An Interpretable XGBoost Model for Diabetes Prediction: Nested Cross-Validation, Calibration, and SHAP Analysis

Zeynep Kucukakcali [1*], Ipek Balikci Cicek [2]
[1,2] Inonu University Faculty of Medicine, Department of Biostatistics and Medical Informatics, 44280, Malatya, Turkey

* Corresponding Author: **Zeynep Kucukakcali**

Article Info

ISSN (online): 2582-8940
Volume: 07
Issue: 03
Received: 07-06-2026
Accepted: 06-07-2026
Published: 05-08-2026
Page No: 168-174

Abstract

Background: Diabetes mellitus is a major global health burden, and its early detection is essential for preventing serious complications. Machine learning, and eXtreme Gradient Boosting (XGBoost) in particular, performs strongly on routine clinical data; however, reported results are frequently optimistic because of hold-out evaluation and data leakage, and the resulting models are often difficult to interpret.

Objective: To develop and rigorously evaluate an interpretable XGBoost model for predicting the presence of diabetes from eight routine diagnostic measurements, using an unbiased, leakage-controlled evaluation design.

Methods: An openly available dataset of 1,168 patient records (771 non-diabetic and 397 diabetic; an approximately 66:34 class imbalance) with eight clinical features was analysed. Physiologically implausible zero values were treated as missing and imputed with the median inside a processing pipeline. Model selection and performance estimation were separated using a 5×5 nested cross-validation scheme, with all preprocessing confined to the training folds to prevent data leakage. Performance was assessed on out-of-fold predictions using accuracy, sensitivity, specificity, precision, F1 score, the area under the ROC curve (ROC-AUC), and the Brier score; probability calibration was examined; and the contribution of each feature was quantified with SHapley Additive exPlanations (SHAP).

Results: The model achieved an ROC-AUC of 0.823, an accuracy of 0.769, a specificity of 0.859, a sensitivity of 0.595, a precision of 0.684, an F1 score of 0.636, and a Brier score of 0.160. The near-diagonal calibration curve and low Brier score indicated that the predicted probabilities were well calibrated and could be interpreted as reliable risk estimates. SHAP analysis identified glucose, body mass index, and age as the most influential predictors, in agreement with established clinical risk factors for diabetes.

Conclusions: Under a leakage-controlled, unbiased evaluation, XGBoost provided moderate but trustworthy discrimination together with well-calibrated probabilities for diabetes prediction, while SHAP confirmed clinically plausible predictors. The comparatively conservative performance underscores the importance of nested cross-validation over simple hold-out splits for realistic model assessment. External validation and decision-threshold or class-imbalance strategies represent promising directions for future work.

DOI: https://doi.org/10.54660/IJMBHR.2026.7.3.168-174

Keywords: diabetes prediction, XGBoost, nested cross-validation, SHAP, model calibration, explainable artificial intelligence, machine learning

1. Introduction

Diabetes mellitus is a chronic metabolic disorder that has become one of the foremost public-health challenges of our time. The 11th edition of the International Diabetes Federation (IDF) Diabetes Atlas estimates that around 589 million adults aged 20–79 years — approximately one in nine — were living with diabetes in 2024, a figure projected to rise to 853 million by 2050 [1].

The disease accounted for an estimated 3.4 million deaths and more than one trillion US dollars in health expenditure in 2024 [1], and a large share of affected individuals remain undiagnosed, delaying treatment and increasing the risk of complications such as cardiovascular disease, nephropathy, retinopathy, and neuropathy [2]. Because many of these complications can be delayed or prevented when the condition is detected early, the timely identification of individuals at elevated risk has become a priority for preventive care.

Traditional risk assessment generally depends on a small set of biomarkers and fixed clinical cut-off values, which may not adequately capture the complex, nonlinear relationships among physiological, demographic, and metabolic factors [3]. Machine learning (ML) provides a data-driven alternative that can learn such relationships directly from routine diagnostic measurements. Recent systematic reviews indicate that ensemble-based methods often attain accuracies in the range of 85–90% and areas under the ROC curve above 0.90, surpassing conventional statistical models [3], while bibliometric analyses point to a growing emphasis on interpretability and multimodal data integration in this research area [4]. Publicly available datasets that comprise the same eight routinely collected features used here — most notably the widely studied Pima Indians Diabetes Database — have long served as benchmarks for developing and comparing such models, while also exposing recurrent difficulties such as missing values and class imbalance [5].

Among ML algorithms, Extreme Gradient Boosting (XGBoost) has proven especially effective for structured (tabular) clinical data, thanks to its regularized gradient-boosted tree framework, its robustness to heterogeneous features, and its consistently strong predictive accuracy [6, 7]. Despite these strengths, two methodological shortcomings frequently undermine the reliability and clinical value of reported results. First, when hyperparameters are tuned and performance is measured on the same data, the resulting estimates are optimistically biased; obtaining an unbiased estimate requires that model selection and performance evaluation be kept separate, as achieved by nested cross-validation [8, 9]. Second, high-performing ensemble models are often perceived as opaque "black boxes," which limits clinical trust and adoption; post-hoc interpretability techniques such as SHapley Additive exPlanations (SHAP) mitigate this concern by attributing each prediction to its contributing features on a sound theoretical basis [10, 11]. Addressing both issues within a single, transparent modelling pipeline remains an open need.

The aim of this study is to develop and rigorously evaluate an interpretable XGBoost model for predicting the presence of diabetes from eight routine diagnostic measurements, using an openly available dataset [12]. To this end, the study pursues three specific objectives: (i) to obtain an unbiased estimate of predictive performance through a 5×5 nested cross-validation design in which all preprocessing steps are confined to the training folds, thereby preventing data leakage; (ii) to assess not only the model's discriminative ability but also the reliability of its predicted probabilities by means of calibration analysis; and (iii) to enhance clinical interpretability by quantifying the contribution of each clinical feature to the model's predictions using SHAP. In pursuing these objectives, the study seeks to address the optimistic bias and limited transparency that characterize many previous XGBoost-based diabetes prediction models.

The remainder of the paper is organized as follows: Section 2 describes the dataset and methodology, and Section 3 presents the experimental results.

2. Materıals And Methods

2.1. Dataset

This study used a dataset comprising 1,168 patient records, designed to predict the presence of diabetes based on routine diagnostic measurements. Each record contains eight clinical features collected during standard health screenings, together with a binary outcome variable (Outcome) indicating whether the patient was diagnosed with diabetes.

The independent variables in the dataset were as follows: Pregnancies (number of pregnancies), Glucose (plasma glucose concentration from a 2-hour oral glucose tolerance test, mg/dL), BloodPressure (diastolic blood pressure, mm Hg), SkinThickness (triceps skinfold thickness, mm), Insulin (2-hour serum insulin level, µU/mL), BMI (body mass index, kg/m²), DiabetesPedigreeFunction (a composite score reflecting the likelihood of diabetes based on family history), and Age (patient age in years). The dependent variable, Outcome, was coded as 1 = diabetes diagnosed and 0 = no diabetes.

The dataset consisted of 771 negative (66.0%) and 397 positive (34.0%) cases, corresponding to an approximately 66:34 class imbalance. Patient ages ranged from 21 to 81 years. The data were prepared and exported from VertexMD, a local-first electronic health records application designed for personal medical record tracking and interoperability research.

The dataset is publicly available as open data in the Mendeley Data repository and can be accessed through the following reference: Hernesto, Christian (2026), "Predicting Diabetes From Tracking Medical Records", Mendeley Data, V3, doi: 10.17632/nxnty5g7y6.3.

2.2. Data Preprocessing

Because zero values are not physiologically valid for the Glucose, BloodPressure, SkinThickness, Insulin, and BMI variables, such zeros were treated as unrecorded (missing) measurements rather than true biological zeros and were converted to missing values (NaN). Missing values were filled using median imputation. To prevent data leakage, both the zero-to-missing conversion and the imputation were embedded within a processing pipeline so that all parameters were learned exclusively from the training data at each cross-validation fold. This ensured that no information leaked from the outer test fold into the training fold. In addition, the dataset was verified to contain no duplicate rows and no missing values outside the pipeline-handled ones.

2.3. Classification Model

The XGBoost (Extreme Gradient Boosting) algorithm, a method based on gradient-boosted decision trees, was used as the classifier. The model was configured with the binary:logistic objective function for binary classification, the logloss evaluation metric, and a histogram-based tree construction method (tree_method = "hist"). To ensure reproducibility, all sources of randomness were fixed with a constant seed (random_state = 42). The preprocessing steps and the classifier were combined into a single pipeline following the order: zero-to-NaN transformation → median imputation → XGBoost classifier.

2.4. Nested Cross-Validation

To obtain an unbiased estimate of model performance, a 5 × 5 nested cross-validation scheme (5 outer folds × 5 inner folds) was applied. The outer loop assessed the model's generalization performance, while the inner loop performed hyperparameter selection. Stratified k-fold (StratifiedKFold) with shuffling was used to preserve the class distribution within each fold. In the inner loop, a grid search was conducted with GridSearchCV using ROC-AUC as the scoring criterion. The searched hyperparameter space included the number of trees (n_estimators: 200, 300), maximum tree depth (max_depth: 3, 5), and learning rate (learning_rate: 0.05, 0.10).

In each outer fold, the best model selected in the inner loop generated predictions on the corresponding test samples, yielding out-of-fold (OOF) probability estimates for all samples. A decision threshold of 0.50 was used for class assignment. Final performance metrics were reported based on these OOF predictions covering the entire dataset.

2.5. Performance Evaluation Metrics

Model performance was evaluated using accuracy, sensitivity (recall), specificity, precision, F1 score, the area under the ROC curve (ROC-AUC), and the Brier score. Specificity was calculated from the confusion matrix as the ratio of true negatives to total negatives. To visualize discriminative performance, the ROC curve was plotted; to assess the reliability of the probability estimates, a calibration curve (with 10 quantile bins) was plotted and the Brier score was reported.

2.6. Feature Importance and SHAP Analysis

To interpret the effects of the features on the model output, a SHAP (SHapley Additive exPlanations) analysis was performed. For this purpose, a final model was trained on the entire dataset using the hyperparameter combination most frequently selected across the outer folds, and SHAP values were computed using the TreeExplainer designed for tree-based models. It should be emphasized that this model was trained solely to interpret feature effects, whereas the performance results were reported from the OOF predictions of the nested cross-validation.

2.7. Software Environment

All analyses were carried out in the Python programming language. The scikit-learn pipeline, cross-validation, and metric functions were used for data processing and modeling; the xgboost library for classification; the shap library for interpretability; and matplotlib for visualization.

3. Results

3.1. Nested Cross-Validation Fold Results

The inner-loop grid search selected the same hyperparameter combination (learning_rate = 0.05; max_depth = 3; n_estimators = 200) in four of the five outer folds; only in the second fold was n_estimators = 300 preferred. This consistency indicates that the selected model configuration is stable. The performance metrics for each outer fold are presented in Table 1.

Table 1: XGBoost model performance metrics by outer fold.

Fold	Accuracy	Sensitivity	Specificity	Precision	F1	ROC-AUC
1	0.799	0.570	0.916	0.776	0.657	0.824
2	0.752	0.563	0.851	0.662	0.608	0.832
3	0.774	0.638	0.844	0.680	0.658	0.825
4	0.760	0.608	0.838	0.658	0.632	0.803
5	0.760	0.595	0.844	0.662	0.627	0.833

3.2. Overall Out-of-Fold (OOF) Performance

The overall performance results computed from the out-of-fold (OOF) predictions covering the entire dataset are given in Table 2. The model achieved good discriminative ability, with 76.9% accuracy and a ROC-AUC of 0.823. The specificity (0.859) being markedly higher than the sensitivity (0.595) indicates that the model was more successful at identifying non-diabetic cases than diabetic ones, which is consistent with the class imbalance in the dataset. The low Brier score of 0.160 suggests that the probability estimates were reasonably well calibrated.

Table 2: Overall performance based on nested cross-validation OOF predictions.

Metric	Value
Accuracy	0.769
Sensitivity	0.595
Specificity	0.859
Precision	0.684
F1 score	0.636
ROC-AUC	0.823
Brier score	0.160

3.3. Confusion Matrix

The confusion matrix for the OOF predictions is shown in Figure 1. The model correctly classified 662 of the 771 non-diabetic cases and mislabeled 109 of them as diabetic. Of the 397 diabetic cases, 236 were correctly identified and 161 were missed (false negatives). This matrix numerically illustrates the model's high success in the negative class and the number of cases overlooked in the positive class.

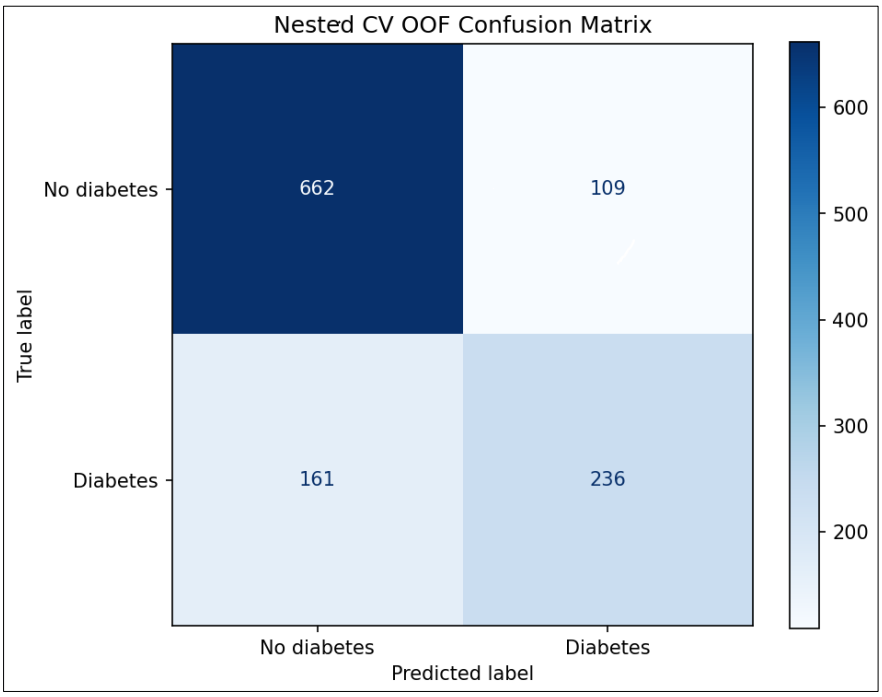

Fig 1: Nested cross-validation OOF confusion matrix.

3.4. ROC Curve

The ROC curve, which reflects the model's discriminative performance, is presented in Figure 2. The area under the curve (AUC) of 0.823 indicates a high probability that the model assigns a greater risk score to a randomly selected diabetic case than to a randomly selected non-diabetic one. The curve lies clearly above the diagonal representing a random classifier.

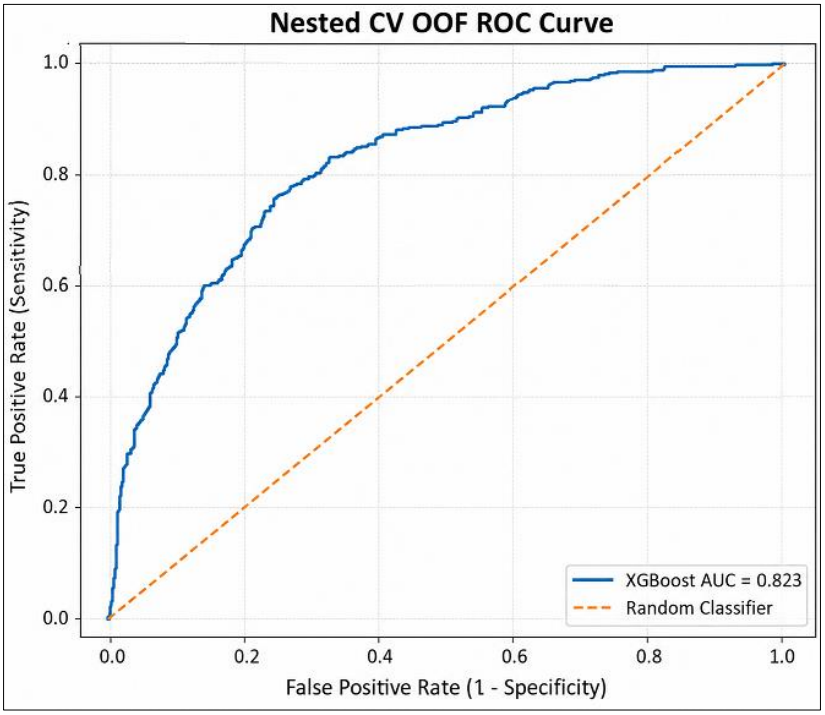

Fig 2: Nested cross-validation OOF ROC curve (AUC = 0.823).

3.5. Calibration Curve

The calibration curve, which evaluates the reliability of the predicted probabilities, is shown in Figure 3. The curve generally follows the diagonal representing perfect calibration closely, indicating that the probability values produced by the model are largely consistent with the observed diabetes rates. The Brier score of 0.160 supports this finding.

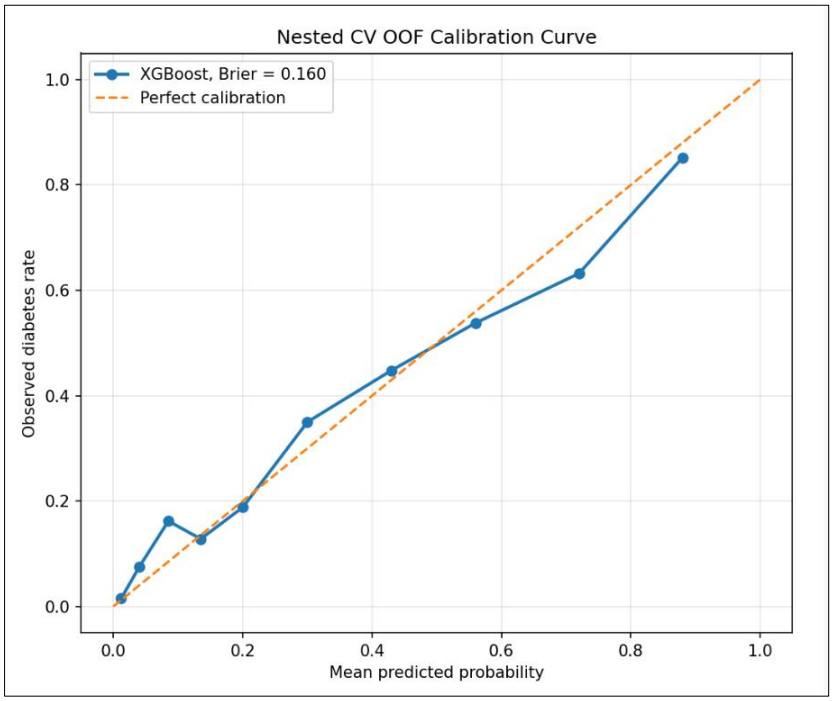

Fig 3: Nested cross-validation OOF calibration curve (Brier = 0.160).

3.6. SHAP Feature Importance Analysis

The results of the SHAP analysis performed on the final model are shown in Figures 4 and 5. According to the mean absolute SHAP values, the most influential feature was Glucose (0.89), followed by BMI (0.58), Age (0.42), Pregnancies (0.31), and DiabetesPedigreeFunction (0.20). Insulin (0.13), BloodPressure (0.10), and SkinThickness (0.03) were the features contributing least to the model's decisions.

The summary (beeswarm) plot in Figure 4 also reveals the direction of the feature effects: high glucose, high BMI, advanced age, and a high number of pregnancies increased the model output toward diabetes (positive SHAP values). These findings are consistent with the known clinical risk factors for diabetes and indicate that the model's predictions are medically interpretable.

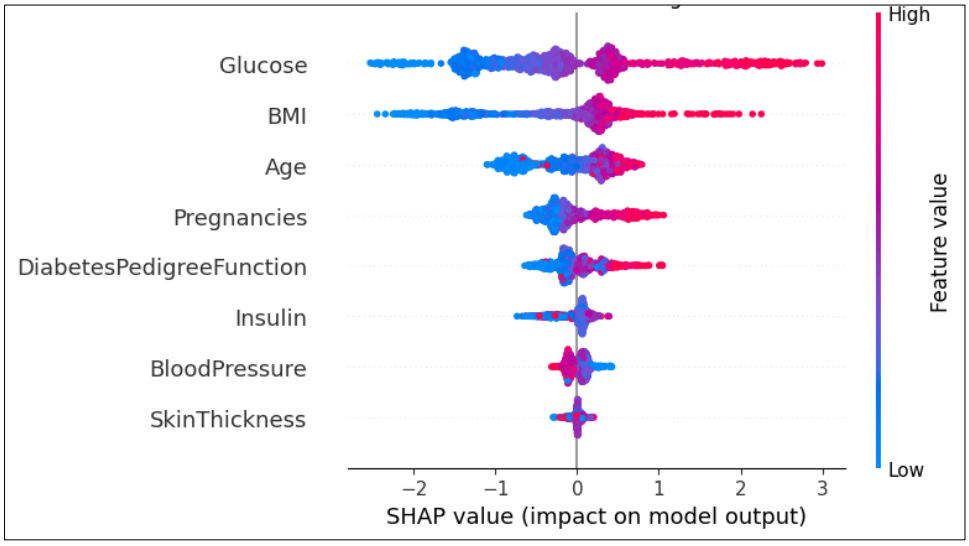

Fig 4: SHAP summary (beeswarm) plot.

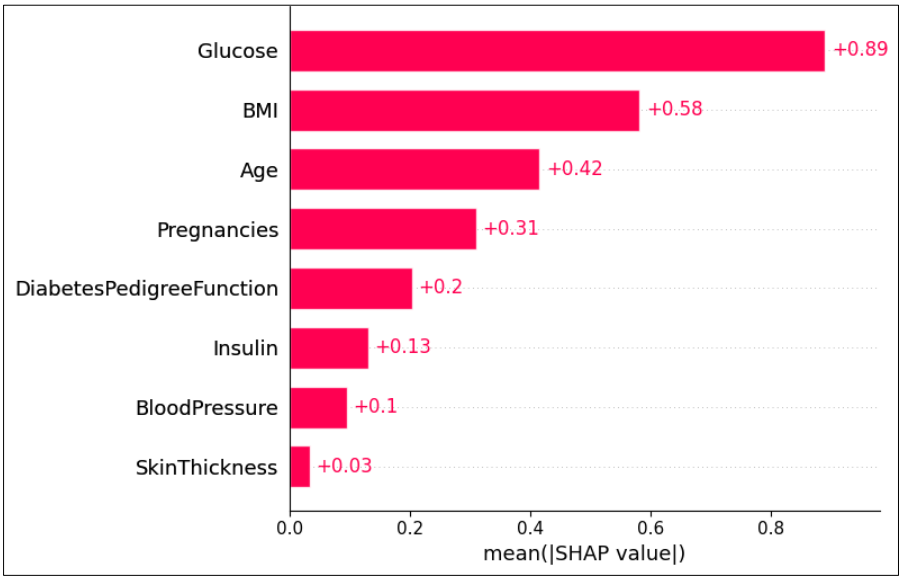

Fig 5: Feature importance ranking based on mean absolute SHAP values.

4. Discussion

This study set out to predict the presence of diabetes from eight routine diagnostic measurements using an XGBoost model, and to do so in a way that is both statistically honest and clinically interpretable. Evaluated through 5×5 nested cross-validation on out-of-fold predictions, the model achieved an ROC-AUC of 0.823, an accuracy of 0.769, a specificity of 0.859, a sensitivity of 0.595, and a Brier score of 0.160, while the SHAP analysis identified glucose, body mass index (BMI), and age as the dominant predictors, followed by the number of pregnancies and the diabetes pedigree function. These headline numbers become more meaningful, however, only when they are placed against what has been reported elsewhere and interpreted in the light of how they were obtained.

Viewed in this way, the discriminative performance obtained here sits comfortably within the range reported for comparable eight-feature diabetes data. An additive–multiplicative neural network evaluated on the Pima Indians Diabetes dataset, for instance, reported an accuracy of about 75.8% and an ROC-AUC of 0.821 [13] figures almost indistinguishable from ours — while a comparison of traditional models on the same data yielded accuracies of roughly 82% for random forest, 78% for gradient boosting, and 76% for logistic regression [14]. At first glance, other studies appear to do considerably better, reporting ROC-AUC values around 0.92–0.93 for random forest and a recursive-feature-elimination GRU model [14]. Yet these more impressive figures are typically produced under conditions that quietly inflate apparent performance: a single train–test split, synthetic oversampling such as SMOTE, and, in some cases, evaluation on the very data used to tune the model [15]. Rather than a shortcoming, then, the more modest estimate reported here is a direct consequence of a deliberately stricter evaluation design.

That design is worth spelling out, because it is the reason our results should be read as conservative rather than disappointing. We chose nested cross-validation over a single hold-out split such as an 80/20 partition for two related reasons. First, a single hold-out estimate depends heavily on which samples happen to land in the test set; for a dataset of roughly 1,168 records, this produces a high-variance figure that can shift substantially from one random split to the next. Second, and more importantly, when hyperparameters are selected by cross-validation and that same score is then reported as the final result, the test data have already influenced model selection, so the estimate becomes optimistically biased [8, 9]. Nested cross-validation removes this bias by nesting an inner loop, used solely for tuning, inside an outer loop used solely for evaluation, so that every outer test fold stays untouched while the model is being chosen. We reinforced this separation by fitting all preprocessing the recoding of physiologically implausible zeros as missing values and their median imputation only on the training folds, thereby closing off the most common route to data leakage. The practical effect is an almost unbiased estimate of generalization performance, which is precisely why our AUC looks lower than that of some hold-out or oversampling-based studies, yet can be trusted further.

If the evaluation design explains how much the model can be believed, the SHAP analysis speaks to why its predictions are believable. The prominence of glucose, BMI, and age echoes both the machine-learning literature, where these same variables repeatedly emerge as the leading predictors on such routine measurements [13, 14], and long-standing clinical understanding of diabetes risk. Reassuringly, the effects also ran in the expected direction: higher plasma glucose, higher BMI, more advanced age, and a greater number of pregnancies each pushed the prediction toward diabetes. This convergence between data-driven importance and established risk factors lends the model a degree of clinical plausibility that a purely accuracy-based account would miss, and it is exactly the kind of transparency increasingly demanded before such tools are trusted in practice.

Interpretability of this kind addresses one half of clinical trust; the reliability of the probabilities themselves addresses the other. Most diabetes prediction studies stop at discrimination metrics such as accuracy and AUC, but a model can rank patients well and still assign probabilities that are systematically too high or too low — a failure of calibration that has been aptly described as the “Achilles heel” of predictive analytics [16]. For this reason we went a step further and examined calibration directly. The near-diagonal calibration curve and the low Brier score of 0.160

indicate that the estimated probabilities correspond reasonably well to observed diabetes frequencies, so they can be read as genuine risk estimates rather than mere ranking scores — a property that also aligns this work with other probability-oriented approaches on the same population, such as Bayesian network models [17].

Reading the outputs as risks also brings the dataset's class imbalance into sharper focus. With roughly a 66:34 split in favour of non-diabetic cases, the model detects non-diabetic individuals more reliably than diabetic ones, as reflected in the gap between its specificity (0.859) and its sensitivity (0.595) at the default 0.50 threshold. We deliberately chose not to correct this with synthetic oversampling, which — when applied outside the cross-validation folds — can introduce leakage and distort the very calibration we sought to preserve; instead, the natural distribution was retained and the folds were stratified. Importantly, the lower sensitivity is not fixed: it can be raised by lowering the decision threshold or by adopting cost-sensitive learning, at the price of some specificity, with the ideal operating point depending on the relative clinical costs of missed cases versus false alarms.

These strengths notwithstanding, the study has clear limitations that temper its conclusions and shape the work ahead. The model was developed and assessed on a single dataset without external validation, so its behaviour on independent populations remains untested; it drew on only eight routine measurements, omitting potentially informative variables such as glycated haemoglobin (HbA1c), lifestyle factors, and longitudinal history; and physiologically implausible zeros were imputed with the median, a principled but simplifying choice. Together with the moderate sensitivity, these factors mean the model is best regarded, for now, as a transparent baseline rather than a stand-alone screening tool. Future work should accordingly pursue external a nd temporal validation, decision-threshold optimization and imbalance-aware methods evaluated strictly within the nested cross-validation to avoid leakage, and the incorporation of richer clinical and lifestyle features — extensions that, building on the honest baseline established here, offer a realistic path toward improving both discrimination and sensitivity.

References

1. Genitsaridi I, Salpea P, Salim A, Sajjadi SF, Tomic D, James S, *et al*. 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. Lancet Diabetes Endocrinol. 2026;14(2):149-56.
2. Teufel F, Orgutsova K, Genitsaridi I, Carrillo-Larco RM, Varghese JS, Marcus ME, *et al*. Global, regional, and national estimates of undiagnosed diabetes in adults: findings from the 2025 IDF Diabetes Atlas. Diabetes Care. 2026;49(3):490-6.
3. Masood S, Al-Bashrawi MA, Khan MA, Dwivedi YK. From data to diagnosis: a systematic review on AI-driven approaches to diabetes prediction. Artif Intell Rev. 2026;59(3):103.
4. Kiran M, Xie Y, Anjum N, Ball G, Pierscionek B, Russell D. Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. Front Digit Health. 2025;7:1557467.
5. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proc Annu Symp Comput Appl Med Care. 1988 Nov 9:261-5.
6. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785-94.
7. Li W, Peng Y, Peng K. Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm. PLoS One. 2024;19(9):e0311222.
8. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91.
9. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res. 2010;11:2079-107.
10. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 2017;30.
11. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, *et al*. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56-67.
12. Hernesto C. Predicting diabetes from tracking medical records [dataset]. Mendeley Data. Version 3. 2020. doi:10.17632/nxnty5g7y6.3.
13. Demirel Tatli Ş, Aytekin K, Agraz M. Enhanced diabetes prediction using novel additive-multiplicative neural networks: a comprehensive machine learning analysis of the PIMA Indians dataset. medRxiv. 2025:2025.09.20.25336250.
14. Nassiwa F, Zeng J. Evaluating traditional machine learning models for predicting diabetes onset using the Pima Indians dataset. SSRN. 2022. Available from: SSRN 4878052.
15. Shams MY, Tarek Z, Elshewey AM. A novel RFE-GRU model for diabetes classification using PIMA Indian dataset. Sci Rep. 2025;15(1):982.
16. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
17. Liang X, Song W, Yang W, Yue Z. Enhancing diabetes risk assessment through Bayesian networks: an in-depth study on the Pima Indian population. Endocr Metab Sci. 2025;17:100212.

How to Cite This Article

Kucukakcali Z, Balikci Cicek I. An interpretable XGBoost model for diabetes prediction: nested cross-validation, calibration, and SHAP analysis. Int J Med All Body Health Res. 2026;7(3):168-174.
doi:10.54660/IJMBHR.2026.7.3.168-174.

Creative Commons (CC) License