SkinThickness, and BloodPressure_have lower mean absolute SHAP values and contribute relatively less to the overall model predictions. This illustrates that these features are not the main contributors to the model's predictions.

In Figure 6, the Beeswarm plot further demonstrates the direction and the distribution of each feature's contribution. Glucose displays a clear directional pattern: as the glucose value increases, SHAP value also increases, with higher glucose values generally corresponding to more positive SHAP values. BMI shows a similar but milder pattern. In contrast, Age and Insulin exhibit a more dispersed color distribution, suggesting their influence on risk is less consistent across all individuals; instead, they may depend on interactions with other variables. The remaining features generally show SHAP values closer to 0, indicating a comparatively weak and inconsistent contribution to individual predictions. These results corroborate the results of the feature-importance plot that Glucose and Insulin are the most important factors.

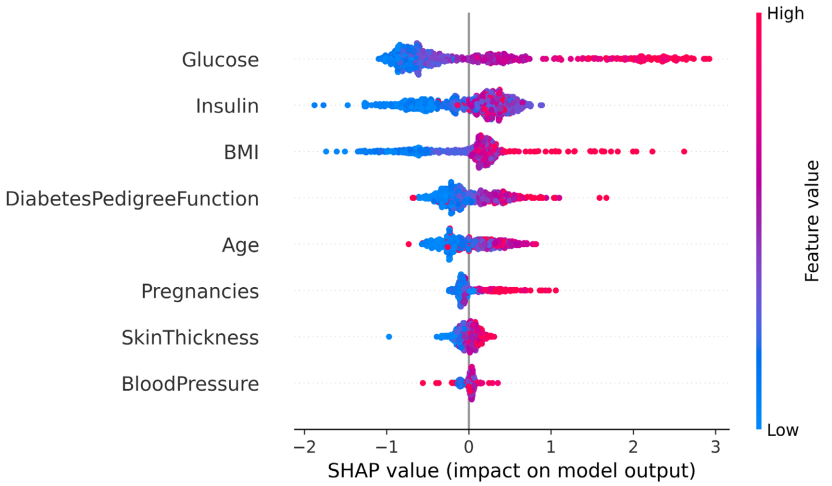


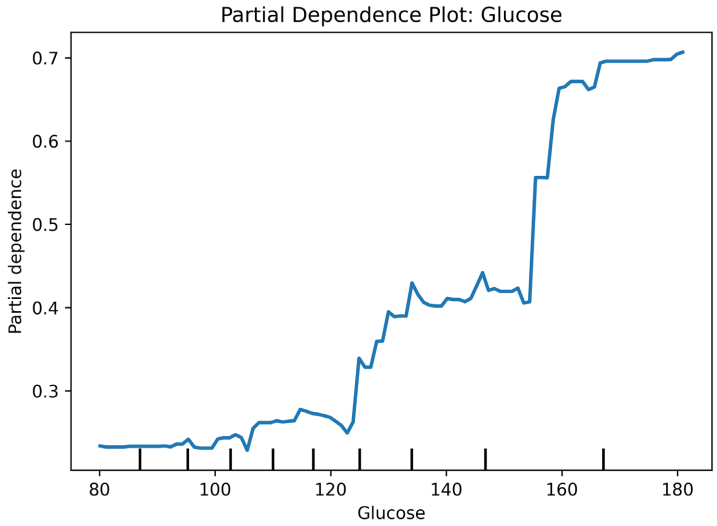

Figure 6. SHAP beeswarm plot showing the magnitude and direction of feature contributions to individual GBM predictions.

## 5.3. Partial Dependence Plots

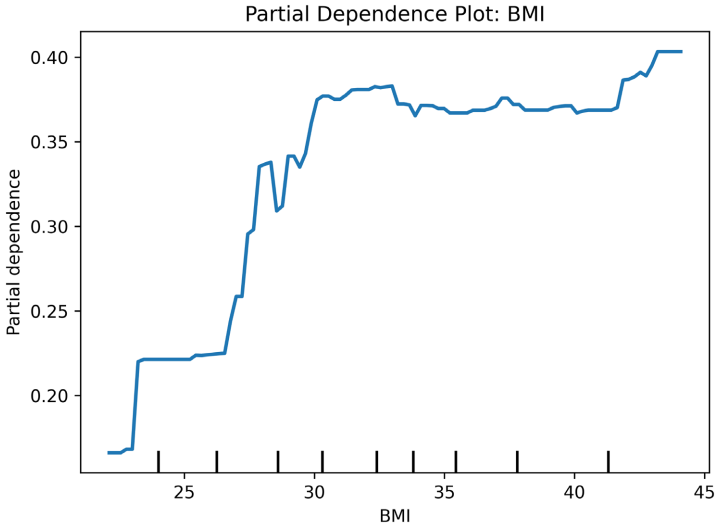
Unlike SHAP analysis in 5.2, the partial dependence (5.3) and interaction plots (5.4) employ the classical partial dependence approach [32]: for each feature being analyzed, all other features remain constant at original values; and the resulting predictions for all the samples are averaged to estimate the feature's overall effect on the model.

Partial dependence plots illustrate the effect of individual features on the model's performance, when holding all other variables constant. The y-axis represents the partial dependence value—higher values indicate higher model-predicted probability of diabetes.

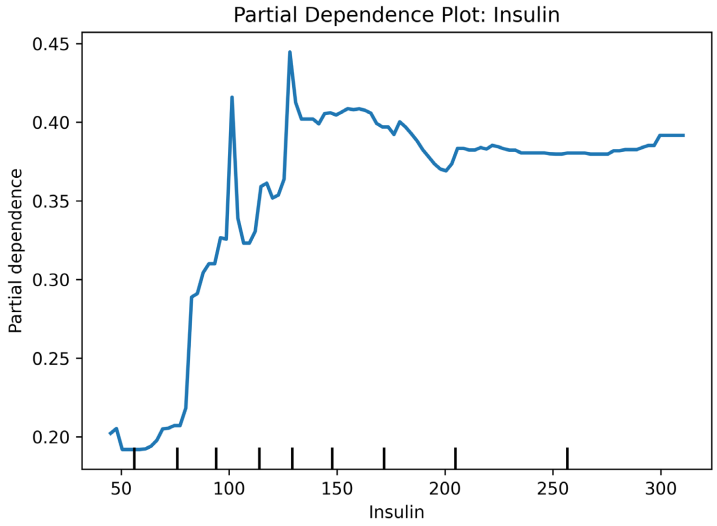
As shown in Figure 7, the partial dependence plots reveal nonlinear patterns across Glucose, BMI, Insulin, and Age.


Partial Dependence Plot: Glucose
Partial dependence
0.7
0.6
0.5
0.4
0.3
80
100
120
140
160
180
Glucose

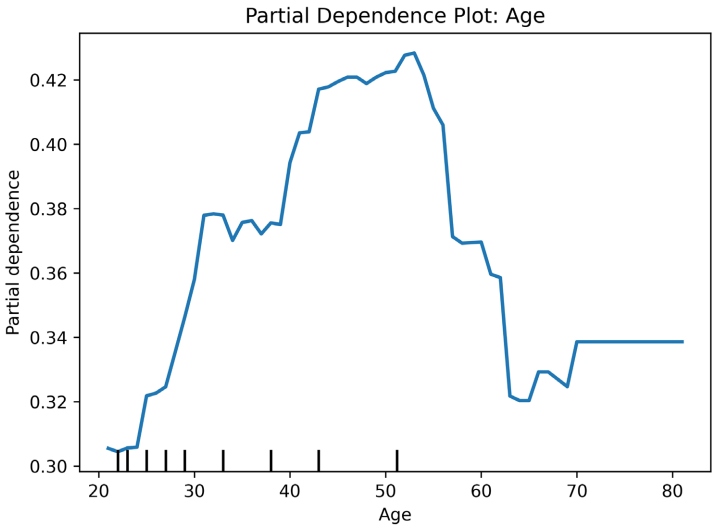
(a)


Partial Dependence Plot: BMI
Partial dependence
0.40
0.35
0.30
0.25
0.20
25
30
35
40
45
BMI

(b)


Partial Dependence Plot: Insulin
Partial dependence
0.45
0.40
0.35
0.30
0.25
0.20
50
100
150
200
250
300
Insulin

(c)




(d)

**Figure 7.** Partial dependence plots for Glucose, BMI, Insulin, and Age. (a) Glucose; (b) BMI; (c) Insulin; (d) Age.

Glucose: The overall trend shows that there's a positive correlation between glucose and risk of diabetes. However, the relationship is not a perfect straight line, but increases through several thresholds. Risk first starts to increase from 105 - 110 mg/dL; then it sharply increases at 125 - 130 mg/dL; it finally reaches a very high risk level at 150 - 160 mg/dL.

BMI: The BMI plot shows an initial increase near 23 kg/m$^2$ and a second increase around 27 - 30 kg/m$^2$. Below BMI at 23 kg/m$^2$, risk is constantly low; then the risk progressively rises through (23 kg/m$^2$ - 30 kg/m$^2$) and levels out over BMI 30 kg/m$^2$ - 45 kg/m$^2$.

Insulin: Below approximately 80 μU/mL, risk remains relatively low. It then rises with several fluctuations between approximately 80 and 140 μU/mL, remains at a relatively high level until around 170 μU/mL, declines toward 200 μU/mL, and then stabilizes with a slight increase near the upper end of the plotted range.

Age: From the ages of 20 to 50, risk increases steadily, reaching a local maximum in the early 50s. It then declines through the late 50s and early 60s, followed by a modest rebound around age 70 and a relatively stable pattern at older ages.

## 5.4. Interaction Plots

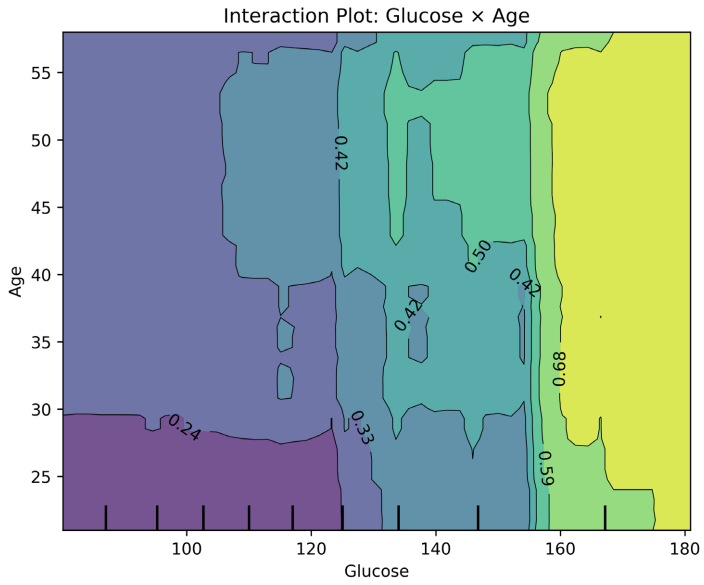
Interaction plots visualize the joint effect of two features on predicted diabetes risk through heatmaps. The color gradient represents partial dependence values, with yellow/bright regions indicating high risk and blue/dark regions indicating low risk.

**Figure 8** presents the interaction plots for Glucose × Age, Glucose × BMI, Glucose × Insulin, and Age × Insulin.

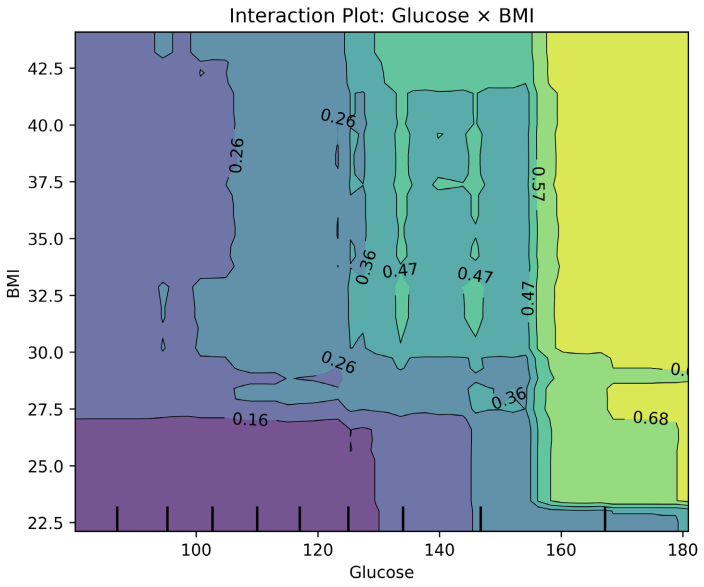
Glucose × Age: The impact of Glucose is highly significant. As Glucose levels increase from 80 mg/dL to 180 mg/dL, the predicted risk increases significantly,

with the highest predicted probabilities at the highest Glucose values. The influence of Age is relatively weak. When Glucose level stays the same, as age increases, the risk does not change significantly. However, in each glucose level, age between 25 - 30 shows a change in the risk, showing a clear interaction effect—Glucose's impact on the outcome is dependent individual's age.

Glucose × BMI: In this plot, Glucose still stands out as an important risk factor: as Glucose increases from 80 mg/dL to 180 mg/dL, the risk increases dramatically. The influence of BMI shows a "threshold" characteristic: within the BMI range of approximately 27.5 to 30, the contour lines undergo a sharp transition, indicating a non-linear shift in the model's response. This shows a clear interaction effect—Glucose's impact on the outcome is dependent on the individual's BMI.

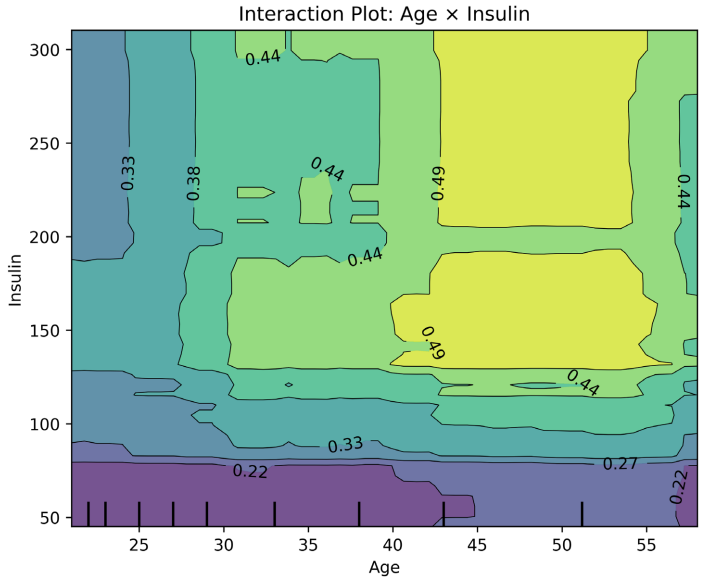



(a)




(b)


(c)




(d)

**Figure 8.** Interaction plots for selected feature pairs in the GBM model. (a) Glucose × Age; (b) Glucose × BMI; (c) Glucose × Insulin; (d) Age × Insulin.

Glucose × Insulin: Similarly, Glucose is an important risk factor: as Glucose increases from 80 mg/dL to 180 mg/dL, the risk increases dramatically. Thresholds occur particularly when Insulin is below 100 and around 200. This shows a clear interaction between Glucose and Insulin.

Age × Insulin: There is a significant increase in risk along the diagonal axis from the bottom-left to the top-right. This means when Age and Insulin Levels both increase simultaneously, the risk increases; when each variable increases independently, the risk almost stays the same. Thus, this plot highlights a strong interaction between Age and Insulin.

## 6. Discussion

This study proposes a sequential two-stage model that reaches high prediction accuracy and low false negative rates which is vital in clinical screening. The model enhances clinical interpretability significantly by showing key interactions between variables and key thresholds in features independently.

In addition, many models are trained and tested on the same dataset. In contrast, the sequential two-stage framework proposed in this study was evaluated on the two entirely independent datasets: the external test set drawn from the Frankfurt Diabetes Dataset differs from the training set drawn from the Pima Indians Diabetes Dataset in both geographic location and patient population. Hence, the high overall performance and low false negative rate support the model's robustness and generalizability.

On top of that, the model utilizes SHAP feature importance and Beeswarm Plot, PDPs, and interaction analysis to enhance model interpretability. The results indicate importance ranking and several important relationships between variables. These methods and graphs are helpful for clinical guidelines, helping address the black-box nature of GBM by interpreting feature effects and interactions, making the model more transparent and interpretable.

These findings also serve as a precursor to causal inference. In the future, to deepen the causal analysis, more iterations will move beyond associative patterns. Causal discovery algorithms can be used to further validate the results and enhance interpretability significantly. With the causal inference frameworks, the model can provide more robust evidence for clinical guidelines and personalized clinical decision-making strategies.

## 7. Conclusions

Diabetes mellitus is a growing global public health threat, over 70% of adults in the United States are currently classified as overweight or obese [35]. Thus, timely detection and appropriate clinical management are critical to prevent severe complications. This study proposed a sequential two-stage Logistic Regression-Gradient Boosting framework for diabetes classification. By implementing an ensemble machine learning method, this study provided key information for clinical diagnosis:

- Glucose and BMI are the two most important risk factors, followed by Age and Insulin, which were consistent with modern research of Type 2 diabetes.
- The study identified key thresholds in several variables, including Glucose at 127 - 130 mg/dL, BMI at 30 kg/m$^2$, Insulin at 75 μU/mL - 150 μU/mL.
- Key interactions were identified: Glucose and BMI jointly amplify diabetic risk; Age and Insulin jointly increase diabetic risk.

The proposed model is considered advanced due to the following characteristics:

- On the external test set drawn from the Frankfurt Diabetes Dataset, the model achieved an overall high accuracy of 89.35%, and a small false negative rate of

20.47%, showing strong model transferability.

- The two-stage structure is advantageous for imbalanced datasets because it first identifies high-confidence positive cases and then applies GBM to the remaining samples, providing an additional opportunity to detect positive cases and reduce false-negative classifications.

Given this robust transferability, future work should focus on large-scale clinical validation through hospital partnerships to further assess real-world applicability. The two-stage model achieved high accuracy and low false negative rate on both internal validation dataset and external test dataset. Given this robust transferability, in the future, the research will collect larger and more diverse diabetes datasets to further evaluate the model's real-world applicability.

## Author Contributions

Moyang Li: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Project Administration, Validation, Writing—original draft, Writing—review & editing. Bowen Cai: Advising, Review, and Editing.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] World Health Organization (2024) Diabetes. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Zhou, B., Rayner, A.W., Gregg, E.W., Sheffer, K.E., Carrillo-Larco, R.M., Bennett, J.E., *et al.* (2024) Worldwide Trends in Diabetes Prevalence and Treatment from 1990 to 2022: A Pooled Analysis of 1108 Population-Representative Studies with 141 Million Participants. *The Lancet*, **404**, 2077-2093. https://doi.org/10.1016/s0140-6736(24)02317-1

[3] Meyers, M., Lerner, B.W. and Tate, P. (2021) Diabetes Mellitus. In: Nemeh, K.H. and Longe, J.L., *The Gale Encyclopedia of Science* (6th ed., Vol. 2), Gale, 1338-1341.

[4] CDC (2024) Diabetes Risk Factors. https://www.cdc.gov/diabetes/risk-factors/index.html

[5] National Institute of Diabetes and Digestive and Kidney Diseases (2022) Risk Factors for Type 2 Diabetes. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes

[6] Wilcox, G. (2005) Insulin and Insulin Resistance. *Clinical Biochemist Reviews*, **26**, 19-39.

[7] OpenML (n.d.) Pima Indians Diabetes Database. OpenML Dataset 37. https://www.openml.org/d/37

[8] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H. (2018) Predicting Diabetes Mellitus with Machine Learning Techniques. *Frontiers in Genetics*, **9**, Article 515. https://doi.org/10.3389/fgene.2018.00515

[9] Kiran, M., Xie, Y., Anjum, N., Ball, G., Pierscionek, B. and Russell, D. (2025) Machine

Learning and Artificial Intelligence in Type 2 Diabetes Prediction: A Comprehensive 33-Year Bibliometric and Literature Analysis. *Frontiers in Digital Health*, **7**, Article 1557467. https://doi.org/10.3389/fdgth.2025.1557467

[10] Abousaber, I., Abdallah, H.F. and El-Ghaish, H. (2025) Robust Predictive Framework for Diabetes Classification Using Optimized Machine Learning on Imbalanced Datasets. *Frontiers in Artificial Intelligence*, **7**, Article 1499530. https://doi.org/10.3389/frai.2024.1499530

[11] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., *et al.* (2019) Global and Regional Diabetes Prevalence Estimates for 2019 and Projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th Edition. *Diabetes Research and Clinical Practice*, **157**, Article 107843. https://doi.org/10.1016/j.diabres.2019.107843

[12] Zheng, Y., Ley, S.H. and Hu, F.B. (2017) Global Aetiology and Epidemiology of Type 2 Diabetes Mellitus and Its Complications. *Nature Reviews Endocrinology*, **14**, 88-98. https://doi.org/10.1038/nrendo.2017.151

[13] Ruze, R., Liu, T., Zou, X., Song, J., Chen, Y., Xu, R., *et al.* (2023) Obesity and Type 2 Diabetes Mellitus: Connections in Epidemiology, Pathogenesis, and Treatments. *Frontiers in Endocrinology*, **14**, Article 1161521. https://doi.org/10.3389/fendo.2023.1161521

[14] Choi, S.G., Oh, M., Park, D., Lee, B., Lee, Y., Jee, S.H., *et al.* (2023) Comparisons of the Prediction Models for Undiagnosed Diabetes between Machine Learning versus Traditional Statistical Methods. *Scientific Reports*, **13**, Article No. 13101. https://doi.org/10.1038/s41598-023-40170-0

[15] Ijaz, M., Alfian, G., Syafrudin, M. and Rhee, J. (2018) Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority over Sampling Technique (SMOTE), and Random Forest. *Applied Sciences*, **8**, Article 1325. https://doi.org/10.3390/app8081325

[16] Hasan, M.K., Das, D., Hossain, E., Hasan, M. and Alam, M.A. (2020) Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access*, **8**, 76516-76531. https://ieeexplore.ieee.org/document/9076634

[17] Ganie, S.M., Pramanik, P.K.D., Bashir Malik, M., Mallik, S. and Qin, H. (2023) An Ensemble Learning Approach for Diabetes Prediction Using Boosting Techniques. *Frontiers in Genetics*, **14**, Article 1252159. https://doi.org/10.3389/fgene.2023.1252159

[18] Bala Manoj Kumar, P., Srinivasa Perumal, R., Nadesh, R.K. and Arivuselvan, K. (2020) Type 2: Diabetes Mellitus Prediction Using Deep Neural Networks Classifier. *International Journal of Cognitive Computing in Engineering*, **1**, 55-61. https://www.sciencedirect.com/science/article/pii/S2666307420300073

[19] Edlitz, Y. and Segal, E. (2022) Prediction of Type 2 Diabetes Mellitus Onset Using Logistic Regression-Based Scorecards. *eLife*, **11**, e71862. https://elifesciences.org/articles/71862

[20] Vakil, V., Pachchigar, S., Chavda, C. and Soni, S. (2021) Explainable Predictions of Different Machine Learning Algorithms Used to Predict Early Stage Diabetes. Cornell University. https://arxiv.org/abs/2111.09939

[21] Talebi Moghaddam, M., Jahani, Y., Arefzadeh, Z., Dehghan, A., Khaleghi, M., Sharafi, M., *et al.* (2024) Predicting Diabetes in Adults: Identifying Important Features in Unbalanced Data over a 5-Year Cohort Study Using Machine Learning Algorithm. *BMC Medical Research Methodology*, **24**, Article No. 220. https://doi.org/10.1186/s12874-024-02341-z

[22] Cowie, C.C. (2019) Diabetes Diagnosis and Control: Missed Opportunities to Improve Health. *Diabetes Care*, **42**, 994-1004. https://doi.org/10.2337/dci18-0047

[23] Seto, H., Oyama, A., Kitora, S., Toki, H., Yamamoto, R., Kotoku, J., *et al.* (2022) Gradient Boosting Decision Tree Becomes More Reliable than Logistic Regression in Predicting Probability for Diabetes with Big Data. *Scientific Reports*, **12**, Article No. 15889. https://doi.org/10.1038/s41598-022-20149-z

[24] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al.* (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[25] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C. and Johannes, R.S. (1988) Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In: *Proceedings of the Symposium on Computer Applications in Medical Care*, IEEE Computer Society Press, 261-265.

[26] Dasilva, J. (2018) Diabetes. Kaggle. https://www.kaggle.com/datasets/johndasilva/diabetes?select=diabetes.csv

[27] Stekhoven, D.J. and Bühlmann, P. (2011) MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics*, **28**, 112-118. https://doi.org/10.1093/bioinformatics/btr597

[28] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[29] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[30] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1023/a:1022627411411

[31] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Representations by Back-Propagating Errors. *Nature*, **323**, 533-536. https://doi.org/10.1038/323533a0

[32] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451

[33] Hastie, T., Tibshirani, R. and Friedman, J. (2017) The Elements of Statistical Learning. 2nd Edition, Springer.

[34] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, **30**, 4765-4774. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[35] National Center for Health Statistics (2026) Obesity and Overweight. National Center for Health Statistics. https://www.cdc.gov/nchs/fastats/obesity-overweight.htm