Health AI Frontier

OPEN ACCESS JOURNAL

Health AI Frontier

Original Research Article

An Explainable and Imbalance-Aware Hybrid Ensemble Framework for Prevalent Diabetes Detection Using the DiaBD Dataset

Umme Habiba and ***Simone A. Ludwig***

Department of Computer Science, North Dakota State University, Fargo, ND, USA

Received Date : *Aug 05, 2026*
Accepted Date : *Nov 01, 2026*
Published Date : *Nov 13, 2026*

***Corresponding author:** Umme Habiba, Department of Computer Science, North Dakota State University, Fargo, ND, USA, E-mail: umme.habiba@ndsu.edu; ORCID: 0000-0002-7997-3555

Citation: Habiba U, Ludwig SA. An Explainable and Imbalance-Aware Hybrid Ensemble Framework for Prevalent Diabetes Detection Using the DiaBD Dataset. Health AI Front. 2026;1(1):40-55.

Copyright: © 2026 Umme Habiba. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

This paper introduces an explainable and imbalance-aware hybrid ensemble framework for prevalent diabetes detection using the DiaBD dataset, comprising 5,288 participants from Bangladesh. The framework integrates clinically motivated feature engineering, leakage-aware model development, optimized ensemble learning, and comprehensive evaluation. We present four key contributions: (1) the construction of six clinically motivated features capturing interactions among glucose, body mass index, age, and blood pressure, sup-ported by feature-ablation and glucose-exclusion analyses; (2) a leakage-aware and imbalance-sensitive learning strategy incorporating stratified data partitioning, fold-specific preprocessing and resampling, and nested cross-validation; (3) optimized probability-level fusion of five classifiers XGBoost, Random Forest, Gradient Boosting, multilayer perceptron, and LightGBM compared with in-dividual classifiers, alternative weighting strategies, and an independent Feature Tokenizer Transformer (FT-Transformer) baseline; and (4) reliability and interpretability assessment through bootstrap confidence intervals, probability calibration, training-derived operating thresholds, decision-curve analysis, Shapley Additive explanations (SHAP), and Local Interpretable Model-Agnostic Explanations (LIME). On the untouched held-out test set, the ensemble achieved an area under the Receiver Operating Characteristic Curve (ROC-AUC) of 0.8722 (95% Confidence Interval: 0.8246-0.9144) and an average precision of 0.3561. At the training-derived Youden threshold, the model achieved 79.41% sensitivity and 84.85% specificity. Glucose-exclusion analysis reduced ROC-AUC to 0.7801, highlighting the contribution of glucose-related predictors. These findings provide a methodological foundation for reliable evaluation and interpretable prevalent diabetes classification in region-specific healthcare datasets.

Keywords: Prevalent diabetes detection; DiaBD dataset; Hybrid ensemble learning; Imbalance-aware learning; Explainable AI

Introduction

Diabetes mellitus remains a major global public health challenge, contributing substantially to morbidity, mortality, healthcare expenditure, and reduced quality of life. According to the International Diabetes Federation, approximately 589 million adults aged 20 to 79 years were living with diabetes in 2024, a number projected to reach 853 million by 2050 [1]. The burden is particularly concerning in low- and middle-income countries, where limited access to screening, diagnosis, and long-term disease management may increase the risk of preventable complications. Timely identification of individuals with diabetes is therefore important for supporting clinical assessment, treatment planning, and preventive healthcare.

Machine learning has emerged as a promising approach for analyzing complex relationships among demographic, clinical, anthropometric, and medical-history variables. Algorithms such as Random Forest, Gradient Boosting, XG-Boost, and neural networks can capture nonlinear associations and feature interactions relevant to diabetes classification [2]. However, many existing studies rely on widely used benchmark datasets, particularly the PIMA Indian Diabetes dataset, whose restricted demographic representation and population coverage limit the generalizability of their findings [3]. Evaluating machine learning approaches using region-specific datasets is therefore important for understanding their performance in populations underrepresented in existing research.

To support diabetes research in Bangladesh, Prama et al. [4] introduced the DiaBD dataset, containing 5,288 records from individuals aged 21 to 80 years across 63 Bangladeshi unions representing urban, semi-urban, and rural communities. The dataset includes demographic, clinical, anthropometric, and medical-history attributes, including age, glucose measurements, body mass index, blood pressure, hypertension, family history, cardiovascular disease, stroke, and recorded diabetic status. These characteristics provide an opportunity to investigate diabetes classification in a region-specific population. Importantly, because the available outcome represents diabetic status at the time of data collection, the present study addresses prevalent diabetes detection rather than prediction of future disease onset. This work analyzes previously collected data without recruiting participants or conducting new clinical measurements.

Despite its potential, DiaBD introduces important methodological challenges. Of the 5,288 records, only 342 (6.47%) belong to the diabetic class, creating substantial class imbalance. Consequently, high overall accuracy may conceal poor identification of diabetic cases. Although resampling methods have been investigated to improve minority-class classification [5,6], their application requires careful validation. In particular, performing preprocessing or synthetic oversampling before data partitioning can introduce information leakage and produce overly optimistic estimates of generalization. Reliable model assessment therefore requires training only preprocessing and resampling, independent test evaluation, and performance measures that reflect both discrimination and minority-class identification.

Interpretability and clinical relevance present additional challenges. Although ensemble and deep-learning models can capture complex feature relationships, their predictions may be difficult to interpret. Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) provide complementary methods for examining individual predictions and feature contributions [7,8]. Their application to diabetes classification has demonstrated the potential to characterize influential predictors and improve model transparency [9,10]. Nevertheless, explainability alone does not establish clinical reliability, and feature attributions should not be interpreted as causal effects. Probability calibration, clinically relevant operating thresholds, and the consequences of false-positive and false-negative classifications must also be considered when assessing the potential utility of predictive models.

The original DiaBD study contributed a region-specific dataset and initial machine learning analyses [4]. Building on this resource, further investigation is warranted to determine how ensemble learning, clinically motivated feature engineering, imbalance-aware training, and explainability perform under a rigorous validation framework. In particular, the added value of combining multiple classifiers should be assessed against tuned individual models and alternative tabular learning approaches rather than assumed. The influence of engineered features and the dependence of discrimination on glucose-related predictors also require systematic examination, particularly because glucose measurements may be closely related to the recorded diabetes outcome.

To address these challenges, this study develops an explainable and imbalance-aware hybrid ensemble framework for prevalent diabetes detection using the DiaBD dataset. The framework combines XGBoost, Random Forest, Gradient Boosting, multilayer perceptron, and LightGBM through optimized probability-level fusion. Six clinically motivated features are constructed to represent relationships among glucose, body mass index, age, and blood pressure related measurements. Model development employs stratified train-test splitting, fold-specific preprocessing, model dependent resampling, and nested cross-validation. An FT-Transformer is independently evaluated as an attention-based baseline for tabular classification. The evaluation further incorporates individual feature ablation, glucose-exclusion analysis, probability calibration, training-derived operating thresholds, decision curve analysis, and explainability assessment.

The main contributions of this study are summarized as follows:

- We develop an explainable, imbalance-aware hybrid ensemble integrating five complementary classifiers through optimized probability-level fusion, with fold-specific preprocessing, resampling, and nested cross-validation to support leakage-free evaluation.
- We investigate the contributions of six clinically motivated engineered features through individual ablation experiments and assess model dependence on glucose-related information through a separate glucose-exclusion analysis.
- We systematically compare the proposed ensemble with tuned individual classifiers and an independent FT-Transformer baseline, incorporating bootstrap confidence intervals, probability calibration, training-derived operating thresholds, and decision-curve analysis.
- We examine model interpretability using SHAP explanations of the tree-based ensemble components and LIME explanations of representative in-dividual predictions, including an assessment of explanation stability.

The remainder of this paper is organized as follows:

- Reviews related work
- The proposed methodology
- The experimental results
- Discusses the findings, limitations, and clinical implications
- Finally, concludes the paper and outlines future research directions

Related Work

Machine learning for diabetes prediction

Machine learning has been widely investigated for diabetes classification and risk prediction because clinical, demographic, anthropometric, and medical history variables often exhibit complex and nonlinear relationships [11,12]. Early studies using the PIMA Indian Diabetes dataset demonstrated the potential of computational approaches for diabetes-related classification [13,14]. Subsequent research has examined Logistic Regression,

Decision Trees, Support Vector Machines, Naive Bayes, Random Forest, Gradient Boosting, XGBoost, neural networks, and deep learning models [15,16].

These approaches have been applied to diabetes screening, risk assessment, and clinical decision support [17-19]. Tree-based ensembles have received particular attention because they can capture nonlinear relationships and interactions among structured clinical variables. XGBoost, for example, incorporates regularization and scalable optimization to improve gradient-boosted learning [20]. However, existing studies frequently rely on established bench-mark datasets, potentially limiting the applicability of their findings to populations with different demographic and clinical characteristics [21]. Recent reviews further emphasize the importance of population diversity, transparent evaluation, and clinically meaningful interpretation in diabetes-related artificial intelligence [22,23].

Diabetes datasets and region-specific modeling

The reliability and generalizability of diabetes classification models depend partly on the representativeness of the underlying data. Although the PIMA Indian Diabetes dataset is widely used, its restriction to female participants from a specific population limits demographic coverage [13]. Other datasets, including those examined by Tigga and Garg [24] and Zou et al. [25] have supported diabetes-related prediction research but remain specific to their respective populations and clinical settings.

To support region-specific research, Prama et al. [4] introduced DiaBD, a Bangladesh-focused dataset containing 5,288 records from individuals aged 21 to 80 years across 63 unions representing urban, semi-urban, and rural communities. The dataset includes demographic, clinical, anthropometric, and medical-history variables, together with recorded diabetic status. Its original publication emphasized dataset development, characterization, validation, and preliminary machine learning analysis [4]. DiaBD therefore provides an opportunity to examine more comprehensive learning frameworks in a Bangladesh specific context. Because the recorded outcome represents existing diabetic status, the present study focuses on prevalent diabetes detection rather than prediction of future disease onset.

Class imbalance in medical prediction

Class imbalance is a persistent challenge in medical classification because disease-positive observations may be substantially less frequent than disease-negative observations [26]. DiaBD contains 342 diabetic and 4,946 non-diabetic records, corresponding to a diabetic prevalence of approximately 6.47% [4]. Under such conditions, high overall accuracy can conceal poor minority-class detection. Sensitivity, precision, F1-score, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and precision-recall analysis provide complementary measures of classification performance [27,28].

Resampling techniques are commonly used to reduce majority-class bias. The Synthetic Minority Over-Sampling Technique (SMOTE) generates synthetic minority-class observations [29], whereas Borderline-SMOTE concentrates on minority observations near the classification boundary [30]. Previous diabetes-related research has combined resampling with ensemble learning to address imbalanced clinical data [31]. Nevertheless, resampling must be restricted to training partitions because applying it before cross validation or train test separation can introduce information leakage and produce overly optimistic performance estimates. Accordingly, imbalance-aware modeling should be accompanied by fold-specific preprocessing, appropriate validation, and evaluation under the original class distribution.

Hybrid ensemble and deep tabular learning

Hybrid ensemble learning combines complementary classifiers to exploit differences in their predictive behavior and potentially improve generalization [32,33]. Random Forest aggregates multiple decision trees to reduce variance [34], whereas Gradient Boosting constructs additive models by sequentially improving earlier predictions [35]. XGBoost extends gradient boosting through regularization and efficient optimization [20]. Multilayer perceptron's and other neural architectures provide alternative approaches for learning nonlinear relationships among clinical variables.

Previous diabetes studies have explored hybrid models, weighted ensemble learning, and combinations of preprocessing and feature selection [31,36,37]. Optimization-guided methods have also been investigated in biomedical classification [38]. Weighted probability-level fusion allows individual classifiers to contribute to a combined prediction according to selected model weights. However, its added value must be assessed against well-tuned individual classifiers and simpler fusion strategies rather than assumed from model complexity alone. Deep learning has also been investigated for diabetes-related classification using structured clinical measurements [39,40]. Attention-based tabular architectures, including the Feature Tokenizer Transformer (FT-Transformer), offer an alternative representation-learning approach for modeling feature relationships. Nevertheless, relatively small and imbalanced clinical datasets may limit the advantages of more complex neural architectures. Consequently, a comparison with an independently trained FT-Transformer provides a useful assessment of the proposed ensemble relative to attention-based tabular learning.

Explainable artificial intelligence in healthcare

Although machine learning models can identify complex relationships among clinical variables, limited interpretability can hinder their use in healthcare decision support. Explainable Artificial Intelligence (XAI) methods seek to characterize the factors influencing model predictions. Local Interpretable Model-Agnostic Explanations (LIME) approximates the behavior of a predictive model near an individual observation using an interpretable surrogate [7]. Shapley Additive Explanations (SHAP) attributes predictions to input features and supports both local explanations and global summaries of feature contributions [8].

Recent diabetes studies have employed these methods to examine influential variables and improve prediction transparency. Tasin et al. [9] investigated LIME and SHAP explanations for diabetes classification, while Hasan et al. [10] combined automated machine learning with explainability analysis. Other work has examined the integration of imbalance-aware learning and interpretability in diabetes-related applications [5]. Recent research also reflects growing interest in adaptive and interpretable diabetes decision-support approaches [41].

Nevertheless, feature-attribution methods explain model

behavior rather than establish causal relationships. Their interpretation may be affected by correlated clinical predictors and engineered interaction features. Accordingly, global attribution summaries should be complemented by representative local explanations and assessments of explanation stability. Furthermore, interpretability should be considered alongside probability calibration, classification thresholds, and the clinical consequences of false-positive and false-negative predictions.

Comparison with prior studies and research gap

Table 1 summarizes representative diabetes classification studies and positions the proposed framework relative to existing approaches. Because these studies differ in population characteristics, datasets, outcome definitions, pre-processing, and validation protocols, their reported performance values should not be interpreted as directly comparable.

Prior studies have established the usefulness of conventional classifiers, ensemble learning, resampling, deep learning, and XAI for diabetes-related classification. However, their evaluation settings vary considerably, and the relative contributions of these approaches remain insufficiently examined for DiaBD. In particular, the effects of optimized ensemble fusion, individual engineered features, glucose-related predictors, probability calibration, and clinically relevant operating thresholds require systematic investigation under leakage-aware validation.

The present study addresses these questions through a five-classifier hybrid ensemble evaluated on an untouched test set preserving the original class distribution. The framework is compared with tuned individual classifiers and an independently trained FT-Transformer baseline [43]. Feature ablation and glucose exclusion experiments assess predictor contributions, while calibration, bootstrap confidence intervals, decision-curve analysis, and SHAP/LIME explanations provide complementary assessments of predictive reliability and interpretability. Rather than assuming that a more complex ensemble necessarily outperforms simpler alternatives, this study examines its performance and limitations within a transparent evaluation framework.

Proposed Methodology

Study design and framework overview

This study conducted a secondary analysis of the previously collected DiaBD dataset to develop and evaluate an explainable and imbalance aware hybrid ensemble framework for prevalent diabetes detection. The framework integrates clinically motivated feature engineering, fold-specific preprocessing, model-dependent imbalance handling, nested cross validation, and optimized probability level fusion. An FT-Transformer was independently evaluated as a deep tabular learning baseline. The overall workflow is illustrated in Figure 1.

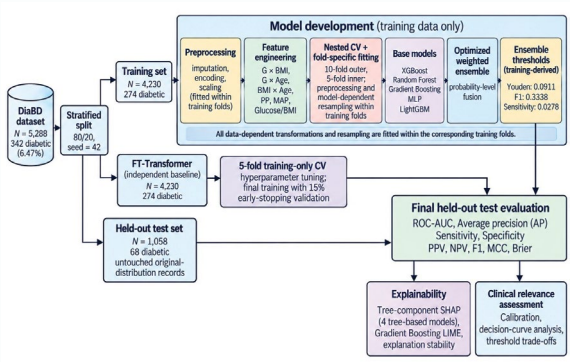


Figure 1: Overview of the proposed leakage-aware, imbalance-aware, and explainable hybrid ensemble framework for prevalent diabetes detection using the DiaBD dataset.

To prevent information leakage, the original dataset was partitioned into training and held-out test sets before any data-dependent preprocessing, resampling, or model selection. All fitted transformations and model optimization were restricted to the training partition. Final discrimination, classification, and calibration performance were evaluated on the untouched test set, which retained the original class distribution. Additional experiments examined individual feature contributions, dependence on glucose-related predictors, threshold-dependent performance, potential clinical net benefit, and model explainability.

Table 1: Comparison of the proposed framework with representative prior diabetes classification studies. Reported performance values are not directly comparable across different datasets and validation protocols.

Study	Dataset	Method	Performance	Scope and Limitations
Prama et al. [4]	DiaBD, Bangladesh	Dataset development, validation, and preliminary machine learning analysis	Preliminary classification results	Introduced DiaBD; the present study extends its analysis through leakage-aware validation, ensemble optimization, feature ablation, calibration, and explainability.
Alghamdi et al. [31]	Henry Ford FIT dataset	SMOTE-based ensemble learning	AUC=0.92	Investigated imbalance-aware ensemble classification using a different clinical dataset.
Dutta et al. [36]	Bangladesh diabetes datasets	Weighted ensemble with preprocessing and feature selection	Accuracy=0.735; AUC=0.832	Examined weighted ensemble learning without the present DiaBD-specific validation and sensitivity analyses.
Ganie et al. [42]	PIMA Indian Diabetes dataset	Boosting-based ensemble classification	Accuracy=92.85%	Focused on conventional benchmark data rather than the DiaBD population.
Hasan et al. [10]	PIMA Indian Diabetes dataset	Automated machine learning with explainable AI	Accuracy=85.01%	Emphasized automated modeling and interpretability in a different dataset.
Proposed study	DiaBD, Bangladesh	Five-model hybrid ensemble; FT-Transformer baseline; fold specific resampling; feature ablation; calibration; SHAP and LIME	ROC-AUC=0.8722; AP=0.3561	Original distribution held-out evaluation with bootstrap uncertainty, glucose-exclusion analysis, calibration, and operating-point assessment.

Dataset and experimental partitioning

The DiaBD dataset contains 5,288 observations collected in Bangladesh, including demographic, anthropometric, clinical, and medical history attributes [4]. Available predictors include age, gender, glucose, Body Mass Index (BMI), systolic and diastolic blood pressure, pulse rate, height, weight, hypertension, family history of diabetes and hypertension, cardiovascular disease, and stroke. The provided binary outcome indicates recorded diabetic status, with diabetic and non-diabetic cases encoded as 1 and 0, respectively. Accordingly, the present analysis addresses prevalent diabetes classification rather than future disease-onset prediction (Figure 2).

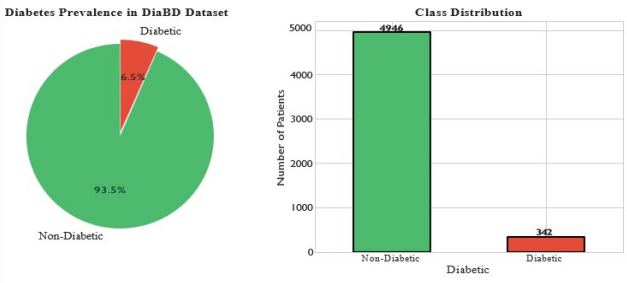


Figure 2: Original class distribution of DiaBD: 4,946 non-diabetic (93.53%) and 342 diabetic (6.47%) observations.

Of the 5,288 observations, 342 (6.47%) were diabetic and 4,946 (93.53%) were non-diabetic. A stratified 80:20 train-test split was performed using random seed 42. The training partition contained 4,230 observations (274 diabetic and 3,956 non-diabetic), whereas the held-out test partition comprised 1,058 observations (68 diabetic and 990 non-diabetic). The test partition was excluded from preprocessing-parameter estimation, synthetic resampling, hyper-parameter tuning, ensemble-weight selection, probability calibration fitting, and operating-threshold selection.

According to Prama et al. [4], DiaBD includes Bangladeshi participants aged 21 to 80 years recruited across 63 unions. Data were collected by trained community health workers, and diabetic-status classifications underwent physician review. However, the published documentation does not clearly specify the diagnostic thresholds or whether the recorded glucose measurements contributed to label assignment. The present study used the published labels without independent clinical reassessment.

Data preprocessing and feature engineering

Preprocessing addressed invalid measurements, missing values, categorical attributes, and variation in numerical feature scales. Zero values in glucose, BMI, systolic blood pressure, diastolic blood pressure, and pulse rate were treated as missing. Categorical attributes were numerically encoded using categories learned from the corresponding training fold, with unseen categories assigned a separate code. Missing numerical values were replaced using training-fold medians.

To reduce the influence of extreme observations, numerical predictors were winsorized using the 0.5^{th} and 99.5^{th} percentiles estimated from the corresponding training fold. The resulting features were transformed using RobustScaler. All fitted preprocessing operations were included in the model pipelines so that validation and test observations were transformed using parameters estimated exclusively from the relevant training data.

Six clinically motivated features were constructed to represent interactions among glucose, BMI, age, and blood pressure. Let G, B, A, S, and D denote glucose, BMI, age, systolic blood pressure, and diastolic blood pressure, respectively. The engineered features were defined as

$$FGB = G \times B, \tag{1}$$

$$F_{G/B} = \frac{G}{B+\in}, \tag{2}$$

$$FAB = A \times B, \tag{3}$$

$$FAG = A \times G, \tag{4}$$

$$FPP = S - D, \tag{5}$$

$$F_{MAP} = D + \frac{S-D}{3}, \tag{6}$$

Where $\varepsilon = 10^{-6}$ was introduced for numerical stability. The last two features represent pulse pressure and mean arterial pressure, respectively.

A logarithmic transformation, $\log(1 + \max(x, 0))$, was applied to glucose, BMI, glucose-BMI interaction, glucose/BMI ratio, age-BMI interaction, and age-glucose interaction. These deterministic transformations were performed without estimating parameters from the validation or test observations. The subsequent imputation, winsorization, encoding, and scaling procedures remained fold-specific. No composite diabetes risk score was included in the final experiments.

Imbalance handling and nested cross validation

Because diabetic observations represented only 6.47% of the dataset, model development incorporated class weighting and model-dependent synthetic re-sampling. The Synthetic Minority Over-Sampling Technique (SMOTE) generates synthetic minority-class observations through interpolation [29], while Borderline-SMOTE concentrates on minority observations near the decision boundary [30].

The final framework employed Borderline-SMOTE for Random Forest and Adaptive Synthetic Sampling (ADASYN) for the Multilayer Perceptron (MLP). Both methods were configured with five nearest neighbors and random seed 42. XGBoost, Gradient Boosting, and LightGBM were trained without synthetic resampling. Class-imbalance weighting was additionally used where supported by the classifier: Random Forest and LightGBM employed balanced class weights, while XGBoost used a training-fold-specific positive-class weight based on the negative-to-positive class ratio. These strategies were implemented within model-specific training pipelines.

Model development employed nested stratified cross-validation within the original training partition. The outer loop contained 10 folds, while the inner hyperparameter-search loop contained five folds. For each outer iteration, hyperparameter tuning was performed using RandomizedSearchCV with eight

sampled configurations and ROC-AUC as the selection criterion. The selected model was fitted using the outer-training fold and subsequently evaluated on its previously unseen outer-validation fold.

All data-dependent transformations and synthetic resampling operations were fitted only within the corresponding training folds. Validation observations were never used to estimate imputation statistics, feature scaling parameters, winsorization limits, or synthetic minority samples. This procedure generated Out-of-Fold (OOF) predicted probabilities for subsequent model comparison and ensemble optimization.

Following nested model development, each classifier underwent a separate five-fold hyperparameter search on the complete training partition. The selected pipeline was then refitted using the training data and applied to the untouched held-out test partition. The same partition was used for evaluating all individual classifiers and the final ensemble.

Hybrid ensemble learning and weight optimization

The proposed framework combines five classifiers: XGBoost, Random Forest, Gradient Boosting, MLP, and LightGBM. These models were selected to incorporate complementary tree-based bagging, boosting, and neural-network learning mechanisms [44,45]. Each classifier generated a diabetic-class probability, which was combined through weighted probability-level fusion:

$$\hat{p}_{ens}(x) = \sum_{m=1}^{5} w_m \hat{p}_m(x), \tag{7}$$

where $p\hat{}_m(x)$ is the predicted probability generated by classifier m, and w_m denotes its ensemble weight.

Four weighting strategies were compared: Equal weighting, ranking-based weighting, performance-based weighting, and optimized weighting. For the optimized candidate, weights were estimated from training-derived OOF probabilities by minimizing a penalized log-loss objective:

$$L(w) = -\frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \left[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i) \right] + \lambda \sum_{m=1}^{5} w_m^2, \tag{8}$$

where

$$\hat{p}_i = \sum_{m=1}^{5} w_m \hat{p}_{im}, \quad \lambda = 0.02. \tag{9}$$

The optimized weights were estimated using Sequential Least-Squares Programming (SLSQP) to minimize out-of-fold log loss with an L_2 penalty ($\lambda = 0.02$), subject to $0 \le w_m \le 0.40$ and $\sum_{m=1}^{5} w_m = 1$. The candidate weighting strategies were compared using training-derived OOF ROC-AUC, and the best-performing strategy was selected.

The final selected weights were 0.2415, 0.2045, 0.4000, 0.0574, and 0.0966 for XGBoost, Random Forest, Gradient Boosting, MLP, and LightGBM, respectively. The weighted probabilities were used for final held-out evaluation.

Because the same OOF predictions were used for weight estimation and fusion-strategy selection, their ensemble performance was considered a model-selection estimate rather than an independent estimate of generalization. The untouched held-out test set provided the final performance assessment.

For a selected probability threshold τ, the predicted class was determined by

$$\hat{y}(x;\tau) = \begin{cases} 1, \hat{p}_{ens}(x) \ge \tau \\ 0, \text{ otherwise} \end{cases}, \tag{10}$$

The complete leakage-aware model development and evaluation procedure is summarized in Algorithm 1.

Algorithm 1: Leakage-aware and explainable hybrid ensemble framework for prevalent diabetes detection.

Require: DiaBD dataset $D = (X, y)$; random seed $s = 42$

Ensure: Held-out predictions, performance metrics, and model explanations

1: Stratify D into training D_{tr} (80%) and held-out test D_{te} (20%)

2: for each classifier M_m in {XGBoost, RF, GB, MLP, LightGBM} **do**

3: Construct six clinically motivated features and fit preprocessing within the corresponding training folds

4: Apply model-specific imbalance handling within training folds: Borderline-SMOTE (RF), ADASYN (MLP), or class weighting where applicable

5: Perform nested stratified CV (10 outer, 5 inner folds), tuning hyperparameters using ROC-AUC

6: Generate training out-of-fold probabilities p_m^{OOF}

7: Perform final five-fold hyperparameter selection and refit M_m on D_{tr}

8: **end for**

9: Compare equal, rank-based, performance-based, and optimized fusion using training OOF predictions

10: Optimize candidate weights using constrained penalized log loss (SLSQP, $\lambda = 0.02$)

11: Select the fusion strategy using training OOF ROC-AUC

12: Determine classification thresholds using training OOF predictions only

13: Fit isotonic calibration using training OOF ensemble probabilities and labels only

14: Generate held-out probabilities p^{te} from the five refitted classifiers

15: Compute ensemble probabilities $\hat{p}_{ens} = \sum_{m=1}^{5} w_m p_m^{te}$

16: Apply the training-fitted isotonic calibrator to held-out ensemble probabilities

17: Independently tune and train the FT-Transformer baseline using training-only five-fold CV

18: Evaluate the ensemble and baseline on the untouched D_{te} using ROC-AUC, AP, and classification metrics

19: Assess held-out probability calibration using calibration curves and Brier score; evaluate bootstrap uncertainty and decision-curve net benefit

20: Generate tree-component SHAP and Gradient Boosting LIME explanations; assess LIME stability

21: **return** Predictions, performance metrics, and explanations

Model hyperparameter settings

The proposed framework employed five base classifiers, while the FT-Transformer was evaluated independently as a deep-learning baseline. Mode1 hyperparameters were selected using training-only cross-validation, with preprocessing and imbalance handling restricted to the corresponding training folds.

The five ensemble classifiers underwent nested stratified cross-validation (10 outer and 5 inner folds). The FT-Transformer was tuned separately using five-fold cross-validation over four predefined configurations, followed by training with a 15% validation subset for early stopping. All models were evaluated on the same untouched held-out test set.

Table 2 summarizes the principal model configurations. Complete hyperparameter search spaces and selected settings are provided in the reproducibility archive.

Performance evaluation and threshold selection

All models were evaluated using the original held-out test partition containing 68 diabetic and 990 non-diabetic observations. Performance metrics included accuracy, sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), F1-score, balanced accuracy, Matthews Correlation Coefficient (MCC), ROC-AUC, and Average Precision (AP).

The principal threshold-dependent measures were defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (12)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (13)$$

$$\text{PPV} = \frac{TP}{TP + FP} \quad (14)$$

$$\text{NPV} = \frac{TN}{TN + FN} \quad (15)$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (16)$$

Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

ROC-AUC summarizes discrimination across classification thresholds:

$$ROC - AUC = \int_0^1 TPR(u)\,du, \quad (17)$$

where u represents the false-positive rate. Precision-recall performance was summarized using AP:

$$AP = \sum_k (R_k - R_{k-1})\,P_k, \quad (18)$$

where P_k and R_k represent precision and recall at the kth operating point. AP was calculated using average precision score, rather than trapezoidal integration of the precision recall curve [27,28].

Ensemble operating thresholds were selected exclusively from training-derived OOF predictions. Three strategies were examined: Maximization of Youden's index, maximization of the F1-score, and selection of a threshold corresponding to a target training sensitivity of 90%. The resulting thresholds were 0.0911, 0.3338, and 0.0278, respectively.

The Youden threshold was treated as the primary statistical operating point. The sensitivity-oriented threshold was examined as an exploratory screening-focused alternative. These thresholds were not interpreted as clinically optimal decision rules, because patient-specific costs and prospective clinical outcomes were unavailable. Default-threshold results were also reported to support comparisons among classifiers.

Calibration, uncertainty, and clinical relevance

Probability calibration was examined to determine whether predicted probabilities were consistent with observed diabetes frequencies. Reliability diagrams and Brier scores were used to evaluate uncalibrated predictions and isotonic-regression calibration.

The Brier score was defined as

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2, \quad (19)$$

Table 2: Overview of model architectures, training strategies, and selected ensemble and FT-Transformer settings.

Model	Configuration
XGBoost	Tree-based gradient boosting; hyperparameters selected by training-only cross-validation.
Random Forest	Bagging-based tree ensemble; Borderline-SMOTE applied within training folds.
Gradient Boosting	Sequential gradient-boosted trees; hyperparameters selected by training-only cross-validation.
MLP	Feedforward neural network; ADASYN applied within training folds.
LightGBM	Histogram-based gradient boosting; class-imbalance weighting used during training.
FT-Transformer	Embedding dimension: 64; Transformer blocks: 2; attention heads: 4; dropout: 0.10; learning rate: 0.001; AdamW optimizer; maximum epochs: 100; early stopping: 10.
Hybrid ensemble	Five-model probability fusion; weights: (0.2415, 0.2045, 0.4000, 0.0574, 0.0966); SLSQP optimization; λ = 0.02.

where $\hat{p}_i$ and y_i denote the predicted probability and observed binary out-come for observation i. Isotonic regression was fitted using training derived OOF predictions and subsequently applied to held-out test probabilities. Raw and calibrated probabilities were evaluated separately to characterize changes in calibration and discrimination.

Uncertainty was quantified using 10,000 bootstrap replicates of the held-out test observations. Bootstrap confidence intervals were calculated for ROC-AUC, AP, and selected operating-point metrics. Paired bootstrap comparisons were used to estimate differences in performance between the proposed ensemble and individual classifiers, particularly XGBoost. These comparisons were used to assess whether observed performance differences were compatible with sampling variability.

Potential clinical relevance was explored using Decision Curve Analysis (DCA). For a threshold probability p_t, net benefit was calculated as

$$NB(pt) = \frac{TP}{N} - \frac{FP}{N} \times \frac{pt}{1-pt} \qquad (20)$$

where N denotes the number of evaluated observations. The model was compared with treat-all and treat-none reference strategies across threshold probabilities. These analyses characterize potential decision trade-offs but do not constitute prospective evidence of clinical effectiveness.

Feature ablation and glucose-exclusion analysis

Individual feature-ablation experiments were performed to evaluate the contribution of the six engineered predictors. In each experiment, one engineered feature was removed, the five base classifiers were retrained using training-only hyperparameter selection, and the previously selected ensemble weights were held fixed. The resulting change in held-out ROC-AUC was calculated relative to the complete feature set. These experiments were intended to evaluate feature sensitivity without reoptimizing the ensemble weights for each ablated model.

A separate glucose-exclusion experiment removed the original glucose measurement and all three glucose-derived features: Glucose-BMI interaction, glucose/BMI ratio, and glucose-age interaction. The remaining predictors were used to retrain the models under the same training-only evaluation approach.

The resulting performance difference quantified the model's dependence on glucose-related information.

This sensitivity analysis does not establish whether glucose was used to define diabetic status in the original dataset. Any overlap between the diagnostic-label criteria and the input measurements remains an important consideration for interpreting performance and generalizability.

Explainability analysis

Model interpretability was investigated using Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). These methods were used to examine global feature contributions, representative individual predictions, and explanation stability.

SHAP analysis was performed separately for XGBoost, Random Forest, Gradient Boosting, and LightGBM. Model-specific TreeExplainer implementations were applied to the fitted tree classifiers using transformed training observations as background references. Background samples of up to 100 training observations were used for the model-specific analysis. Feature contributions were evaluated in probability space were supported and checked for consistency with model predictions.

A weighted tree-component summary was additionally constructed by normalizing the selected ensemble weights over the four tree-based models. This representation accounted for approximately 94.26% of the total five-model ensemble weight. Because MLP was excluded from this aggregation, the resulting explanations were interpreted as a tree-component summary rather than exact SHAP explanations of the complete hybrid ensemble.

LIME was applied to the Gradient Boosting classifier, which received the largest individual ensemble weight. Representative true-positive, false-positive, false-negative, and true-negative predictions were selected for local explanation using the fitted feature representation. The local surrogate used 10 explanatory features and 5,000 perturbation samples for illustrative cases.

Explanation stability was examined for 25 test observations using 10 random seeds per observation. For each run, the top 10 explanatory features were retained, and pairwise Jaccard similarity was calculated between the resulting feature sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \qquad (21)$$

where A and B represent the selected explanation feature sets. Repeated explanations used 1,000 perturbation samples per run. Mean stability was re-ported to characterize the consistency of local explanations. Neither SHAP nor LIME attributions were interpreted as evidence of causal clinical relationships.

Reproducibility and reporting

The experiments were implemented in Python using scikit-learn, imbalanced-learn, XGBoost, LightGBM, PyTorch, SHAP, and LIME. Reproducibility was supported through fixed random seeds, stratified data partitioning, and fold-specific preprocessing. The five ensemble classifiers were evaluated using nested cross-validation with 10 outer and five inner folds, with eight randomized hyper-parameter configurations per search. The FT-Transformer was independently optimized using five-fold cross-validation on the training data.

All preprocessing, resampling, hyperparameter optimization, ensemble-weight selection, calibration fitting, and threshold selection were restricted to the training partition. The accompanying reproducibility archive contains the analysis code, hyperparameter search spaces, selected model configurations, software dependencies, trained models, and experimental results, supporting transparent evaluation and independent replication of the proposed framework.

The manuscript was reviewed against the TRIPOD+AI reporting guideline [46], and the completed checklist is submitted

as an accompanying reporting document.

Results

Held-out performance and model comparison

The proposed framework was evaluated on the untouched held-out test set containing 1,058 observations, including 68 diabetic and 990 non-diabetic cases. The test set retained the original class distribution, with a diabetes prevalence of 6.43%, and was not subjected to synthetic resampling. Table 3 summarizes the performance of the five individual classifiers, optimized hybrid ensemble, and independent FT-Transformer baseline.

The optimized ensemble achieved a ROC-AUC of 0.8722 and an AP of 0.3561. Its discrimination was closely comparable to XGBoost (ROC-AUC = 0.8714), whereas Gradient Boosting achieved the highest AP (0.3896). At the default classification threshold, LightGBM obtained the highest F1-score (0.4503) and Matthews Correlation Coefficient (MCC = 0.4109). These findings indicate that no model consistently achieved the best performance across all metrics.

To assess the contribution of ensemble-weight optimization, four probability-level fusion strategies were compared (Table 4). Optimized weighting achieved the highest training OOF ROC-AUC (0.8798) and was therefore selected before held-out evaluation. On the independent test set, rank-based weighting achieved the highest ROC-AUC (0.8733), whereas optimized weighting achieved the highest AP (0.3561). These small differences indicate that weight optimization did not consistently outperform simpler fusion strategies.

Figure 3 presents the Receiver Operating Characteristic (ROC) and precision-recall curves. The ensemble and XGBoost showed similar ROC discrimination, while Gradient Boosting exhibited stronger precision-recall performance. The precision-recall reference level corresponded to the test-set diabetes prevalence of approximately 0.064.

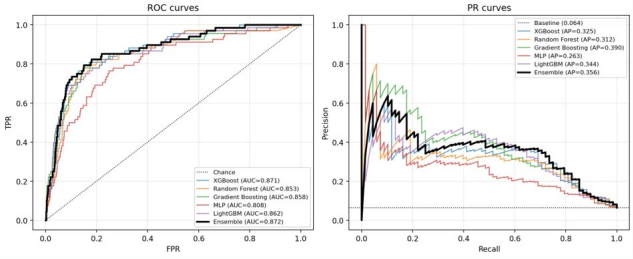


Figure 3: ROC and precision recall curves of the individual classifiers and optimized hybrid ensemble on the untouched held-out test set.

Bootstrap analysis with 10,000 replicates yielded a 95% Confidence Interval (CI) of 0.8246-0.9144 for ensemble ROC-AUC and 0.2602-0.4849 for AP. Relative to XGBoost, the ensemble achieved a ROC-AUC difference of 0.0009 (95% CI: −0.0132 to 0.0155) and an AP difference of 0.0315 (95% CI: −0.0042 to 0.0717). Both confidence intervals included zero, indicating that the observed differences did not establish statistically significant superiority over XGBoost.

The independent FT-Transformer achieved a held-out ROC-AUC of 0.8415 and AP of 0.3231. At the default threshold of 0.5, its sensitivity, specificity, F1-score, and MCC were 0.6324, 0.8838, 0.3805, and 0.3552, respectively. Although the FT-Transformer showed lower overall discrimination than the optimized ensemble, it identified more diabetic cases at the default threshold. This difference underscores the importance of evaluating classification thresholds alongside discrimination metrics.

Table 3: Held-out test performance of individual classifiers, the optimized hybrid ensemble, and the independent FT-Transformer baseline. Threshold-dependent metrics were calculated at the default threshold of 0.5. AP denotes average precision.

Model	ROC-AUC	AP	Sensitivity	Specificity	F1-score	MCC
XGBoost	0.8714	0.325	0.2647	0.9707	0.313	0.28
Random Forest	0.8535	0.312	0.3088	0.9606	0.3281	0.286
Gradient Boosting	0.8583	0.39	0.2353	0.9859	0.3265	0.327
MLP	0.808	0.263	0.2206	0.9586	0.2419	0.196
LightGBM	0.8622	0.344	0.5	0.9505	0.4503	0.411
Optimized ensemble	0.8722	0.356	0.2206	0.9778	0.2857	0.265
FT-Transformer	0.8415	0.323	0.6324	0.8838	0.3805	0.355

Table 4: Comparison of ensemble fusion strategies. OOF ROC-AUC was computed on training-derived out-of-fold predictions, whereas test metrics were evaluated on the untouched held-out set. AP denotes average precision.

Fusion Strategy	OOF AUC	Test AUC	Test AP
Equal weighting	0.8748	0.8714	0.3409
Rank-based weighting	0.8792	0.8733	0.3475
Performance-based weighting	0.8783	0.8728	0.3444
Optimized weighting	0.8798	0.8722	0.3561

Operating thresholds and classification errors

To examine threshold-dependent classification behavior, four operating points were considered: The default threshold of 0.5 and three thresholds selected exclusively from training Out-of-Fold (OOF) predictions. The training-derived thresholds were 0.0911 for Youden's index, 0.3338 for F1-score optimization, and 0.0278 for a sensitivity-oriented strategy targeting 90% sensitivity in training.

Table 5 reports the resulting held-out confusion-matrix counts and classification metrics. The F1-derived threshold is explicitly identified as training-selected rather than test-optimal.

At the default threshold, the ensemble correctly identified only 15 of the 68 diabetic cases, corresponding to 22.06% sensitivity and 97.78% specificity. The training-selected Youden threshold substantially increased sensitivity to 79.41%, identifying 54 diabetic cases, while specificity declined to 84.85%. This operating point produced 150 false positives and 14 false negatives.

The training-selected F1-optimal threshold achieved a held-out F1-score of 0.4029, slightly higher than the 0.3971 obtained at the Youden threshold. However, its sensitivity was lower (41.18%), indicating a greater number of missed diabetic cases.

At the sensitivity-oriented threshold, 58 of 68 diabetic cases were identified, yielding 85.29% sensitivity. This improvement

was accompanied by 314 false positives, 68.28% specificity, and 15.59% positive predictive value. Although the threshold was selected to target 90% sensitivity during training, the held-out sensitivity did not reach that target.

Figure 4 presents the corresponding confusion matrices.

At the Youden threshold, the negative predictive value was approximately 0.9836. Bootstrap analysis yielded 95% CIs of 0.6935-0.8857 for sensitivity and 0.8258-0.8705 for specificity. These intervals highlight uncertainty in minority-class detection, particularly because the test set contained only 68 diabetic observations.

Overall, threshold selection substantially altered the trade-off between sensitivity, specificity, and positive predictive value. The Youden threshold provided a statistical balance between sensitivity and specificity, whereas the sensitivity-oriented threshold reduced missed diabetic cases at the cost of considerably more false-positive predictions.

Probability calibration and clinical relevance

Calibration was assessed using reliability diagrams, Brier scores, and a ten-bin calibration-error summary. Table 6 reports the raw and isotonic-calibrated results for the ensemble and XGBoost.

The raw ensemble achieved a Brier score of 0.0503 (95% CI: 0.0407-0.0604), which decreased to 0.0493 (95% CI: 0.0395-0.0597) after isotonic calibration. For XGBoost, the corresponding Brier scores decreased from 0.0575 (95% CI: 0.0463-0.0693) to 0.0504 (95% CI: 0.0405-0.0610).

The mean absolute deviation between predicted and observed frequencies across ten quantile-based calibration bins decreased from 0.0165 to 0.0157 for the ensemble and from 0.0345 to 0.0226 for XGBoost. These values provide a descriptive bin-based calibration measure and are not interpreted as a smoothing-based Integrated Calibration Index.

The ensemble exhibited lower Brier scores and bin-based calibration deviations than XGBoost before and after isotonic calibration. Nevertheless, the absolute improvement in ensemble Brier score following isotonic calibration was small (0.0010), and no paired-bootstrap confidence interval was calculated for this difference. Therefore, statistical superiority in calibration was not established.

Figure 5 illustrates the raw and isotonic-calibrated reliability curves across the full probability range and the low-probability region. Both models exhibited deviations from ideal calibration, with variation particularly evident among low predicted probabilities. Isotonic calibration modified the probability distributions but did not eliminate all calibration discrepancies.

Decision Curve Analysis (DCA) was performed to explore potential clinical net benefit across threshold probabilities (Figure 6).

The raw ensemble exhibited positive net benefit across a range of lower threshold probabilities and generally outperformed the treat-all and treat-none strategies within much of this region. Net benefit declined as the decision threshold increased and approached zero near a threshold probability of 0.38. Differences among the ensemble, its calibrated variant, and XGBoost were generally modest across the lower threshold range.

These findings suggest that potential decision benefit depends on the selected probability threshold and the relative

Table 5: Held-out performance of the optimized ensemble at default and training-OOF-selected operating thresholds.

Operating threshold	TN	FP	FN	TP	Sensitivity	Specificity	PPV	F1
Default (0.5000)	968	22	53	15	0.2206	0.9778	0.405	0.286
Youden (training-OOF; 0.0911)	840	150	14	54	0.7941	0.8485	0.265	0.397
F1-optimal (training-OOF; 0.3338)	947	43	40	28	0.4118	0.9566	0.394	0.403
Sensitivity-targeted (training-OOF; 0.0278)	676	314	10	58	0.8529	0.6828	0.156	0.264

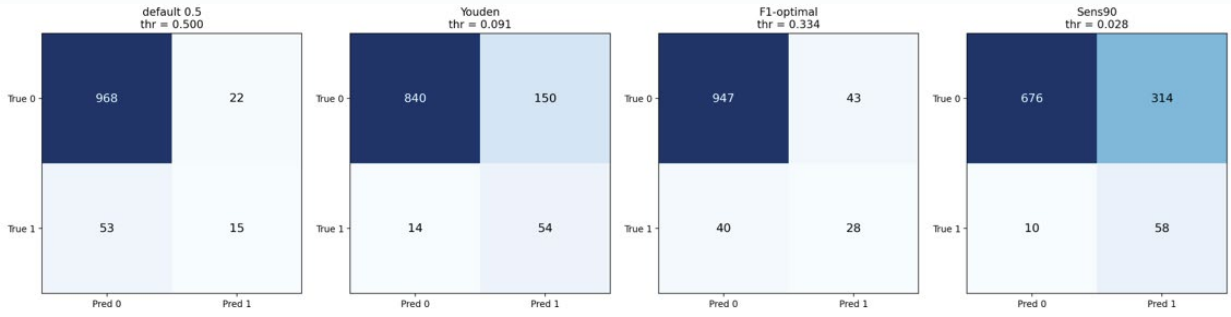


Figure 4: Held-out confusion matrices of the optimized ensemble at the default, training-selected Youden, training-selected F1-optimal, and sensitivity-oriented thresholds.

Table 6: Calibration performance on the untouched held-out test set. Confidence intervals were obtained using 10,000 bootstrap replicates. The last column reports the mean absolute calibration deviation across ten quantile bins.

Model	Brier	95% CI	Bin error
Ensemble (raw)	0.05	0.0407–0.0604	0.0165
Ensemble (isotonic)	0.049	0.0395–0.0597	0.0157
XGBoost (raw)	0.058	0.0463–0.0693	0.0345
XGBoost (isotonic)	0.05	0.0405–0.0610	0.0226

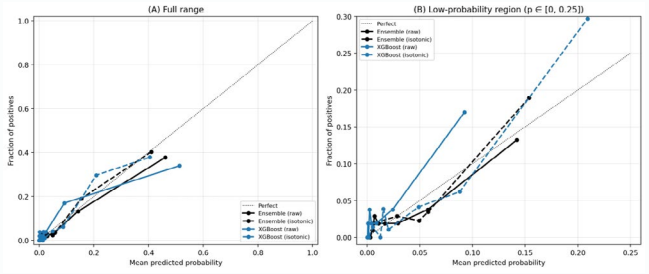


Figure 5: Reliability diagrams for the optimized ensemble and XGBoost before and after isotonic calibration: (A) full probability range and (B) low-probability region. The diagonal represents perfect calibration.

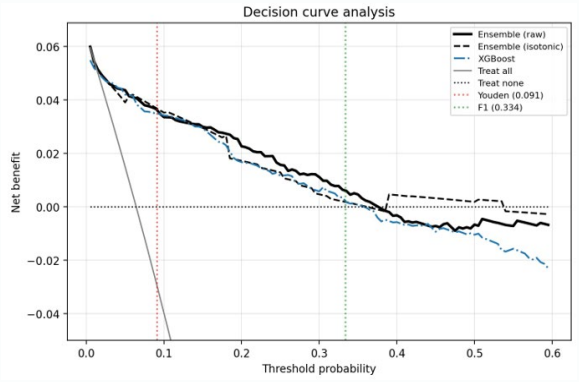


Figure 6: Decision-curve analysis comparing the raw ensemble, isotonic-calibrated ensemble, XGBoost, and treat-all and treat-none reference strategies.

Six individual feature-ablation experiments were performed to investigate the contribution of clinically motivated engineered predictors. Each experiment removed one engineered feature, retrained the classifiers, and evaluated the ensemble using the previously selected weights. The complete ensemble achieved a held-out ROC-AUC of 0.8722.

Figure 7 presents the change in ROC-AUC following each feature removal. Removing the age-glucose interaction produced the largest reduction in discrimination ($\Delta AUC = -0.0059$), followed by mean arterial pressure (−0.0026) and glucose/BMI ratio (−0.0015). Removing age-BMI interaction resulted in a negligible decrease (−0.0003).

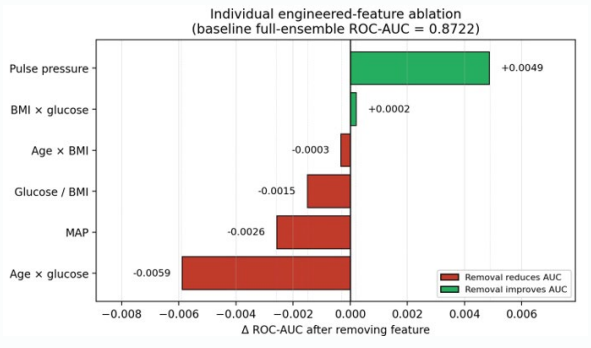


Figure 7: Individual engineered-feature ablation results expressed as changes in held-out ROC-AUC relative to the complete ensemble (ROC-AUC = 0.8722).

Conversely, excluding pulse pressure increased ROC-AUC by 0.0049, while removing the BMI-glucose interaction increased ROC-AUC by only 0.0002. These findings indicate that the engineered features contributed unequally to model discrimination and that adding all derived variables did not uniformly improve performance.

A separate glucose-exclusion experiment removed the original glucose variable and all three glucose-derived features: Glucose-BMI interaction, glucose/BMI ratio, and glucose-age interaction. The resulting ROC-AUC declined from 0.8722 to 0.7801, a reduction of 0.0921.

This larger reduction indicates substantial dependence on glucose-related predictors. However, it does not determine whether glucose measurements contributed to the original definition of diabetic status. The observed dependence must therefore be interpreted in the context of the dataset's diagnostic-label provenance.

SHAP-based global and local interpretation

Shapley Additive Explanations (SHAP) were calculated for the four tree-based ensemble members: XGBoost, Random Forest, Gradient Boosting, and LightGBM. Their attributions were combined using normalized weights to produce a weighted tree-component summary. The four tree-based models accounted for approximately 94.26% of the selected ensemble weight.

Figure 8 presents the weighted tree-component SHAP summary. Hypertension status and glucose were among the most influential predictors, followed by glucose-age interaction, glucose-BMI interaction, glucose/BMI ratio, and age-BMI interaction. Higher values of several glucose-related features generally contributed toward increased predicted diabetes probabilities, although the magnitude and direction of attribution varied across observations. The individual tree-model analyses showed broadly similar patterns, with differences in feature rankings and attribution magnitudes. The prominence of hypertension and glucose-related variables indicates that these predictors influenced the fitted models substantially.

Because SHAP TreeExplainer was applied to the four tree-based classifiers, the MLP component was excluded from the weighted attribution summary. The resulting explanations represent approximately 94.26% of the ensemble weight rather than the complete five-model ensemble. These attributions describe model behavior, not causal clinical relationships, particularly in the presence of correlated and engineered predictors.

LIME-based local interpretation and stability

Local Interpretable Model-Agnostic Explanations (LIME) were used to examine representative predictions from the Gradient Boosting classifier. Four cases were selected to illustrate true positive, false-positive, true-negative, and false-negative outcomes for which the Gradient Boosting classifier and optimized ensemble agreed at the default threshold of 0.5.

Figure 9 presents the corresponding local explanations.

In the true positive case, the ensemble and Gradient Boosting assigned diabetic probabilities of 0.892 and 0.797, respectively. Hypertension status, age-glucose interaction, and glucose contributed positively toward the Gradient Boosting prediction.

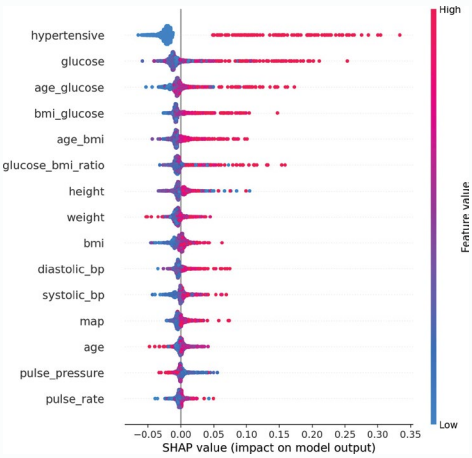


Figure 8: Weighted SHAP summary of the four tree-based ensemble components. Positive values indicate contributions toward higher predicted diabetes probabilities.

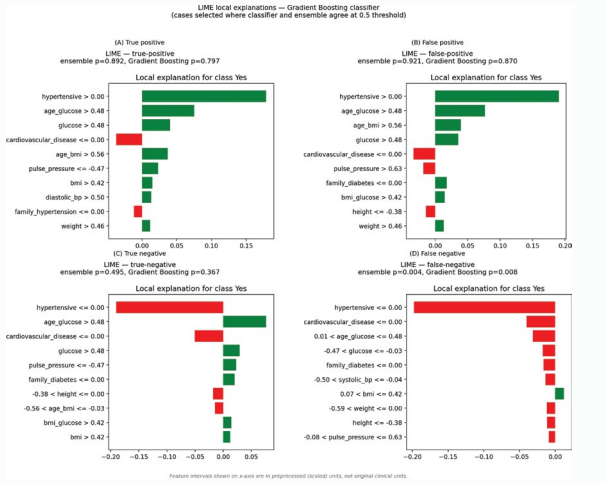


Figure 9: LIME explanations of representative Gradient Boosting predictions: (A) true positive, (B) false positive, (C) true negative, and (D) false negative. Feature intervals are expressed in preprocessed rather than original clinical measurement units.

In the false positive case, the corresponding probabilities were 0.921 and 0.870. Hypertension and glucose-related features contributed toward a diabetic classification despite the observed non-diabetic status.

For the true negative case, the ensemble and Gradient Boosting probabilities were 0.495 and 0.367. The absence of hypertension contributed against a diabetic classification, although some features supported the positive class.

For the false-negative case, the ensemble and Gradient Boosting probabilities were 0.004 and 0.008. Several local feature conditions reduced the predicted probability despite the observed diabetic status.

Explanation stability was assessed using 25 test observations and 10 random seeds per observation. The mean Jaccard similarity among the top 10 explanatory feature sets was 0.499, indicating moderate overlap across repeated explanations.

This variability demonstrates that LIME feature rankings may depend on the local perturbation procedure. Consequently, individual explanations should be interpreted cautiously, particularly for clinical decision-support applications. Taken together, SHAP and LIME provided complementary insights into global feature contributions and individual classification behavior. The analyses highlighted the influence of hypertension, glucose-related measurements, and engineered interactions, while also illustrating misclassifications and limitations in explanation stability.

Discussion

Principal findings and comparative performance

This study investigated an explainable and imbalance aware hybrid ensemble framework for prevalent diabetes detection using the DiaBD dataset. The frame work employed stratified train-test partitioning, fold-specific preprocessing and resampling, and training-only model optimization to minimize information leak-age. Evaluation on the untouched held-out test set, which preserved the original diabetic prevalence of 6.43%, yielded a ROC-AUC of 0.8722 (95% CI: 0.8246-0.9144) and an Average Precision (AP) of 0.3561 (95% CI: 0.2602-0.4849). This evaluation provides a more reliable assessment of model discrimination under the original class distribution while reducing the risk of information leakage.

The proposed ensemble combined XGBoost, Random Forest, Gradient Boosting, multilayer perceptron, and LightGBM through optimized probability-level fusion. However, its ROC-AUC was only marginally higher than that of tuned XGBoost (0.8722 *vs.* 0.8714), with a paired bootstrap difference of 0.0009 (95% CI: −0.0132 to 0.0155). Although the ensemble achieved higher AP than XGBoost (0.3561 *vs.* 0.3250), the corresponding difference of 0.0315 (95% CI: −0.0042 to 0.0717) was also uncertain. Furthermore, Gradient Boosting achieved the highest individual AP (0.3896), while LightGBM provided the highest F1-score (0.4503) at the default threshold. Thus, the findings do not establish consistent superiority of ensemble fusion over well-tuned individual classifiers. Rather, they demonstrate that the integrated framework achieves competitive discrimination while supporting systematic evaluation of calibration, feature contributions, and classification trade-offs.

The independently evaluated FT-Transformer achieved a ROC-AUC of 0.8415 and AP of 0.3231, both lower than the corresponding ensemble values. Nevertheless, its sensitivity of 63.24% at the default threshold exceeded that of the ensemble (22.06%). This difference primarily illustrates the importance of operating-point behavior: A model with higher overall discrimination does not necessarily achieve higher sensitivity at a common numerical threshold. Moreover, the FT-Transformer underwent five-fold training only tuning rather than the nested cross validation used for the ensemble members. Consequently, the comparison supports the competitiveness of the proposed approach but does not establish a general advantage of ensemble learning over transformer based tabular models.

Class imbalance, feature engineering, and glucose dependence

The substantial class imbalance in DiaBD, with only 342 diabetic observations among 5,288 records, creates an important challenge for reliable minority-class identification

[4,26]. The revised framework addressed imbalance through model dependent class weighting and training-fold resampling, including Borderline-SMOTE for Random Forest and ADASYN for the multilayer perceptron. Restricting these procedures to training folds prevented synthetic observations from entering validation or test partitions, addressing a central limitation of the original experimental design. This approach is consistent with the established role of synthetic oversampling in imbalanced classification [29,30]. However, the present results do not independently establish the superiority of any particular resampling method because a controlled comparison of alternative imbalance-handling strategies was not performed.

Feature-ablation experiments showed that the six clinically motivated engineered variables contributed unequally to discrimination. Removing the age-glucose interaction produced the largest reduction in ROC-AUC (−0.0059), followed by mean arterial pressure (−0.0026) and the glucose/BMI ratio (−0.0015).

Conversely, removing pulse pressure increased ROC-AUC by 0.0049, while excluding the glucose-BMI interaction produced a negligible increase of 0.0002. These relatively small changes indicate that feature engineering did not uniformly improve model performance. Accordingly, the value of derived predictors should be assessed empirically rather than assumed from their clinical plausibility.

Excluding glucose and its derived features reduced ROC-AUC from 0.8722 to 0.7801, indicating substantial model dependence on glucose-related information. Although the original DiaBD study reported physician-reviewed diabetic-status classifications, the exact diagnostic thresholds and the role of glucose in label assignment were not explicitly documented [4]. Therefore, potential overlap between predictors and the outcome definition cannot be excluded, limiting conclusions about the model's independent diagnostic value.

Calibration, explainability, and clinical relevance

The results demonstrate that classification performance depends substantially on the selected operating threshold. At the default threshold of 0.5, the ensemble identified only 15 of 68 diabetic cases, corresponding to 22.06% sensitivity. Applying the training-derived Youden threshold (0.0911) increased sensitivity to 79.41%, while specificity decreased to 84.85%. The sensitivity-oriented threshold (0.0278) further increased sensitivity to 85.29%, but reduced specificity to 68.28% and positive predictive value to 15.59%. These findings illustrate the consequences of prioritizing diabetic-case identification in a low-prevalence population.

Although higher sensitivity may be desirable in screening-oriented applications, additional false-positive results can increase unnecessary follow-up testing, resource use, and patient burden. The training selected F1-optimal threshold (0.3338) achieved a slightly higher held-out F1-score than the Youden threshold (0.4029 versus 0.3971), but substantially lower sensitivity (41.18% *vs.* 79.41%). Therefore, maximizing F1-score alone does not necessarily produce the most appropriate operating point for a screening-oriented application. Clinical threshold selection would require explicit consideration of follow-up costs, the consequences of missed cases, and the intended care setting.

Probability calibration provided additional insight into prediction reliability. The ensemble's Brier score decreased from 0.0503 to 0.0493 following isotonic calibration, while XGBoost's Brier score decreased from 0.0575 to 0.0504. The corresponding ten-bin mean absolute calibration deviations were also lower for the ensemble than for XGBoost. Nevertheless, isotonic calibration produced only a modest absolute improvement for the ensemble, and the statistical significance of these differences was not established. Decision-curve analysis suggested positive net benefit for the ensemble over a range of relatively low threshold probabilities, although its advantage over XGBoost was generally small. These findings support evaluating discrimination, calibration, and clinical trade-offs jointly rather than relying on a single performance metric.

Explainability analysis further characterized the behavior of the fitted mod-els. SHAP identified hypertension status, glucose, and several glucose-related interactions features as influential predictors across the tree-based classifiers [8]. However, the weighted SHAP summary represented only the four tree-based ensemble components, which accounted for approximately 94.26% of the selected ensemble weight; it was not an exact explanation of the complete five-model ensemble. Moreover, correlations among glucose and its engineered interaction features may redistribute SHAP attribution across related predictors. Therefore, individual feature rankings should be interpreted cautiously, particularly without a separate SHAP stability assessment.

LIME explanations of representative Gradient Boosting predictions illustrated how individual feature conditions contributed to correct classifications and errors [7]. The explanation stability analysis yielded a mean Jaccard similarity of 0.499 across repeated runs, indicating that influential feature sets were not fully consistent. Moreover, correlations among glucose, BMI, and their engineered interactions may affect the distribution of feature attributions. Consequently, SHAP and LIME should be interpreted as descriptions of fitted model behavior rather than causal explanations or evidence of clinical reliability.

Taken together, these findings highlight the importance of integrating probability calibration, clinically relevant operating thresholds, and explainability into imbalance-aware diabetes classification. The proposed framework provides a comprehensive approach for assessing predictive reliability, interpreting model behavior, and evaluating classification trade-offs in region-specific healthcare datasets.

Limitations and future directions

This study has several limitations. First, the framework was evaluated using a single Bangladesh-specific dataset without external validation, which may limit generalizability to other populations and healthcare settings. Although stratified hold-out evaluation and nested cross-validation minimized information leakage, the relatively small number of diabetic cases in the test set introduced uncertainty in performance estimates.

Second, the study relied on previously collected diabetic-status labels for which the exact diagnostic criteria and potential overlap with glucose measurements were not fully documented. Consequently, the results should be interpreted as prevalent diabetes classification rather than prediction of future disease onset. Additionally, preprocessing and synthetic resampling

introduce modeling assumptions, while the selected ensemble weights require independent validation to assess their stability across datasets. Median imputation may obscure clinically meaningful variability and does not account for uncertainty associated with missing measurements, potentially affecting model performance and generalizability.

Finally, the explainability and clinical utility analyses were retrospective. The SHAP summary represented the tree-based ensemble components, while LIME explanations exhibited moderate stability. Subgroup fairness and prospective clinical effectiveness were not evaluated. Future research will focus on external validation, subgroup performance assessment, and prospective evaluation, alongside improvements in model calibration, ensemble optimization, and explanation reliability.

Conclusion and Future Work

This study developed an explainable and imbalance-aware hybrid ensemble framework for prevalent diabetes detection using the DiaBD dataset. The framework integrated XGBoost, Random Forest, Gradient Boosting, multilayer perceptron, and LightGBM through optimized probability-level fusion, along-side clinically motivated feature engineering and model-dependent imbalance handling. An FT-Transformer was independently evaluated as a deep tabular learning baseline. To minimize information leakage, model development employed stratified train-test partitioning, fold-specific preprocessing and resampling, and nested cross-validation.

On the untouched held-out test set, the ensemble achieved a ROC-AUC of 0.8722 (95% CI: 0.8246-0.9144) and an average precision of 0.3561 (95% CI: 0.2602-0.4849). Although its discrimination was comparable to that of tuned XGBoost, the observed improvement was not statistically established. At the training-derived Youden threshold, the ensemble achieved 79.41% sensitivity and 84.85% specificity, highlighting the importance of operating-point selection. Feature-ablation and glucose-exclusion analyses revealed substantial dependence on glucose-related information, while calibration, decision-curve analysis, and SHAP/LIME explanations provided complementary assessments of predictive reliability, potential clinical trade-offs, and model behavior.

Overall, the findings demonstrate the importance of leakage-aware validation and comprehensive evaluation when developing explainable diabetes classification models for region-specific populations. Future research should prioritize external validation using independent datasets with clearly documented diagnostic criteria, assessment of demographic subgroups, and prospective evaluation of clinical performance. Further investigation of alternative imbalance-handling strategies, calibration methods, and explanation stability may improve model reliability and generalizability. Together, these advances provide a foundation for developing more reliable, interpretable, and clinically relevant AI-assisted diabetes classification frameworks.

Credit Authorship Contribution Statement

Umme Habiba: Conceptualization, Methodology, Software, Writing - original draft preparation.

Simone A. Ludwig: Supervision, Writing - review and editing.

Data and Code Availability

The DiaBD dataset is publicly available through Prama et al. [4] (DOI: 10.1016/j.dib.2025.111746).

The analysis code, trained models, hyperparameter configurations, soft-ware dependencies, and experimental outputs are available in the accompanying GitHub reproducibility repository.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interests

The authors declare that they have no conflict of interest.

References

1. International Diabetes Federation, IDF Diabetes Atlas, 11th Edn. International Diabetes Federation, 2025.
2. Guan Z, Li H, Liu R, Cai C, Liu Y, Li J, et al. Artificial intelligence in diabetes management: Advancements, opportunities, and challenges. Cell Rep Med. 2023;4(10):101213.
3. Thotad PN, Bharamagoudar GR, Anami BS. Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods. Diabetes Metab Syndr. 2023;17(1):102690.
4. Prama TT, Rahman MJ, Zaman M, Sarker F, Mamun KA. DiaBD: A diabetes dataset for enhanced risk analysis and research in Bangladesh. Data Brief. 2025:61:111746.
5. Netayawijit P, Chansanam W, Sorn-In K. Interpretable machine learning framework for diabetes prediction: Integrating SMOTE balancing with SHAP explainability for clinical decision support. Healthcare. 2025;13(20):2588.
6. Al-Zebari A, Sengur A. Performance comparison of machine learning techniques on diabetes disease detection. In: 2019 1st International Informatics and Software Engineering Conference (UBMYK), IEEE; 2019. p. 1-4.
7. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM; 2016. p. 1135-44.
8. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems; 2017. p. 4765-74.
9. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters. 2023;10(1-2):1-10.
10. Hasan R, Dattana V, Mahmood S, Hussain S. Towards transparent diabetes prediction: Combining AutoML and explainable AI for improved clinical insights. Information. 2025;16(1):7.
11. Sadhasivam J, Muthukumaran V, Raja JT, Joseph RB, Munirathanam M, Balajee JM. Diabetes disease prediction using decision tree for feature selection. J Phys Conf Ser. 2021;1964(6):062116.

12. Khaleel FA, Al-Bakry AM. Diagnosis of diabetes using machine learning algorithms. Mater Today Proc. 2023;80:3200-3.

13. 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. 1988:261-5.

14. Abdulhadi N, Al-Mousa A. Diabetes detection using machine learning classification methods. In: 2021 International Conference on Information Technology (ICIT), IEEE; 2021. p. 350-4.

15. Wee BF, Sivakumar S, Lim KH, Wong WK, Juwono FH. Diabetes detection based on machine learning and deep learning approaches. Multimed Tools Appl. 2024;83(8):24153-85.

16. Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J Diabetes Metab Disord. 2020;19(1):391-403.

17. Yahyaoui A, Jamil A, Rasheed J, Yesiltepe M. A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International Informatics and Software Engineering Conference (UBMYK), IEEE; 2019. p. 1-4.

18. Sarwar MA, Kamal N, Hamid W, Shah MA. Prediction of diabetes using machine learning algorithms in healthcare. In: 2018 24th International Conference on Automation and Computing (ICAC), IEEE; 2018. p. 1-6.

19. Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Comput Sci. 2019;165:292-9.

20. 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, ACM; 2016. p. 785-94.

21. Khan FA, Zeb K, Al-Rakhami M, Derhab A, Bukhari SAC. Detection and prediction of diabetes using data mining: A comprehensive review. IEEE. 2021;9:43711-35.

22. 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.

23. Mbuya E, Mokheleli T, Bokaba T, Ndayizigamiye P. The application of artificial intelligence in diabetes prediction: A bibliometric analysis, in: Implications of Information and Digital Technologies for Development, IFIP Advances in Information and Communication Technology, Springer, 2024.

24. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods, Procedia Comput Sci. 2020;167:706-16.

25. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018:9:515.

26. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009;21(9):1263-84.

27. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, ACM; 2006. p. 233-40.

28. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.

29. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-57.

30. Han H, Wang WY, Mao BH. Borderline-SMOTE: A new oversampling method in imbalanced data sets learning. In: Huang DS, Zhang XP, Huang GB, editors. Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg; 2005. p. 878-87.

31. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The henry ford ExercIse testing project. PLoS One. 2017;12(7):e0179805.

32. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241-59.

33. Kuncheva LI. Combining pattern classifiers: Methods and algorithms. Wiley; 2004.

34. Breiman L. Random forests. Machine Learning. 2001;45:5-32.

35. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist. 2001;29(5):1189-232.

36. Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, et al. Early prediction of diabetes using an ensemble of machine learning models. Int J Environ Res Public Health. 2022;19(19):12378.

37. Edeh MO, Khalaf OI, Tavera CA, Tayeb S, Ghouali S, Abdulsahib GM, et al. A classification algorithm-based hybrid diabetes prediction model. Front Public Health. 2022:10:829519.

38. Habiba U, Ludwig SA. PSO-optimized TabTransformer architecture with feature engineering for enhanced cervical cancer risk prediction. Expert Syst Appl. 2026;300:130484.

39. Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express. 2018;4(4):243-6.

40. García-Ordás MT, Benavides C, Benítez-Andrades JA, Alaiz-Moretón H, García-Rodríguez I. Diabetes detection using deep learning techniques with oversampling and feature augmentation. Comput Methods Programs Biomed. 2021;202:105968.

41. Benito D, Aguilar J, Ramírez JM, Cantallops MM, Anta AF. A dynamic explainability method for fuzzy cognitive maps based on causal and temporal evolution analysis. Appl Soft Comput. 2026;202:115925.

42. Ganie SM, Dutta Pramanik PK, Bashir Malik M, Mallik S, Qin H. An ensemble learning approach for diabetes prediction using boosting techniques. Front Genet. 2023;14:1252159.

43. Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting deep learning models for tabular data, in: Advances in Neural Information Processing Systems. 2021;34:18932-43.

44. Ganji MF, Abadeh MS. A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis. Expert Syst Appl. 2011;38(12):14650-9.

45. Lekha S, Suchetha M. Real-time non-invasive detection and classification of diabetes using modified convolution neural network. IEEE J Biomed Health Inform. 2018;22(5):1630-6.

46. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.