AIMS Bioengineering, 13(4): 483–503.
DOI: 10.3934/bioeng.2026023
Received: 25 August 2026
Revised: 29 September 2026
Accepted: 09 October 2026
Published: 23 October 2026

Bioengineering

https://www.aimspress.com/journal/Bioengineering

Research article

A leakage-free and risk-aware machine learning framework for reliable diabetes prediction

Didar Abdalrahman Ismael*and Aram Rasul

Department of Computer Science, University of Raparin, Kurdistan Region, Iraq

* **Correspondence:** Email: didar.abdulrahman@uor.edu.krd; Tel: +9647729219595.

Abstract: Reliable diabetes prediction requires more than high discrimination; clinical risk models should also provide reliable probability estimates and account for the unequal consequences of prediction errors. This study proposes a leakage-free and risk-aware machine learning framework that integrates training-only preprocessing, probability calibration, cost-sensitive thresholding, statistical uncertainty assessment, subgroup stability analysis, and decision curve analysis (DCA). The framework was independently evaluated on the Pima Indians Diabetes Dataset (PIMA; $N = 768$) and the CDC Diabetes Health Indicators dataset ($N = 253{,}680$) using outer cross-validation, with preprocessing, model fitting, calibration, and threshold tuning performed without using the corresponding outer test folds. Logistic regression achieved the highest ROC-AUC on PIMA (0.8431; 95% CI: 0.8151–0.8695), while gradient boosting achieved the highest ROC-AUC on the CDC dataset (0.8298; 95% CI: 0.8276–0.8319). The calibration performance varied between methods and metrics, and there was no single method that consistently outperformed the others. Cost-sensitive analysis showed that lowering the decision threshold under higher assumed false-negative costs increased recall and reduced missed positive cases, but at the expense of more false positives. This trade-off was consistent across cost-ratio sensitivity analyses, while DCA indicated a positive model net benefit around the decision threshold corresponding to the primary 5:1 cost scenario in both datasets. In general, the results indicate that reliable evaluation of the diabetes risk model should extend beyond discrimination alone and jointly consider leakage control, probability calibration, asymmetric error costs, statistical uncertainty, subgroup stability, and potential decision utility.

Keywords: diabetes prediction; machine learning; data leakage; probability calibration; beta calibration; cost-sensitive learning; decision threshold; decision curve analysis

Abbreviations

AUC, area under the receiver operating characteristic curve; BS, Brier score; CDC, Centers for Disease Control and Prevention; CI, confidence interval; DCA, decision curve analysis; ECE, expected calibration error; FN, false negative; FNR, false-negative rate; FP, false positive; FPR, false-positive rate; GB, gradient boosting; LR, logistic regression; MCE, maximum calibration error; ML, machine learning; PIMA, Pima Indians Diabetes Dataset; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine.

1. Introduction

Diabetes mellitus is a common chronic disease associated with significant long-term complications and use of healthcare resources. Early identification of individuals at high risk of diabetes can facilitate early intervention and guide screening decisions. Thus, ML methods have been extensively studied for diabetes prediction and clinical decision support [1, 2].

Although predictive results are promising, several methodological and decision-related limitations remain. The first is data leakage, where preprocessing parameters are obtained using information from validation or test observations [3, 4]. Second, a classifier may be able to rank patients well but not produce probability estimates that match the observed outcome frequencies [5, 6]. Third, the conventional 0.5 classification threshold does not necessarily reflect the unequal consequences of false-negative and false-positive decisions [7, 8]. Fourth, point estimates alone can obscure substantial sampling uncertainty, particularly in small benchmark datasets. Finally, conventional predictive metrics do not directly determine whether model-guided decisions provide greater benefits than treating all or treating none [9, 10].

Recent diabetes-prediction research has increasingly used ensemble learning, deep neural networks, dimensionality reduction, biomarker-driven feature ranking, and explainability. These developments can improve discrimination and interpretation; however, high predictive performance does not by itself establish that preprocessing is leakage-safe, probabilities are calibrated, operating thresholds reflect unequal error consequences, or decisions have positive net benefit. This motivates an integrated framework in which model fitting, calibration, and decision-threshold tuning are performed without using the corresponding outer test folds.

The contributions of this study are sixfold. First, data-dependent preprocessing and model-fitting operations are confined to the corresponding outer-training folds, while calibration and threshold tuning are also performed without using the outer test observations. Second, probability reliability is assessed using Platt scaling, isotonic regression, beta calibration, and a Platt–isotonic ensemble, together with calibration intercept and slope. Third, theoretical cost-based thresholds are explicitly distinguished from empirically tuned thresholds. Fourth, a sensitivity analysis evaluates several false-negative-to-false-positive cost ratios rather than relying on a single operating assumption. Fifth, bootstrap confidence intervals and subgroup analyses quantify statistical uncertainty and stability across clinically relevant subgroups. Sixth, DCA evaluates potential decision utility across a broad threshold-probability range, and the complete framework is independently replicated on the large-scale CDC Diabetes Health Indicators dataset.

2. Related work

Data leakage has long been recognized as a source of biased evaluation. Kaufman *et al.* [3] formalized leakage in data mining, and Kuhn and Johnson [11] emphasized that data-dependent preprocessing should be placed within resampling procedures. Kapoor and Narayanan [4] also identified leakage as an important reproducibility problem. More recently, Alturayeif and Hassine [12] investigated automated leakage detection in ML code, while Rosenblatt *et al.* [13] demonstrated the inflationary effect of leakage in connectome-based prediction.

Diabetes prediction has been investigated using various conventional classifiers and increasingly sophisticated ensemble and deep-learning approaches [14–18]. Khan *et al.* [19] proposed a deep intelligent predictive model combining PseKNC, principal component analysis, and a deep neural network. More recently, Khan [20] proposed a biochemical biomarker-driven deep-learning framework with SHAP-based feature interpretation for diabetes classification. These studies illustrate the current emphasis on advanced representation learning, feature reduction, and explainability. In broader computational biomedicine, Almusallam *et al.* [21] proposed an integrated-feature deep-learning framework for RNA 5-methyluridine prediction; this is not a diabetes study and is considered here only as a methodological example of recent robustness-oriented biomedical prediction.

Khushal and Fatima [22] integrated fuzzy logic and machine learning for diabetes risk prediction based on modifiable parameters such as food, physical activity, and BMI. They developed a fuzzy transformation with an intermediate "possible diabetes risk" category to facilitate preventive action. This complements the present study's focus on reliable probability estimation and decision-oriented assessment rather than redefining the outcome space.

Probability calibration addresses a different problem from discrimination. Platt scaling provides a parametric sigmoid transformation [23], whereas isotonic regression provides a flexible monotonic mapping [24]. Beta calibration extends parametric calibration through a richer family of mappings that includes the identity transformation [25]. Niculescu-Mizil and Caruana [5] showed that strong classifiers can still produce unreliable probabilities, while Van Calster *et al.* [6] emphasized calibration as a core requirement for clinical risk prediction.

Cost-sensitive decision-making and DCA extend model evaluation beyond accuracy. Elkan [7] formalized classification under unequal costs, and Vickers and Elkin [9] introduced DCA to quantify net benefit across thresholds. These foundations motivated the unified framework proposed in this study.

3. Materials and methods

3.1. Datasets and prediction task

The Pima Indians Diabetes Dataset contains 768 observations, each with eight clinical predictors: Pregnancies, Glucose, Blood pressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age [26]. The true binary outcome indicates diabetic or non-diabetic status. Among the 768 observations, 268 (34.90%) were classified as diabetic and 500 (65.10%) as non-diabetic.

Let

$$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \qquad y_i \in \{0, 1\}, \tag{3.1}$$

and let the model estimate

$$\hat{p}(\mathbf{x}) = P(Y = 1 \mid \mathbf{x}). \tag{3.2}$$

Zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI were treated as medically invalid measurements and therefore converted to missing values before imputation.

The affected counts were 5/768 (0.65%) for Glucose, 35/768 (4.56%) for BloodPressure, 227/768 (29.56%) for SkinThickness, 374/768 (48.70%) for Insulin, and 11/768 (1.43%) for BMI. Zero values for Pregnancies were preserved because they are clinically valid. A preprocessing sensitivity analysis compared the primary training-fold median-imputation strategy with KNN imputation ($k = 5$) and with a diagnostic comparator that retained the original zero values. All three alternatives used the same stratified 10-fold outer cross-validation splits, the same deterministic feature engineering and scaling steps, and the selected PIMA classifier, with all data-dependent preprocessing fitted only on the corresponding training folds.

For independent-dataset replication, the complete framework was also evaluated on the CDC Diabetes Health Indicators dataset, which contained 253,680 observations and 21 health-related predictors [27]. The observed positive-class prevalence was 13.93%. Because the CDC and PIMA predictor spaces differ, the PIMA-trained classifier was not transferred directly to CDC. Instead, the same leakage-free and risk-aware evaluation framework was applied independently to CDC. Thus, the CDC experiment evaluates replication of the methodology rather than transportability of a single fitted PIMA model.

This study analyzed existing publicly available datasets. No new intervention, participant recruitment, or collection of identifiable human data was performed.

3.2. Leakage-controlled model development

A stratified 10-fold outer cross-validation procedure was used to preserve the class distribution across folds and to obtain out-of-fold predictions. For each outer fold k, the data were divided into an outer-training set, $D_{\text{train}}^{(k)}$, and an outer test set, $D_{\text{test}}^{(k)}$, such that

$$D_{\text{train}}^{(k)} \cap D_{\text{test}}^{(k)} = \varnothing. \tag{3.3}$$

Data-dependent preprocessing, model fitting, calibration, and threshold tuning were performed without using the corresponding outer test fold. For PIMA, medically invalid zero values were first replaced with missing values, and median imputation was performed using parameters estimated from the outer-training data only. The predefined feature-engineering transformations were then computed from the imputed predictors using fixed formulas without the outcome labels. Finally, the standardization parameters were estimated from the outer-training data. All training-derived preprocessing parameters were then applied unchanged to the outer test fold, thereby preventing information leakage from the test data during preprocessing.

Within each outer-training set, a stratified 5-fold inner cross-validation procedure was used for probability calibration, ensemble-weight selection, and empirical threshold tuning. Consequently, the corresponding outer test fold was not used for preprocessing parameter estimation, classifier fitting, calibration fitting, ensemble-weight selection, or empirical threshold selection. After completing model development using the outer-training data, the final trained and calibrated model was applied to the untouched outer test fold to generate predicted probabilities.

This process was repeated across all 10 outer folds. The resulting outer-test predictions were combined to form a complete set of out-of-fold (OOF) predictions, such that each observation received a prediction from a model that was not trained on that observation. The aggregated OOF predictions were then used for discrimination, calibration, cost-sensitive analysis, bootstrap uncertainty, subgroup stability, and decision-curve analyses.

The complete leakage-free and risk-aware evaluation framework is summarized in Figure 1.

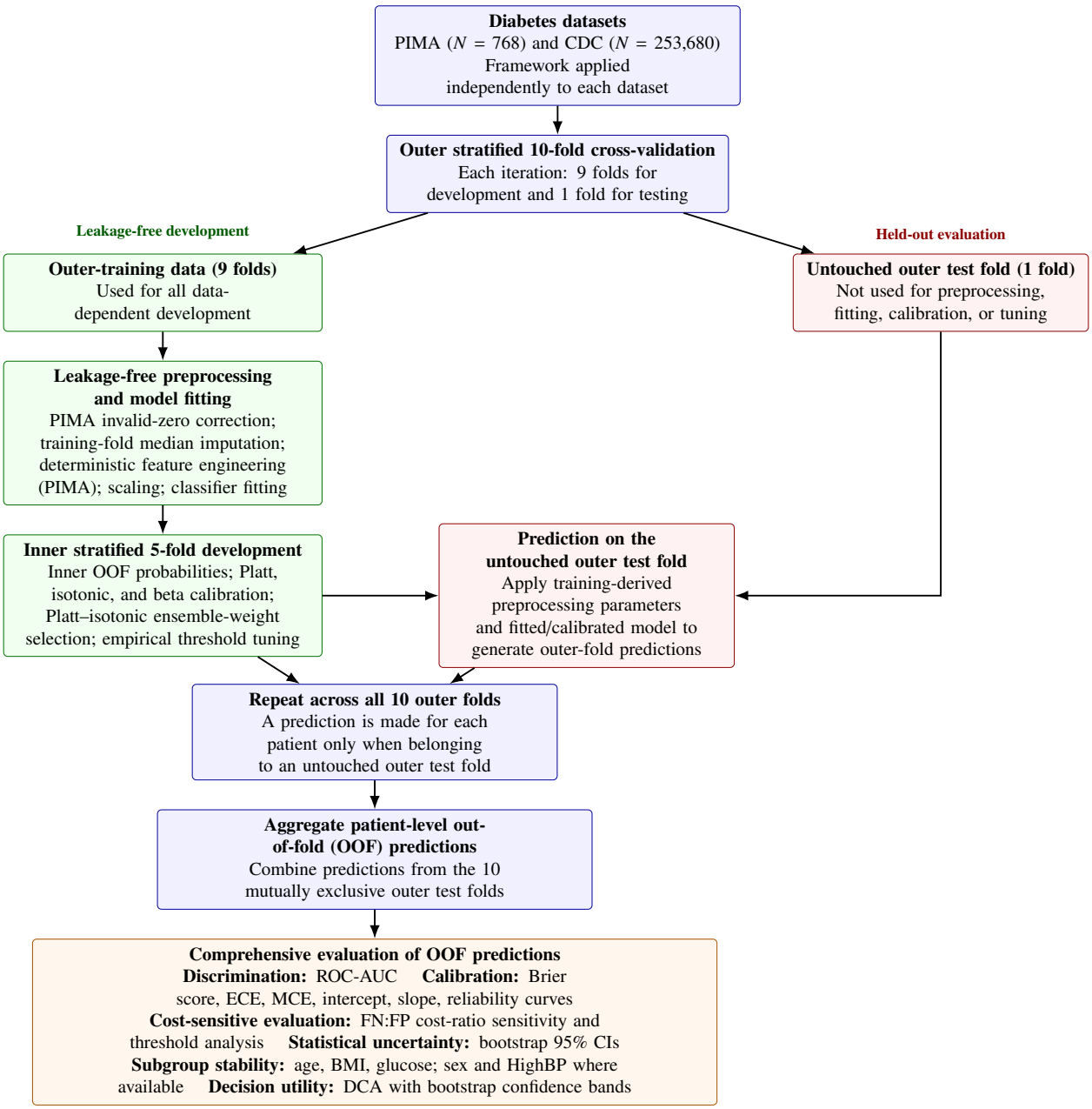


Figure 1. Overview of the proposed leakage-free and risk-aware machine learning framework, from outer-fold development to aggregated out-of-fold prediction and comprehensive evaluation.

Missing values were imputed using the training-fold median,

$$\tilde{x}_j = \text{median}\left(x_{ij} : i \in D_{\text{train}}\right), \tag{3.4}$$

and standardization used training-fold statistics,

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}. \tag{3.5}$$

No feature selection, SMOTE, oversampling, or undersampling was applied. Stratified cross-validation was used to approximately maintain the original class proportions across the training and test folds. The higher cost of false negatives was addressed through cost-sensitive decision thresholds rather than by resampling the data. Random forest additionally used balanced class weights during training.

3.3. Deterministic feature engineering

Four deterministic PIMA-specific features were introduced:

$$\begin{aligned} F_1 &= \text{BMI} \times \text{Age}, \\ F_2 &= \frac{\text{Glucose}}{\text{BMI} + \epsilon}, \\ F_3 &= \text{Pregnancies} \times \text{Age}, \\ F_4 &= \log(\text{Insulin} + 1). \end{aligned} \tag{3.6}$$

Note: The constant $\epsilon = 10^{-6}$ was used to avoid division by zero. These transformations depend solely on the current observations and are therefore leakage-safe when applied after missing values have been imputed. They were not reproduced on the CDC dataset because its predictor space differs.

The transformations were prespecified and clinically motivated rather than selected by searching for combinations that maximized outer-test performance. BMI-age and pregnancy-age terms represent plausible interactions between established risk-related characteristics; the glucose-to-BMI ratio provides a simple relative measure of glycemia to adiposity, and the logarithmic insulin transformation reduces the influence of the strongly right-skewed insulin scale. Their contribution was evaluated by comparing model performance with and without feature engineering. Because the CDC dataset does not contain the same glucose, insulin, and pregnancy characteristics, the same feature-engineering transformations could not be applied to the CDC dataset and were therefore not required for cross-dataset framework replication.

3.4. Classification models and probability calibration

Four supervised learning algorithms were considered for PIMA: LR, SVM with an RBF kernel, RF, and GB. Because of the substantially larger CDC sample, LR, RF, and GB were used for the complete framework replication on the CDC dataset. The models were compared using outer-fold predictions, and the classifier with the highest pooled ROC-AUC was selected for calibration and risk-aware evaluation within each dataset. This selection was not nested within the outer-training folds.

The RBF-SVM was not included in the CDC experiment because probability-enabled kernel SVM training becomes substantially more computationally demanding at the CDC sample size, particularly

when embedded within repeated nested fitting and calibration. This was a computational-scope decision rather than a claim that SVM was unsuitable for use with CDC data. The purpose of the CDC study was to apply the complete evaluation framework to a substantially larger dataset rather than to enforce the same set of classifiers across datasets.

Probability calibration was performed using Platt scaling, isotonic regression, beta calibration, and a Platt–isotonic ensemble. Platt scaling uses a logistic function to transform the model output into a calibrated probability,

$$\hat{p}_{\text{Platt}} = \frac{1}{1 + \exp[-(Af(\mathbf{x}) + B)]}, \tag{3.7}$$

whereas isotonic regression estimates a monotonic mapping,

$$\hat{p}_{\text{Iso}} = g(\hat{p}_{\text{raw}}). \tag{3.8}$$

Beta calibration was implemented through logistic regression on $\log(\hat{p}_{\text{raw}})$ and $-\log(1 - \hat{p}_{\text{raw}})$, following Kull *et al.* [25].

The ensemble was defined as

$$\hat{p}_{\text{Ens}} = \alpha \hat{p}_{\text{Platt}} + (1 - \alpha)\hat{p}_{\text{Iso}}, \tag{3.9}$$

with $\alpha \in \{0, 0.1, \ldots, 1\}$. Importantly, α was not selected for the outer test predictions. Within each outer-training fold, temporary Platt and isotonic mappings were fitted using inner-training predictions, and the weight was selected by the Brier score using held-out inner information; the calibration mapping was then refitted within the outer-training data and applied to the untouched outer test fold. Consequently, the weight was allowed to vary across outer folds rather than being treated as a globally fixed optimum.

Calibration was evaluated using the Brier score (BS), expected calibration error (ECE), maximum calibration error (MCE), calibration intercept, calibration slope, and reliability curves. BS, ECE, and MCE were defined as

$$\begin{aligned} BS &= \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2, \\ ECE &= \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \\ MCE &= \max_m \left|\text{acc}(B_m) - \text{conf}(B_m)\right|. \end{aligned} \tag{3.10}$$

Note: ECE and MCE were computed using $M = 10$ equal-width probability bins.

These measures were preserved because they provide complementary views of probabilistic performance. The Brier score summarizes the overall squared probability error and therefore reflects both calibration and discrimination-related probability accuracy; ECE summarizes the average absolute discrepancy between predicted and observed event frequencies across probability bins; and MCE records the largest such bin-level discrepancy. No single one of these measures is treated as the definitive clinically informative calibration measure. For that reason, they are interpreted jointly with calibration intercept, calibration slope, and reliability curves. ECE and MCE are also acknowledged as bin-dependent summaries.

The calibration intercept and slope were obtained from

$$\text{logit}\, P(Y = 1) = a + b\ \text{logit}(\hat{p}), \tag{3.11}$$

for which ideal calibration corresponds to $a = 0$ and $b = 1$.

3.5. Cost-sensitive decision rule

For a calibrated probability $p = P(Y = 1 \mid X = x)$, the conditional expected loss of predicting negative is $C_{FN}p$, whereas the expected loss of predicting positive is $C_{FP}(1 - p)$. A positive decision is preferred when

$$C_{FP}(1 - p) < C_{FN}p, \tag{3.12}$$

which yields the cost-sensitive decision threshold

$$t^* = \frac{C_{FP}}{C_{FN} + C_{FP}}. \tag{3.13}$$

For the primary experimental assumption $C_{FN} : C_{FP} = 5 : 1$, $t^* = 1/6 \approx 0.1667$. The cost ratio is a decision assumption rather than a universal clinical constant.

The theoretical threshold was distinguished from a data-dependent empirical threshold. Within each outer-training fold, the empirical threshold was selected from inner out-of-fold probabilities calibrated using those same predictions, as follows:

$$t_{\text{emp}} = \arg\min_{t \in \mathcal{T}} \left[C_{FN}FN(t) + C_{FP}FP(t) \right], \tag{3.14}$$

with $\mathcal{T} = \{0.01, 0.02, \ldots, 0.99\}$, and was then applied unchanged to the corresponding outer test fold. The sensitivity analysis considered $C_{FN} : C_{FP} \in \{1 : 1, 2 : 1, 3 : 1, 5 : 1, 10 : 1\}$.

3.6. Statistical uncertainty and subgroup stability

Statistical uncertainty was quantified using bootstrap resampling of observations and their corresponding out-of-fold (OOF) predictions. We used 2,000 bootstrap replicates for PIMA and 1,000 for CDC because of the substantially larger size and computational cost of the CDC dataset. 95% percentile confidence intervals were reported for ROC-AUC, Brier score, ECE, recall, false-negative count/rate, calibration intercept, and calibration slope. DCA confidence bands were obtained by bootstrap resampling of the full net-benefit curve in each replicate, rather than by independently bootstrapping individual threshold points.

Subgroup stability was evaluated at the prespecified 5:1 theoretical threshold. PIMA analyses considered age, BMI, and glucose subgroups. CDC analyses considered sex, the dataset's ordinal age categories, BMI subgroups, and high-blood-pressure status. For each subgroup, the ROC-AUC, Brier score, recall, FNR, FPR, and bootstrap 95% confidence intervals were calculated. These analyses were intended to assess whether model performance remained stable across subgroups and should not be interpreted as a complete fairness audit. Sex-specific analysis is not possible in PIMA because the dataset contains only female participants.

3.7. Decision curve analysis

The DCA net benefit for the model was calculated as

$$NB_{\text{model}} = \frac{TP}{N} - \frac{FP}{N}\left(\frac{p_t}{1 - p_t}\right). \tag{3.15}$$

A treat-none approach offers no net benefit, while a treat-all approach depends on prevalence and p_t.

DCA was evaluated from $p_t = 0.01$ to 0.80 with bootstrap 95% confidence bands. The broad range was retained to show that clinical utility is threshold-dependent rather than restricted to the primary 5:1 operating point.

4. Experimental results and discussion

4.1. Baseline discrimination

The baseline model performance is summarized in Table 1. On PIMA, LR achieved the highest ROC-AUC of 0.8431 (95% CI: 0.8151–0.8695), whereas on the CDC dataset, GB achieved the highest ROC-AUC of 0.8298 (95% CI: 0.8276–0.8319); therefore, these models were selected for subsequent calibration and risk-aware analysis. The wider confidence interval for PIMA is consistent with its smaller sample size.

Table 1. Baseline outer-fold model performance.

Dataset	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Brier	ECE	MCE
PIMA	LR	0.7565	0.6884	0.5522	0.6128	**0.8431**	0.1548	0.0529	0.1262
	SVM	0.7565	0.6919	0.5448	0.6096	0.8288	0.1608	0.0511	0.1482
	RF	0.7578	0.6314	0.7351	0.6793	0.8302	0.1620	0.0548	0.1065
	GB	0.7747	0.6892	0.6455	0.6667	0.8336	0.1598	0.0632	0.1786
CDC	LR	0.8635	0.5350	0.1555	0.2410	0.8219	0.0991	0.0124	0.4365
	RF	0.8245	0.3872	0.4451	0.4142	0.8012	0.1209	0.0926	0.6776
	GB	0.8664	0.5713	0.1659	0.2571	**0.8298**	0.0967	0.0057	0.1390

Note: Accuracy, precision, recall, and F1-score are calculated at the conventional decision threshold of 0.5; ROC-AUC is threshold-independent.

LR and GB achieved ROC-AUC values of 0.8219 and 0.8298 on CDC; however, their recall at the conventional threshold of 0.5 was only 0.1555 and 0.1659, respectively. This does not contradict their discrimination performance, as ROC-AUC measures ranking over all thresholds, whereas recall depends on a single decision cutoff. Many positive cases received predicted probabilities below 0.5, despite the models maintaining useful ranking ability. The positive-class prevalence in the CDC was approximately 13.9%. Therefore, the low recall at 0.5 motivated the subsequent cost-sensitive analysis. Lowering the decision threshold identified more positive cases but also increased false positives.

4.2. PIMA feature-engineering comparison

Table 2 displays the performance of the PIMA model with and without the proposed feature engineering. The impact of feature engineering was not universally positive but depended on the particular model and statistic. Feature engineering improved the ROC-AUC and Brier score for LR and SVM. For GB, accuracy, recall, F1, and ROC-AUC increased somewhat, but the Brier score increased slightly. For RF, recall and F1 increased, while accuracy and the Brier score decreased marginally, and the ROC-AUC remained almost the same. These results suggest that the contribution of designed features depends on the classifier and the evaluation metric.

Table 2. PIMA model performance with and without the proposed feature engineering.

Model	Feature engineering	Accuracy	Recall	F1-score	ROC-AUC	Brier
Logistic regression	Without	0.7669	0.5560	0.6247	0.8345	0.1573
	With	0.7565	0.5522	0.6128	0.8431	0.1548
SVM	Without	0.7487	0.5299	0.5954	0.8233	0.1636
	With	0.7565	0.5448	0.6096	0.8288	0.1608
Random forest	Without	0.7604	0.7127	0.6749	0.8303	0.1614
	With	0.7578	0.7351	0.6793	0.8302	0.1620
Gradient boosting	Without	0.7708	0.6269	0.6563	0.8318	0.1595
	With	0.7747	0.6455	0.6667	0.8336	0.1598

4.3. Preprocessing sensitivity analysis

The PIMA preprocessing sensitivity analysis is summarized in Table 3. Median and KNN imputation produced very similar discrimination and overall probability error. The primary median-imputation pipeline achieved an ROC-AUC of 0.8431 and a Brier score of 0.1548, whereas KNN imputation achieved an ROC-AUC of 0.8429 and a Brier score of 0.1545. The recall was 0.5522 with median imputation and 0.5634 with KNN imputation. When the medically invalid zero values were preserved as a diagnostic comparator, the ROC-AUC dropped to 0.8344, and the Brier score increased to 0.1569. The ECE was lower for the retained-zero comparator (0.0302), showing that a single bin-based calibration summary cannot be considered alone. Overall, the close results for median and KNN imputation indicate that the principal PIMA findings were not strongly dependent on the specific imputation method, while the diagnostic comparator confirms that the zero-handling decision can affect individual performance measures.

Table 3. PIMA preprocessing sensitivity analysis using leakage-free 10-fold out-of-fold predictions.

Preprocessing	ROC-AUC	Brier	ECE	Accuracy	Recall	F1-score
Zero→NaN + median imputation (primary)	0.8431	0.1548	0.0529	0.7565	0.5522	0.6128
Zero→NaN + KNN imputation ($k = 5$)	0.8429	0.1545	0.0594	0.7539	0.5634	0.6151
Zeros retained (diagnostic comparator)	0.8344	0.1569	0.0302	0.7643	0.5597	0.6237

4.4. Calibration comparison

Table 4 compares the four calibration methods with the corresponding uncalibrated model selected for each dataset. No method consistently outperformed the others across all calibration criteria. On PIMA, isotonic regression achieved the lowest Brier score (0.1519), while the Platt–isotonic ensemble achieved the lowest ECE (0.0304). Platt and beta calibration produced calibration slopes that were closest to the ideal value of one. For the CDC dataset, differences in Brier score were very small; the ensemble achieved the lowest Brier score (0.0967), while isotonic regression achieved the lowest ECE (0.0009). The relatively high MCE for beta calibration on PIMA (0.4242) was associated with a

sparsely populated probability bin, despite its favorable overall calibration intercept and slope.

Table 4. Calibration comparison for the selected model in each dataset.

Dataset	Method	ROC-AUC	Brier	ECE	MCE	Intercept	Slope
PIMA	Raw	0.8431	0.1548	0.0529	0.1262	-0.0303	0.9416
	Platt	0.8431	0.1548	0.0433	0.1527	-0.0026	0.9998
	Isotonic	0.8390	**0.1519**	0.0335	0.1922	-0.0796	0.8080
	Beta	0.8426	0.1542	0.0357	0.4242	-0.0020	1.0068
	Ensemble	0.8421	0.1526	**0.0304**	0.2257	-0.0268	0.9561
CDC	Raw	0.8298	0.0967	0.0057	0.1390	0.0408	1.0308
	Platt	0.8298	0.0968	0.0055	0.0466	0.0007	1.0009
	Isotonic	0.8296	0.0967	**0.0009**	0.5667	-0.0065	0.9951
	Beta	0.8298	0.0967	0.0034	**0.0399**	-0.0001	1.0000
	Ensemble	0.8297	**0.0967**	0.0025	0.1187	0.0011	1.0012

Bootstrap results further quantify the uncertainty. For PIMA, the ensemble Brier score was 0.1526 (95% CI: 0.1387–0.1666), ECE was 0.0304 (95% CI: 0.0277–0.0686), the calibration intercept was -0.0268 (95% CI: -0.2169–0.1736), and the slope was 0.9561 (95% CI: 0.7985–1.1452). For CDC, the ensemble Brier score was 0.096714 (95% CI: 0.095954–0.097476), ECE was 0.002477 (95% CI: 0.001673–0.003807), the calibration intercept was 0.0011 (95% CI: -0.0165–0.0204), and the slope was 1.0012 (95% CI: 0.9902–1.0122). The ideal intercept of 0 and slope of 1 fall within the ensemble intervals in both datasets; these results support the position that there is an absence of strong evidence of gross overall miscalibration but should not be interpreted as proof of calibration equivalence.

Importantly, the nested ensemble weight varied across the outer folds. On PIMA, the selected α values ranged from 0.1 to 1.0 (mean 0.50), indicating considerable sample dependence. For the CDC dataset, the selected values were more concentrated, ranging from 0.3 to 0.5. Therefore, a fixed value such as $\alpha = 0.30$ should not be interpreted as a universal optimum.

Figures 2 and 3 present the reliability diagrams for PIMA and CDC, respectively. The plots compare predicted diabetes probabilities with observed outcome frequencies, using the diagonal line as the reference for perfect calibration. Deviations from this line indicate overestimation or underestimation of risk, providing a visual assessment of probability calibration across both datasets.

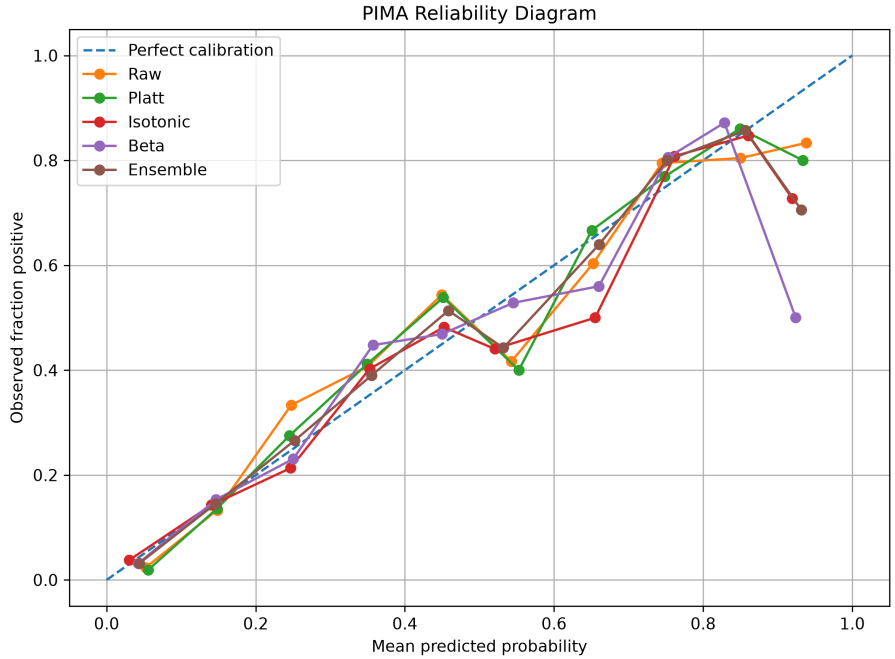


Figure 2. PIMA reliability comparison for raw, Platt, isotonic, beta, and nested ensemble probabilities. High-probability deviations should be interpreted cautiously because PIMA contains only 768 observations.

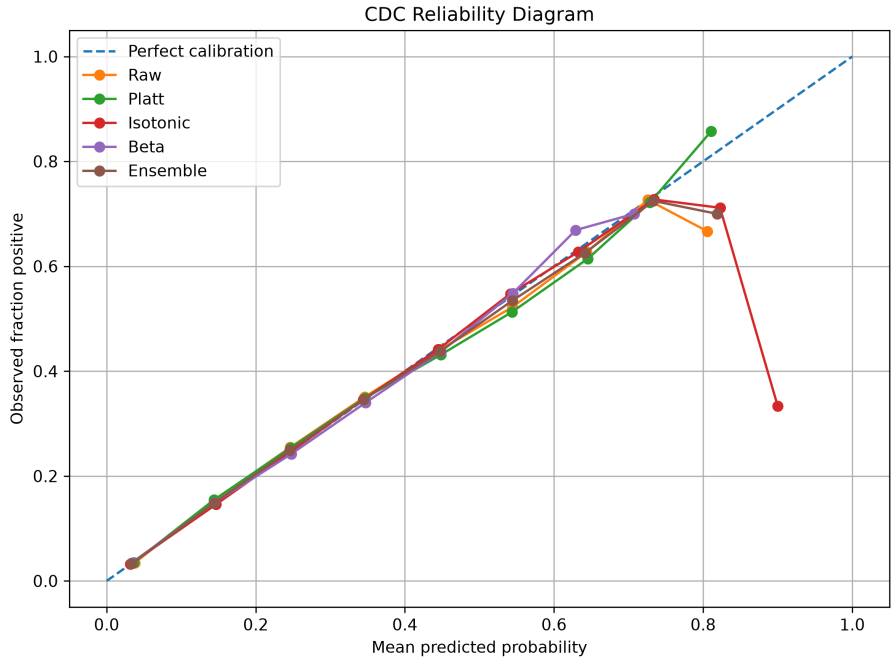


Figure 3. CDC reliability comparison for raw, Platt, isotonic, beta, and nested ensemble probabilities.

4.5. Cost-ratio sensitivity and empirical threshold tuning

Table 5 illustrates how the theoretical decision threshold and the resulting recall, false positives, and false negatives change as the false-negative-to-false-positive cost ratio increases. In both datasets, increasing $C_{FN} : C_{FP}$ lowers the theoretical decision threshold, increases recall, and reduces false negatives while increasing the number of false positives. This indicates that the 5:1 cost ratio represents an assumed clinical scenario rather than a universally applicable setting for diabetes prediction.

Table 5. Theoretical cost-ratio sensitivity using nested ensemble probabilities.

FN:FP	Theoretical threshold (t^*)	PIMA			CDC		
		Recall	FN	FP	Recall	FN	FP
1:1	0.5000	0.5634	117	68	0.1578	29,769	4,051
2:1	0.3333	0.8022	53	140	0.4223	20,418	18,051
3:1	0.2500	0.8881	30	189	0.5823	14,765	31,489
5:1	0.1667	0.9478	14	244	0.7387	9,236	52,848
10:1	0.0909	0.9813	5	321	0.8819	4,175	86,635

Under the primary 5:1 cost scenario, the two approaches (i.e., the prespecified theoretical threshold and the nested empirical threshold) produced similar results (Table 6). On PIMA, the theoretical threshold resulted in a slightly lower expected cost than the nested empirical threshold (314 vs. 322). For the CDC dataset, the nested empirical threshold resulted in a slightly lower expected cost than the theoretical threshold (98,862 vs. 99,028). The empirical thresholds varied from 0.08 to 0.19 across the PIMA outer folds but were concentrated at 0.16–0.17 across the CDC folds. These findings do not support presenting a single empirically selected threshold from the full dataset as an independently validated optimum value.

Table 6. Comparison of theoretical and empirically tuned decision thresholds under the 5:1 cost scenario.

Dataset	Threshold approach	Threshold	Recall	FP	FN	Expected cost
PIMA	Theoretical	0.1667	0.9478	244	14	314
	Empirical (nested)	fold-specific (0.08–0.19)	0.9440	247	15	322
CDC	Theoretical	0.1667	0.7387	52,848	9,236	99,028
	Empirical (nested)	fold-specific (0.16–0.17)	0.7492	54,537	8,865	98,862

The operational interpretation is also important. Under the 5:1 cost scenario, the lower decision threshold substantially increases recall but also increases the number of false positives. In a deployment setting, these additional positive predictions may require further clinical assessment, testing, or follow-up, thereby increasing the workload for healthcare services. The decision threshold should therefore be selected by considering both local clinical capacity and clinically justified error costs rather than recall alone.

4.6. Decision curve analysis

DCA was evaluated across threshold probabilities from 0.01 to 0.80. At $p_t = 0.17$, which is close to the theoretical threshold of 0.1667 under the primary 5:1 cost scenario, the PIMA model achieved a net benefit of 0.2636 (95% CI: 0.2254–0.3023), compared with 0.2156 for treat-all and 0 for treat-none. On CDC, at the same threshold probability, the model achieved a net benefit of 0.0603 (95% CI: 0.0590–0.0615), compared with -0.0369 for treat-all and 0 for treat-none. Therefore, at a threshold probability of approximately 0.17, corresponding to the primary 5:1 cost scenario, the model provided greater estimated net benefit than both the treat-all and treat-none strategies. However, the full decision curves show that the relative net benefit depends on the selected threshold probability.

Figures 4 and 5 present the decision curve analyses for PIMA and CDC, respectively. The plots compare model net benefit with the treat-all and treat-none strategies across threshold probabilities, with bootstrap 95% confidence bands indicating estimation uncertainty. A higher model net benefit indicates greater potential clinical utility at the corresponding decision threshold.

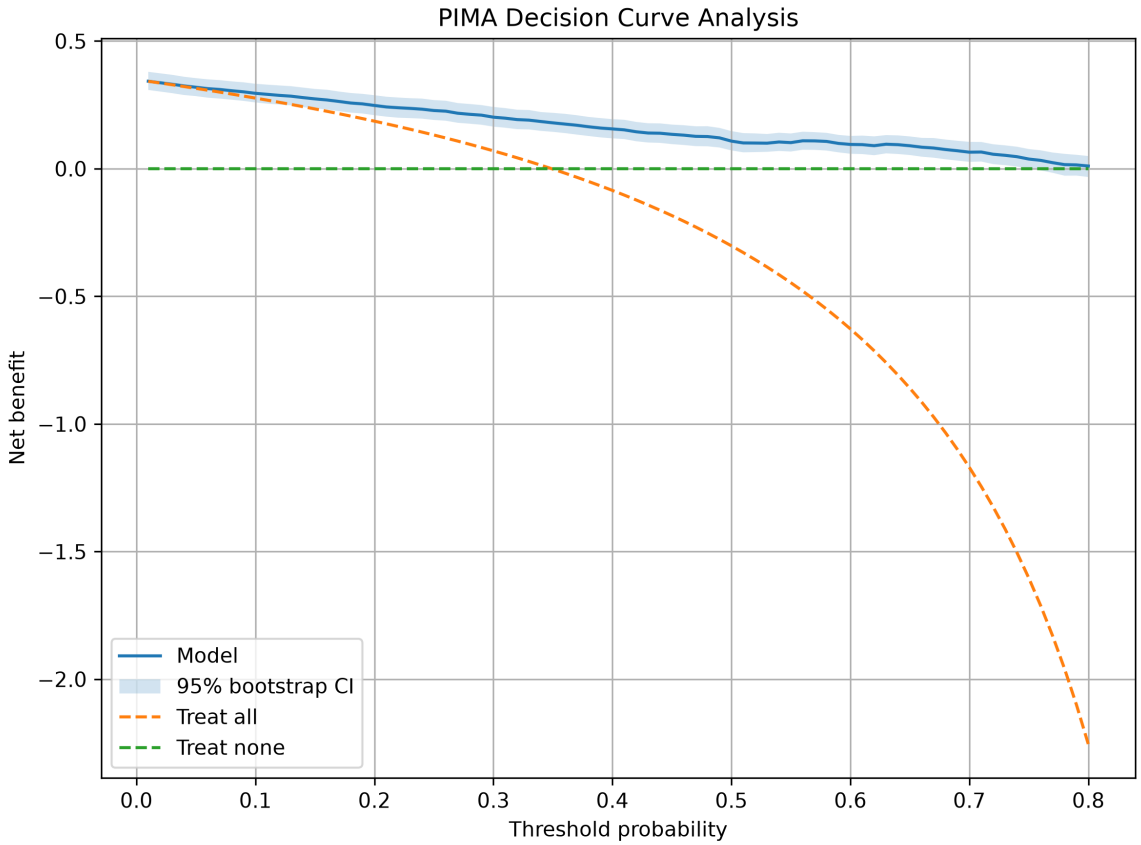


Figure 4. PIMA decision curve analysis across threshold probabilities with bootstrap 95% confidence bands.

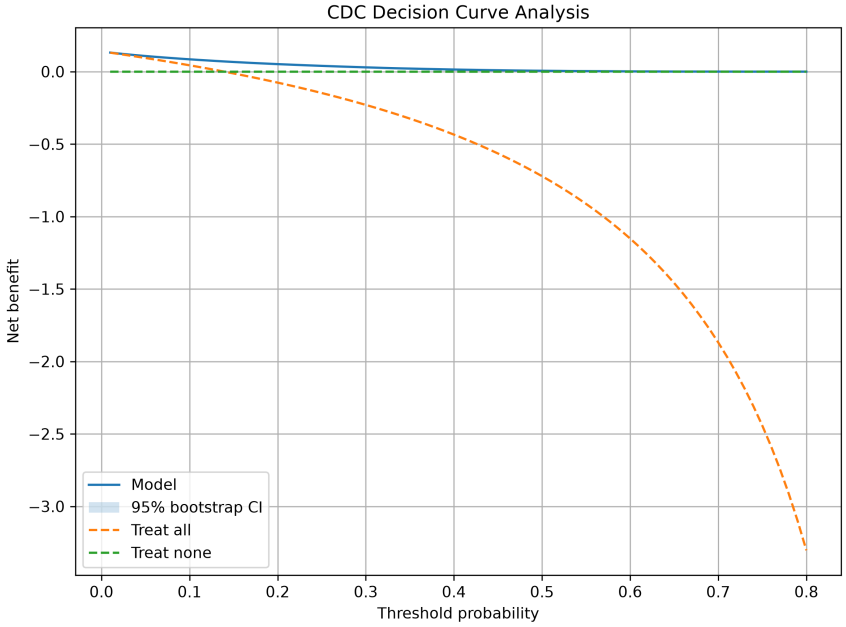


Figure 5. CDC decision curve analysis across threshold probabilities with bootstrap 95% confidence bands.

4.7. Subgroup stability

Subgroup analyses showed that model performance varied across patient subgroups despite the overall performance observed in the full datasets. Table 7 illustrates model performance across selected subgroups at the 5:1 theoretical threshold; complete subgroup results, including bootstrap intervals for ROC-AUC, Brier score, recall, FNR, and FPR, are provided as supplementary material.

Table 7. Representative subgroup operating characteristics at the theoretical 5:1 threshold.

Dataset	Subgroup	N	Recall (95% CI)	FPR (95% CI)
PIMA	Age < 40	561	0.931 (0.890–0.968)	0.414 (0.364–0.463)
	Age ≥ 60	32	1.000 (1.000–1.000)	0.913 (0.783–1.000)
	BMI < 25	117	0.889 (0.666–1.000)	0.194 (0.120–0.271)
	Glucose < 100	197	0.625 (0.385–0.857)	0.144 (0.092–0.197)
	Glucose ≥ 126	297	1.000 (1.000–1.000)	0.909 (0.852–0.956)
CDC	Female	141,974	0.735 (0.729–0.742)	0.220 (0.217–0.222)
	Male	111,706	0.742 (0.736–0.749)	0.271 (0.268–0.274)
	BMI < 25	72,080	0.404 (0.390–0.418)	0.076 (0.074–0.078)
	BMI ≥ 30	87,851	0.858 (0.853–0.863)	0.451 (0.447–0.455)
	HighBP: no	144,851	0.395 (0.385–0.405)	0.074 (0.073–0.075)
	HighBP: yes	108,829	0.852 (0.848–0.856)	0.520 (0.517–0.523)

On PIMA, recall was high in many subgroups, but FPR varied substantially. For example, the glucose ≥ 126 subgroup had a recall of 1.000 (95% CI: 1.000–1.000) and an FPR of 0.909 (95% CI:

0.852–0.956), whereas the glucose < 100 stratum had a recall of 0.625 (95% CI: 0.385–0.857) and an FPR of 0.144 (95% CI: 0.092–0.197). Estimates for small categories were more uncertain; for example, the age ≥ 60 group contained only 32 observations and had an ROC-AUC of 0.749 (95% CI: 0.519–0.937).

On CDC, the much larger sample produced narrower confidence intervals. Female and male recall were 0.735 (95% CI: 0.729–0.742) and 0.742 (95% CI: 0.736–0.749), respectively, while FPR was 0.220 (95% CI: 0.217–0.222) in females and 0.271 (95% CI: 0.268–0.274) in males. BMI-stratified recall increased from 0.404 for BMI < 25 to 0.858 for BMI ≥ 30, accompanied by an increase in FPR from 0.076 to 0.451. These differences indicate variation in model performance across subgroups and should not be interpreted as evidence of algorithmic unfairness. However, they highlight the importance of evaluating model performance across relevant patient subgroups before local clinical deployment.

4.8. Cross-dataset summary

Table 8 summarizes the main findings. Although the selected classifier differed between the two datasets, the overall methodological pattern was consistent: calibration affected probability calibration more than discrimination; the assumed cost ratio determined the theoretical decision threshold; nested empirical threshold tuning produced results similar to the theoretical threshold; and DCA showed positive net benefit around the threshold corresponding to the primary 5:1 cost scenario.

Table 8. Cross-dataset comparison under the reported evaluation.

Characteristic	PIMA	CDC diabetes health indicators
Number of observations	768	253,680
Number of original predictors	8	21
Positive-class prevalence	34.90%	13.93%
Best baseline model	LR	GB
Best baseline ROC-AUC	0.8431	0.8298
Lowest Brier after calibration	0.1519 (Isotonic)	0.096714 (Ensemble)
Lowest ECE after calibration	0.0304 (Ensemble)	0.000872 (Isotonic)
Theoretical 5:1 threshold	0.1667	0.1667
Recall at theoretical threshold	0.9478	0.7387
False negatives at theoretical threshold	14	9,236
Expected cost at theoretical threshold	314	99,028
Nested empirical 5:1 threshold range	0.08–0.19	0.16–0.17

5. Discussion

The findings support a multi-dimensional view of model reliability. Leakage-free evaluation is not intended to guarantee better numerical performance but to prevent held-out observations from influencing model development. The outer-training procedure extends this principle beyond preprocessing to calibration, ensemble-weight selection, and empirical threshold tuning. Under this stricter evaluation, LR achieved the highest ROC-AUC on PIMA, whereas GB achieved the highest

ROC-AUC on CDC. The wider PIMA bootstrap confidence interval also reflects the greater statistical uncertainty associated with its substantially smaller sample size.

The calibration results showed that no method was universally superior. On PIMA, isotonic regression achieved the lowest Brier score, whereas the Platt–isotonic ensemble achieved the lowest ECE; Platt and beta calibration produced slopes closest to 1. On CDC, the ensemble achieved the lowest Brier score, although the differences were very small, while isotonic regression achieved the lowest ECE. The substantial variation in PIMA fold-specific α values also cautions against treating a fixed ensemble weight as a stable optimum, whereas the CDC weights were more concentrated.

The distinction between theoretical and empirical threshold selection is important. The theoretical threshold follows directly from the assumed decision costs, whereas an empirical threshold is data-dependent and must be tuned without using the observations upon which it is evaluated. Under the primary 5:1 cost scenario, the prespecified theoretical threshold performed comparably to empirical threshold tuning within the outer-training folds on both datasets. The sensitivity analysis further showed that 0.1667 is not a universal diabetes decision threshold: increasing the assumed cost of false negatives relative to false positives lowered the theoretical threshold, increased recall, and reduced false negatives while increasing false positives.

This trade-off has important clinical workflow implications. A lower threshold can reduce the number of missed high-risk individuals but also increase the number of people who require further assessment. For example, the 5:1 theoretical threshold produced 52,848 false positives in the CDC analysis. Such a workload may be acceptable in some screening pathways but impractical in others. Therefore, the choice of decision threshold should consider clinically justified error costs, additional clinical testing, clinician workload, patient burden, and local resource capacity.

Subgroup analyses further showed that model performance varied across patient groups. At the theoretical threshold corresponding to the 5:1 cost scenario, recall and FPR varied across the evaluated age, BMI, glucose, sex, and high-blood-pressure subgroups. Some PIMA estimates were uncertain because of small subgroup sizes, whereas CDC estimates were more stable. These findings do not establish algorithmic unfairness but highlight the importance of evaluating model performance across relevant patient subgroups before deployment. Sex-based analysis was not possible on PIMA because its participants were female only.

DCA provided a complementary evaluation of the potential decision utility. The net benefit of the model was greater than the treat-all and treat-none strategies on PIMA and CDC at approximately $p_t \approx 0.17$, with bootstrap confidence bands quantifying the statistical uncertainty. However, these findings demonstrate potential decision utility under the studied assumptions rather than clinical effectiveness. Prospective clinical evaluation remains necessary before deployment. Finally, the CDC analysis should be interpreted as an independent evaluation of the proposed framework rather than evidence that a PIMA-trained model or numerical threshold is directly transportable.

PIMA and CDC differ substantially in terms of sample size, predictor space, prevalence, and population composition. Model development, calibration, and empirical threshold tuning were therefore performed independently within each dataset. The common theoretical threshold of 0.1667 was obtained solely by applying the same prespecified 5:1 cost scenario to both datasets. Local recalibration and clinically justified cost assumptions would still be required before applying the framework in a specific clinical setting.

The preprocessing sensitivity analysis further showed that the main PIMA results were stable to

replacing training-fold median imputation with KNN imputation: ROC-AUC changed only from 0.8431 to 0.8429 and the Brier score from 0.1548 to 0.1545. Retaining medically invalid zeros produced mixed metric changes, including a lower ROC-AUC (0.8344) and a higher Brier score (0.1569), despite a lower ECE. This finding reinforces the need to report several complementary probability-quality measures and to make the handling of structurally invalid values explicit.

Future work can investigate meta-learning approaches for better cross-dataset adaptation and generalizability. Such approaches would need to account for differences in predictor spaces, outcome definitions, disease prevalence, and calibration across cohorts. Because PIMA and the CDC dataset do not share identical feature sets, cross-dataset adaptation remains an important direction for future investigation rather than a component evaluated in the present study.

6. Limitations

The primary PIMA benchmark is small, restricted to female participants from a specific population, and contains only eight original clinical predictors. Therefore, subgroup estimates can have wide confidence intervals, and calibration estimates in sparsely populated probability regions can be unstable. Although using the CDC dataset substantially increases sample size and population diversity, its predictor space and data-collection characteristics differ from PIMA; therefore, the CDC analysis represents independent replication of the framework rather than direct external validation of a single fitted model.

Accordingly, the addition of the CDC dataset substantially broadens the empirical evaluation but does not establish generalizability across all healthcare settings. Validation of further contemporary clinical cohorts with different geographic, demographic, and care delivery characteristics is necessary. In particular, future studies should assess transportability, local recalibration, and decision thresholds in datasets whose variables and outcome definitions better match the intended deployment settings.

The assumed error costs are experimental and should not be interpreted as universal clinical values. Sensitivity analysis demonstrates that the decision threshold and false-positive/false-negative trade-off depend directly on the chosen cost ratio. Clinically appropriate error costs should therefore be established for the intended screening pathway, considering factors such as additional clinical testing after a positive prediction and the consequences of delayed diagnosis. ECE and MCE depend on the chosen binning scheme, and bootstrap confidence intervals quantify sampling variation but do not account for all sources of uncertainty, such as model specification, dataset shift, or clinical implementation. Dataset-level classifier selection from pooled outer-fold outcomes may introduce selection-induced optimism; empirical threshold tuning also reused calibration-fitting prediction pairs. Subgroup analysis is exploratory and should not be interpreted as a formal fairness assessment. Finally, this study is retrospective and based on public datasets; prospective clinical evaluation, local recalibration, and health-economic assessment remain necessary before clinical deployment.

7. Conclusions

In this study, we proposed a leakage-free and risk-aware ML framework for the prediction of diabetes, including training-only preprocessing, probability calibration, cost-sensitive thresholding, statistical uncertainty assessment, subgroup stability analysis, and DCA. In the outer-fold classifier

comparison, LR achieved the highest ROC-AUC on the PIMA dataset (0.8431; 95% CI: 0.8151–0.8695), while GB achieved the highest ROC-AUC on the CDC dataset (0.8298; 95% CI: 0.8276–0.8319). Calibration comparisons did not show a universal winner in the metrics evaluated. On the PIMA dataset, isotonic regression achieved the lowest Brier score (0.1519), while the ensemble achieved the lowest ECE (0.0304). On the CDC dataset, the ensemble obtained the lowest Brier score (0.0967), while isotonic regression achieved the lowest ECE (0.0009).

Under the prespecified 5:1 error-cost ratio, the theoretical threshold $t^* = 0.1667$ produced a recall of 0.9478 on PIMA and 0.7387 on CDC, with 14 and 9,236 false negatives, respectively. Empirical threshold tuning within the outer-training folds produced similar results without using the outer test observations for threshold selection. Sensitivity analysis confirmed that these results depend on the assumed cost ratio, while DCA with bootstrap confidence bands showed a positive net benefit around the threshold corresponding to the primary 5:1 cost scenario in both datasets. In general, the results indicate that the reliability of a diabetes prediction model should not be evaluated only on the basis of discrimination. Leakage prevention ensures methodological validity, calibration evaluates the reliability of predicted probabilities, cost-sensitive thresholding makes decision assumptions explicit, bootstrap intervals quantify statistical uncertainty, subgroup analysis detects variation in model performance across patient groups, and DCA evaluates potential decision utility. Together, these elements allow for a more complete and transparent evaluation of diabetes risk models. However, further validation and clinically justified cost assumptions are required before clinical deployment.

Data availability

The PIMA dataset is a public benchmark described by Smith *et al.* [26]. The CDC Diabetes Health Indicators dataset is publicly available through the UCI Machine Learning Repository [27].

Code availability

The complete Python code used for the reported analysis, together with the generated calibration, cost-sensitivity analysis, DCA, bootstrap, subgroup, and patient-level out-of-fold output files, is available as supplementary material accompanying the revised submission. The analysis uses standard Python scientific-computing packages and fixed random seeds to support reproducibility.

Use of AI tools declaration

The authors declare that OpenAI ChatGPT was used to assist in language editing, manuscript restructuring, and LaTeX formatting. The authors reviewed and revised the generated text and assume full responsibility for the content, analysis, interpretations, and conclusions of the manuscript. No generative-AI tool was used to generate or alter the experimental datasets or numerical results reported in this study.

Conflict of interest

The authors declare that they have no conflicts of interest with respect to this paper.

References

1. Rajkomar A, Oren E, Chen K, et al. (2018) Scalable and accurate deep learning with electronic health records. *npj Digit Med* 1: 18. https://doi.org/10.1038/s41746-018-0029-1
2. Sisodia D and Sisodia DS (2018) Prediction of diabetes using classification algorithms. *Procedia Comput Sci* 132: 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122
3. Kaufman S, Rosset S, Perlich C, et al. (2012) Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data* 6: 1–21. https://doi.org/10.1145/2382577.2382579
4. Kapoor S and Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4: 100804. https://doi.org/10.1016/j.patter.2023.100804
5. Niculescu-Mizil A and Caruana R (2005) Predicting good probabilities with supervised learning, *Proceedings of the International Conference on Machine Learning*, 625–632. https://doi.org/10.1145/1102351.1102430
6. Van Calster B, McLernon DJ, van Smeden M, et al. (2019) Calibration: the Achilles heel of predictive analytics. *BMC Med* 17: 230. https://doi.org/10.1186/s12916-019-1466-7
7. Elkan C (2001) The foundations of cost-sensitive learning, *Proceedings of the International Joint Conference on Artificial Intelligence*, 973–978.
8. Provost F and Fawcett T (1997) Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 43–48.
9. Vickers AJ and Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 26: 565–574. https://doi.org/10.1177/0272989X06295361
10. Vickers AJ, van Calster B, Steyerberg EW (2019) A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 3: 18. https://doi.org/10.1186/s41512-019-0064-7
11. Kuhn M and Johnson K (2013) *Applied Predictive Modeling*. New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3
12. Alturayeif N and Hassine J (2025) Data leakage detection in machine learning code: transfer learning, active learning, or low-shot prompting? *PeerJ Comput Sci* 11: e2730. https://doi.org/10.7717/peerj-cs.2730
13. Rosenblatt M, Tejavibulya L, Jiang R, et al. (2024) Data leakage inflates prediction performance in connectome-based machine learning models. *Nat Commun* 15: 1829. https://doi.org/10.1038/s41467-024-46150-w
14. Amma BNG (2024) En-RfRsK: an ensemble machine learning technique for prognostication of diabetes mellitus. *Egypt Inform J* 25: 100441. https://doi.org/10.1016/j.eij.2024.100441
15. Kawarkhe M and Kaur P (2024) Prediction of diabetes using diverse ensemble learning classifiers. *Procedia Comput Sci* 235: 403–413. https://doi.org/10.1016/j.procs.2024.04.040
16. Jena SK, Behera DK, Jena AK, et al. (2025) A novel ensemble machine learning framework for improved diabetes prediction and complication prevention. *Procedia Comput Sci* 258: 4008–4017. https://doi.org/10.1016/j.procs.2025.04.652

17. Rizal, Azzahra AF, Bandaso RMS, et al. (2025) Type 2 diabetes prediction using recall-optimized XGBoost. *Procedia Comput Sci* 269: 1309–1318. https://doi.org/10.1016/j.procs.2025.09.072
18. Phan LT, Rakkiyappan R, Manavalan B (2025) ReMed-T2D: a robust ensemble learning model for early detection of type 2 diabetes using healthcare dataset. *Comput Biol Med* 187: 109771. https://doi.org/10.1016/j.compbiomed.2025.109771
19. Khan S, Naeem M, Qiyas M (2023) Deep intelligent predictive model for the identification of diabetes. *AIMS Math* 8: 16446–16462. https://doi.org/10.3934/math.2023840
20. Khan S (2026) Biochemical biomarker-driven deep learning framework with SHAP-based feature interpretation for diabetes classification. *Biophys Chem* 334: 107601. https://doi.org/10.1016/j.bpc.2026.107601
21. Almusallam N, Khan S, Alarfaj FK, et al. (2025) A robust deep learning framework for RNA 5-methyluridine modification prediction using integrated features. *BMC Biol* 23: 328. https://doi.org/10.1186/s12915-025-02433-2
22. Khushal R and Fatima U (2024) Fuzzy logic and machine learning for diabetes risk prediction using modifiable factors. *Int J Adv Appl Sci* 11: 225–231. https://doi.org/10.21833/ijaas.2024.12.025
23. Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, In: Smola AJ, Bartlett P, Schölkopf B, et al., *Advances in Large Margin Classifiers*, 61–74. https://doi.org/10.7551/mitpress/1113.003.0008
24. Zadrozny B and Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699. https://doi.org/10.1145/775047.775151
25. Kull M, Silva Filho TM, Flach P (2017) Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. *Proc Mach Learn Res* 54: 623–631. Available from: https://proceedings.mlr.press/v54/kull17a.html.
26. Smith JW, Everhart JE, Dickson WC, et al. (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proc Annu Symp Comput Appl Med Care* 9: 261–265. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2245318/.
27. UCI Machine Learning Repository (2017) CDC Diabetes Health Indicators. Available from: https://doi.org/10.24432/C53919.



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)