

Trustworthy Early Diabetes Risk Screening Using Clinically Informed Probability Fusion with Calibration, Explainability, Decision-Curve, and Subgroup Analysis

Nouh Alhindawi

Department of Software Engineering, Faculty of Information Technology, Jadara University,
Irbid, Jordan

Abstract

Early diabetes-risk models are often reported using single train-test splits, incomplete treatment of physiologically impossible zero values and accuracy alone. This study develops a clinically informed probability-fusion (CI-Fusion) framework for the Pima Indians Diabetes Dataset ($n = 768$). Within each cross-validation fold, impossible zeros were treated as missing, median-imputed and represented by missingness indicators. Elastic-net logistic regression, RBF-SVM, MLP, Extra Trees, HistGradientBoosting, XGBoost, LightGBM and CatBoost were compared with a weighted four-model fusion under repeated stratified five-fold cross-validation. CI-Fusion achieved the highest mean AUROC (0.842, 95% CI 0.825-0.859), 0.765 accuracy and a Brier score of 0.156. Threshold analysis increased sensitivity to 0.810 while retaining 0.740 specificity. SHAP, calibration, decision-curve and subgroup analyses support transparent use as a screening, not diagnostic, tool.

Keywords: diabetes screening; ensemble learning; probability fusion; explainable AI; calibration.

1 Introduction

Diabetes mellitus is a major and expanding public-health challenge. The International Diabetes Federation estimates that 589 million adults aged 20-79 were living with diabetes in 2024 and projects 853 million by 2050, with a substantial share remaining undiagnosed (International Diabetes Federation 2025). Earlier identification of people who may benefit from confirmatory testing, preventive counselling or closer monitoring is therefore important. Machine learning can support this task by integrating routinely collected clinical indicators, but its role should be framed as risk screening rather than autonomous diagnosis.

The Pima Indians Diabetes Dataset is one of the most frequently reused benchmarks in medical machine learning. Its compact size and eight numeric predictors make it suitable for methodological comparison, yet the same properties create important risks. Results can vary substantially with a single train-test split, physiologically impossible zero values are commonly processed inconsistently and complex models can overfit a dataset of only 768 records. Moreover, many studies emphasise accuracy while giving limited attention to probability calibration, clinically meaningful operating thresholds, net benefit or performance variation across subgroups.

Modern boosting systems such as XGBoost, LightGBM and CatBoost can capture nonlinear relationships and interactions in tabular data (Chen and Guestrin 2016; Ke et al. 2017; Prokhorenkova et al. 2018). Nevertheless, no single algorithm is guaranteed to dominate. Linear models can remain competitive and offer stable probability estimates, while diverse tree ensembles may capture complementary structure. Probability-level fusion is therefore attractive when the component models make partly different errors (Wolpert 1992). In healthcare, however, an ensemble is useful only when its evaluation is leakage-safe, transparent and connected to a plausible decision pathway.

This study replaces the draft's conventional nine-model benchmark with a broader and more trustworthy framework. The contributions are fivefold. First, all zero correction, imputation, missingness-indicator generation and scaling are learned inside each training fold. Second, eight modern models spanning regularised linear, kernel, neural, extremely randomised and gradient-boosted families are compared. Third, a clinically informed probability-fusion model (CI-Fusion) combines elastic-net logistic regression, Extra Trees, XGBoost and CatBoost. Fourth, evaluation extends beyond discrimination to calibration, Brier score, decision-curve analysis, threshold trade-offs, SHAP-based feature relevance,

subgroups and ablation. Fifth, paired repeated-cross-validation comparisons are reported without claiming superiority where differences are not statistically supported.

The intended contribution is not a new diagnostic device. It is a reproducible small-data evaluation pattern showing how modern tabular AI can be used more responsibly on a heavily studied benchmark. This distinction is especially important because the dataset represents adult women of Pima heritage and cannot establish performance in men, other ethnic groups or contemporary clinical populations without external validation.

2 Related work and research gap

2.1 Diabetes prediction with machine learning

Earlier work on the Pima data established that comparatively simple algorithms can be competitive. Alzboon et al. (2023) compared logistic regression, decision trees, random forests, k-nearest neighbours, naive Bayes, support-vector machines, gradient boosting and neural networks, finding logistic regression and neural networks strongest across several criteria. Sisodia and Sisodia (2018) likewise showed meaningful but moderate performance from conventional classifiers. These studies are valuable baselines because they reveal that model complexity alone does not guarantee better generalisation on a small clinical table.

Recent research has increasingly used ensembles, resampling and optimisation. Kibria et al. (2022) combined a soft-voting ensemble with SMOTETomek and SHAP and reported balanced accuracy around 90%. Simaiya et al. (2022) proposed a multistage ensemble, although records with implausible zero values were removed, reducing the analysable sample. Oliullah et al. (2023) reported 92.91% accuracy from a stacked ensemble of modern learners after feature engineering. Zhang et al. (2024) combined harmony-search optimisation, feature selection and stacking and reported 93.09% accuracy on the Pima dataset. Hasan et al. (2024) used AutoML and several explanation methods, reporting 85.01% accuracy and demonstrating growing interest in transparent deployment.

These headline values are not directly interchangeable. Some studies use synthetic resampling, feature-selection procedures, data augmentation, local data, simulated observations or optimisation regimes whose separation from evaluation is not always clear. Reza et al. (2024), for example, investigated stacking with Pima, simulated and local healthcare data and reported a cross-validated Pima result close to the traditional mid-to-high 70% range. Such variation illustrates why evaluation design is as influential as algorithm

choice. A result obtained after resampling the entire dataset or selecting features before cross-validation may be optimistic even when the classifier itself is appropriate.

Work by Al-Batah and colleagues provides related evidence for rigorous transformation, multi-model evaluation and explainability across diagnostic and decision-support tasks. Adaptive transformation materially changed clustering behaviour in a diagnostic setting (Al-Batah 2025), while ensemble and explainability components have been integrated in other high-stakes classification domains (Al-Batah and Aldrous 2026). Comparative healthcare frameworks have also emphasised multi-metric evaluation and accessible deployment in heart disease and brain-tumour prediction (Alqaraleh et al. 2026; Alzboon and Al-Batah 2026). These methodological ideas are relevant here, but citations from unrelated application domains are not treated as direct evidence of diabetes performance.

2.2 Trustworthiness, validation and explainability

Clinical prediction research requires more than ranking models by accuracy. TRIPOD+AI emphasises transparent reporting of data, modelling choices, validation and intended use (Collins et al. 2024). Calibration matters because a model can discriminate cases well while systematically overestimating or underestimating risk (Van Calster et al. 2019). Decision-curve analysis complements statistical performance by evaluating net benefit over thresholds at which a person would be referred for further assessment (Vickers and Elkin 2006). Explainability methods such as SHAP can summarise which variables influence model predictions, although they do not establish causal effects (Lundberg and Lee 2017).

The main gap is therefore not simply the absence of another classifier. Few Pima studies combine fold-wise correction of invalid zeros, repeated cross-validation, contemporary boosting models, probability fusion, calibration, clinically oriented threshold analysis, decision curves, explainability, subgroup checks and ablation in one coherent protocol. The present study addresses this gap and explicitly distinguishes exploratory benchmark evidence from clinical validation.

3 Materials and methods

3.1 Data source and cohort

A secondary analysis was conducted using the public Pima Indians Diabetes Database, originally assembled by the National Institute of Diabetes and Digestive and Kidney Diseases and described by Smith et al. (1988). The supplied comma-separated file contained 768

records, eight predictors and a binary Outcome variable. Outcome 1 denotes diabetes and Outcome 0 denotes no diabetes. There were 268 positive cases and 500 negative cases, giving a prevalence of 34.9%.

The predictors were number of pregnancies, two-hour plasma glucose, diastolic blood pressure, triceps skin-fold thickness, two-hour serum insulin, body mass index (BMI), diabetes pedigree function and age. Table 1 reports their clinical meaning, distribution and invalid-zero burden. Representative raw observations included a non-diabetic profile with glucose 85 mg/dL, BMI 26.6 kg/m² and age 31, and a diabetic profile with glucose 148 mg/dL, BMI 33.6 kg/m² and age 50. Because the dataset is tabular rather than image-based, no pixel-level feature extraction or clinical-image sampling was applicable; Figure 2 provides distributional data visualisations instead.

3.2 Leakage-safe preprocessing and feature representation

Zero is physiologically implausible for glucose, blood pressure, skin thickness, insulin and BMI in this context. These values were therefore converted to missing values. The operation was performed separately inside each training fold. Median values were learned from the training portion only, and binary missingness indicators were added so that models could retain information about measurement absence. This design prevents held-out data from influencing imputation and allows missingness itself to contribute when informative.

Elastic-net logistic regression used robust scaling, while RBF-SVM and MLP used standard scaling; tree-based learners used unscaled values. Scaling parameters were also fitted only on training folds. No synthetic oversampling was used, preserving the observed prevalence and probability scale. Class weighting was used for logistic regression, SVM and Extra Trees to reduce dominance by the non-diabetic class. Six clinically plausible interaction terms were explored in an ablation analysis, including glucose-BMI and glucose-age products. Because the original eight predictors yielded slightly better validation performance, the final framework retained the original predictors plus fold-generated missingness indicators.

3.3 Candidate models and proposed CI-Fusion

Eight candidate learners were implemented. Elastic-net logistic regression used $C = 0.7$, l_1 ratio = 0.2 and balanced class weights (Zou and Hastie 2005). RBF-SVM used $C = 1.5$, $\gamma = \text{'scale'}$, probability estimation and balanced weights (Cortes and Vapnik 1995). The MLP contained two hidden layers of 32 and 16 units, L2 regularisation $\alpha = 0.01$, learning rate 0.002 and early stopping. Extra Trees used 350 trees, square-root feature sampling and

minimum leaf size 2 (Geurts et al. 2006). HistGradientBoosting used 220 iterations, learning rate 0.04, 15 maximum leaf nodes, minimum leaf size 20 and L2 regularisation 1.0.

XGBoost used 280 trees, maximum depth 3, learning rate 0.035, subsampling and column sampling of 0.85, minimum child weight 3, $L2 = 5$ and $L1 = 0.1$. LightGBM used 280 trees, learning rate 0.035, 15 leaves, minimum child size 20, row and column sampling of 0.85, $L2 = 5$ and $L1 = 0.1$. CatBoost used 280 iterations, depth 5, learning rate 0.035 and L2 leaf regularisation 5. Settings were prespecified as conservative small-data configurations rather than optimised against the held-out folds.

CI-Fusion combines probability outputs from four structurally different components: elastic-net logistic regression, Extra Trees, XGBoost and CatBoost. The final probability is $p(\text{CI-Fusion}) = [1.2 p(\text{LR}) + 1.4 p(\text{ET}) + 1.5 p(\text{XGB}) + 1.5 p(\text{CAT})] / 5.6$. The weights were prespecified to preserve a substantial linear contribution while giving slightly greater influence to boosted tree models. No fold-specific weight search was performed. Fusion was conducted at probability level because calibrated risk ranking and threshold adjustment are more useful for screening than majority voting on hard labels.

3.4 Validation, statistical analysis and trustworthy-AI assessment

The primary comparison used repeated stratified five-fold cross-validation with three repeats, producing 15 paired held-out test folds. Performance was summarised by AUROC, precision-recall AUC, accuracy, balanced accuracy, precision, sensitivity, specificity, F1, Matthews correlation coefficient, log loss and Brier score. Fold-level 95% confidence intervals were calculated as the mean plus or minus 1.96 standard errors. Because repeated cross-validation folds are correlated, these intervals and tests are interpreted as comparative evidence rather than independent-sample inference.

The CI-Fusion AUROC was compared with each candidate using paired two-sided Wilcoxon signed-rank tests across folds, followed by Holm correction. A separate shuffled stratified ten-fold procedure generated one out-of-fold probability for every record. These pooled predictions were used for ROC curves, calibration summaries, the confusion matrix, threshold analysis, decision-curve analysis and exploratory subgroups.

Three operating points were evaluated: the conventional 0.50 threshold, the Youden threshold maximising sensitivity plus specificity minus one and the most specific grid threshold that retained sensitivity of at least 0.80. Calibration was summarised by Brier score and by regressing outcome on the logit of predicted probability to obtain calibration intercept

and slope. Decision-curve net benefit was calculated from referral thresholds 0.05-0.60, with a positive prediction interpreted as referral for confirmatory testing or clinical review.

Global feature relevance was estimated from absolute SHAP values of a CatBoost component fitted to the complete dataset after the same zero correction and median imputation. Missingness-indicator contributions were aggregated with their parent variable. Age subgroups (21-29, 30-49 and 50+) and BMI subgroups (<30 and ≥ 30 kg/m²) were evaluated using out-of-fold predictions. These analyses are descriptive and were not powered as formal fairness tests. An ablation compared the final original-feature pipeline with engineered interactions and with a pipeline that incorrectly treated implausible zeros as valid.

Analyses were performed in Python 3.13.5 using pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0 (Pedregosa et al. 2011), XGBoost 3.1.3, LightGBM 4.6.0, CatBoost 1.2.8, SciPy 1.17.0, statsmodels 0.14.6 and SHAP 0.50.0. Random states were fixed for reproducibility.

4 Results

4.1 Data-quality and descriptive findings

The dataset contained substantially more invalid zeros for insulin (374 records, 48.7%) and skin thickness (227, 29.6%) than for blood pressure (35), BMI (11) or glucose (5). This pattern justified both median imputation and explicit missingness indicators. Among valid values, the median glucose was 117 mg/dL (interquartile range 99-141), median BMI was 32.3 kg/m² (27.5-36.6) and median age was 29 years (24-41). Figure 2 shows that diabetic records tended to occupy higher glucose and BMI ranges, while age distributions were right-skewed in both classes.

The data audit also illustrates why zero handling cannot be treated as a minor cosmetic step. Almost half of insulin values would be interpreted as genuine zero measurements if the raw table were used without clinical validity checks. At the same time, deleting every record containing an invalid zero would discard much of the cohort. Fold-wise imputation with missingness indicators therefore provided a pragmatic compromise for this benchmark.

4.2 Comparative model performance

CI-Fusion achieved the highest mean AUROC, 0.842 (95% CI 0.825-0.859), and mean precision-recall AUC, 0.730. Mean accuracy was 0.765, sensitivity 0.644, specificity 0.829, F1 0.655 and Brier score 0.156. CatBoost was the strongest single model by mean AUROC (0.838), followed closely by elastic-net logistic regression (0.836), Extra Trees (0.835) and

XGBoost (0.833). The narrow separation among these models shows that the fusion gain was incremental rather than transformative.

Paired tests confirmed this interpretation. After Holm correction, CI-Fusion had higher fold-level AUROC than RBF-SVM (adjusted $p = 0.0009$), MLP ($p = 0.0005$), HistGradientBoosting ($p = 0.0018$) and LightGBM ($p = 0.0076$). Differences from elastic-net logistic regression, Extra Trees, XGBoost and CatBoost were not statistically significant. The MLP was unstable, with mean AUROC 0.547 and a wide interval, indicating that this small dataset did not support the selected neural architecture reliably.

Ten-fold pooled out-of-fold analysis yielded AUROC 0.839, precision-recall AUC 0.709, accuracy 0.757, sensitivity 0.631, specificity 0.824 and Brier score 0.156. At threshold 0.50, the model correctly classified 412 of 500 non-diabetic records and 169 of 268 diabetic records, with 88 false positives and 99 false negatives. Table 2 and Figure 3 provide the complete comparison.

4.3 Thresholds, calibration and clinical utility

The conventional threshold prioritised specificity over sensitivity. The Youden threshold of 0.377 increased sensitivity from 0.631 to 0.836, with specificity 0.718, accuracy 0.759 and F1 0.708. A threshold of 0.395 was selected to maintain sensitivity of at least 0.80; it produced sensitivity 0.810, specificity 0.740, accuracy 0.764 and F1 0.706. This operating point may be more suitable for an initial screening pathway where missed high-risk cases are costlier than additional confirmatory tests.

Calibration was reasonable but not perfect. The pooled calibration intercept was -0.197 and the slope was 1.212, indicating mild average overprediction and probabilities that were slightly too moderate. Decision-curve analysis showed positive net benefit relative to referring nobody and generally higher net benefit than referring everybody across much of the 0.05-0.60 threshold range. These results support clinical review as the action following a positive screen, not treatment based solely on the model.

4.4 Explainability, subgroups and ablation

The CatBoost SHAP analysis ranked glucose as the most influential variable, accounting for 30.5% of aggregated absolute relevance, followed by BMI (21.0%), age (14.0%), diabetes pedigree function (10.7%), pregnancies (7.5%), insulin (7.0%), skin thickness (5.9%) and blood pressure (3.4%). This ranking is clinically plausible and consistent with prior diabetes-prediction studies, but it should be read as model attribution rather than causal importance.

Subgroup results revealed materially different operating characteristics at threshold 0.50. For ages 21-29, AUROC was 0.846 and specificity 0.926, but sensitivity was only 0.440 because prevalence was low. For ages 50+, sensitivity rose to 0.814 while specificity fell to 0.587. The BMI <30 subgroup had AUROC 0.838 and specificity 0.924 but sensitivity 0.383, whereas BMI $\geq$ 30 had sensitivity 0.685 and specificity 0.727. These differences reinforce the need for threshold selection, calibration and monitoring in the intended population.

Ablation did not support retaining engineered interactions. Across three stratified folds, the original-feature pipeline achieved mean AUROC 0.848 and accuracy 0.770, compared with AUROC 0.846 and accuracy 0.763 when interactions were added. Treating invalid zeros as genuine values reduced mean AUROC to 0.840 and accuracy to 0.759. The final model therefore favoured a simpler representation with clinically informed missing-value handling.

The four fusion components were highly but not perfectly correlated. Pairwise probability correlations ranged from 0.895 to 0.980 and hard-prediction disagreement ranged from 5.6% to 15.1%. This residual diversity explains why averaging could modestly improve mean AUROC and Brier score, while the high correlations also explain why improvements over the strongest single models were small.

5 Discussion

5.1 Principal findings

The central finding is that a restrained probability fusion can improve average discrimination without erasing the value of simpler models. CI-Fusion achieved the highest mean AUROC and competitive calibration, but it did not significantly outperform the four strongest components. This is scientifically more informative than declaring a universal winner: on a small tabular dataset, regularised logistic regression, Extra Trees and modern boosting systems formed a statistically similar leading group.

The more important contribution lies in the evaluation architecture. Fold-wise handling of invalid zeros prevented preprocessing leakage, repeated cross-validation reduced dependence on a fortunate split and out-of-fold predictions enabled calibration, decision curves and threshold analysis. The ablation showed that clinically informed data cleaning added value whereas extra interaction features did not. This result cautions against equating novelty with additional complexity.

5.2 Comparison with previous studies

The observed accuracy of approximately 76% is close to traditional Pima benchmarks and to the Pima-only cross-validation result reported by Reza et al. (2024). It is lower than the 85-93% values reported by several optimised, AutoML or stacked systems (Hasan et al. 2024; Oliullah et al. 2023; Zhang et al. 2024). Table 4 makes clear that these studies differ in resampling, feature engineering, optimisation, dataset combinations and validation design. A lower but leakage-resistant estimate can be more credible than a higher estimate obtained under a less transparent protocol.

The feature ranking also agrees with prior literature in identifying glucose, BMI and age as major predictive variables. In contrast to studies that remove zero-valued rows or create large synthetic cohorts, the present analysis retained all 768 records and modelled missing measurements explicitly. The study further adds calibration, net-benefit, threshold and subgroup evidence that is often absent from accuracy-centred comparisons.

5.3 Social and practical implications

A screening model can alter access to care, referral volume and the distribution of false reassurance. The threshold should therefore be chosen by the service context rather than by software default. In a resource-constrained primary-care programme, a sensitivity-oriented threshold may help identify more people for low-cost confirmatory testing, but it also creates additional appointments and anxiety. Decision-curve analysis makes this trade-off explicit and allows planners to connect model output to service capacity.

Population specificity is equally important. The dataset includes only women of Pima heritage and reflects a historical data-collection setting. Deploying the model in Jordan, Europe or another community without recalibration could reproduce inequities or miss local risk patterns. Subgroup differences observed even within this small sample show that a single threshold can behave differently across age and BMI groups. Responsible use would require local validation, transparent communication, clinician oversight, periodic auditing and an appeal route when model output conflicts with clinical judgement.

5.4 Strengths and limitations

Strengths include a leakage-safe pipeline, contemporary model library, repeated paired validation, probability fusion, multi-metric reporting, threshold analysis, calibration, decision curves, SHAP, subgroup checks, ablation and complete hyperparameter disclosure. The

analysis also avoids selective reporting: the unstable MLP and non-significant comparisons with strong baselines are retained.

Limitations are substantial. The sample is small, historical and demographically narrow, and no external cohort was available. Several predictors contain extensive missingness encoded as zero, and important variables such as HbA1c, medication, socioeconomic context and longitudinal history are absent. Hyperparameters and fusion weights were prespecified rather than selected in a fully nested procedure. The threshold analyses reused out-of-fold predictions and should be confirmed prospectively. Repeated-cross-validation folds are not independent, subgroup estimates are imprecise and SHAP relevance is descriptive rather than causal. Finally, a public benchmark cannot substitute for assessment of workflow impact, patient acceptability, fairness and safety.

Future work should evaluate the framework on multi-centre, contemporary cohorts with men and diverse ethnic and socioeconomic groups. Model recalibration and threshold selection should be performed separately at each site, with prospective decision-impact evaluation. Nested optimisation, uncertainty quantification, patient-level explanations and open analytical code would further strengthen reproducibility.

6 Conclusion

A clinically informed probability fusion of elastic-net logistic regression, Extra Trees, XGBoost and CatBoost achieved the highest mean AUROC in a leakage-safe comparison of modern models on the Pima Indians Diabetes Dataset. Its advantage over the strongest individual models was modest and not statistically significant, but it delivered useful probability estimates for threshold adjustment and decision analysis. Glucose, BMI and age were the leading predictive features, while subgroup and ablation results demonstrated that data quality and operating thresholds matter as much as algorithm selection. The framework should be considered a transparent benchmark for early-risk screening research, not a validated diagnostic system. External, prospective and population-specific validation is essential before clinical use.

Acknowledgement

The authors are grateful to the Deanship of Scientific Research at Jadara University for providing financial support for this publication.

Sources

National Institute of Diabetes and Digestive and Kidney Diseases / UCI Machine Learning Repository: Pima Indians Diabetes Database. Public mirror available from: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database> (Accessed 22.7.2026).

References

- Al-Batah, Mohammad Subhi (2025): Adaptive Data Transformation for Enhanced Clustering Performance in Diagnostic Systems. *Journal of Computer Science*, 21 (9): 2074-2080. <https://doi.org/10.3844/jcssp.2025.2074.2080>.
- Al-Batah, Mohammad Subhi, and Aldrous, Taqieddin (2026): Ensemble Learning-Based Hybrid Explainable Intrusion Detection System With SMOTE for Enhanced Detection and Interpretability in Cybersecurity. *Journal of Computer Science*, 22 (5): 1569-1576. <https://doi.org/10.3844/jcssp.2026.1569.1576>.
- Alqaraleh, Muhyeeddin, Al-Batah, Mohammad Subhi, Alzboon, Mowafaq Salem, and Alourani, Abdullah (2026): Brain Tumor Detection with Real-World Predictions in Jordan Hospitals. *Scientific Reports*, 16: 3321. <https://doi.org/10.1038/s41598-025-33215-z>.
- Alzboon, Mowafaq Salem, and Al-Batah, Mohammad Subhi (2026): A Clinically Accessible Heart Disease Prediction Framework: Multi-Model Evaluation with Ensemble Learning and Low-Code Deployment. *Discover Applied Sciences*, 8: 339. <https://doi.org/10.1007/s42452-026-08353-2>.
- Alzboon, Mowafaq Salem, Al-Batah, Mohammad Subhi, Alqaraleh, Muhyeeddin, Abuashour, Ahmad, and Bader, Ahmad Fuad Hamadah (2023): Early Diagnosis of Diabetes: A Comparison of Machine Learning Methods. *International Journal of Online and Biomedical Engineering*, 19 (15): 144-165. <https://doi.org/10.3991/ijoe.v19i15.42417>.
- Geurts, Pierre, Ernst, Damien, and Wehenkel, Louis (2006): Extremely Randomized Trees. *Machine Learning*, 63: 3-42. <https://doi.org/10.1007/s10994-006-6226-1>.
- Chen, Tianqi, and Guestrin, Carlos (2016): XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 785-794. <https://doi.org/10.1145/2939672.2939785>.
- Collins, Gary S., et al. (2024): TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ*, 385: e078378. <https://doi.org/10.1136/bmj-2023-078378>.

- Cortes, Corinna, and Vapnik, Vladimir (1995): Support-Vector Networks. *Machine Learning*, 20: 273-297. <https://doi.org/10.1007/BF00994018>.
- Hasan, Raza, Dattana, Vishal, Mahmood, Salman, and Hussain, Saqib (2024): Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights. *Information*, 16 (1): 7. <https://doi.org/10.3390/info16010007>.
- International Diabetes Federation (2025): IDF Diabetes Atlas. 11th edition. Brussels: International Diabetes Federation. Available from: <https://diabetesatlas.org/> (Accessed 22.7.2026).
- Ke, Guolin, et al. (2017): LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30: 3146-3154.
- Kibria, Hafsa Binte, Nahiduzzaman, Md, Goni, Md Omaer Faruq, Ahsan, Mominul, and Haider, Julfikar (2022): An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. *Sensors*, 22 (19): 7268. <https://doi.org/10.3390/s22197268>.
- Lundberg, Scott M., and Lee, Su-In (2017): A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30: 4765-4774.
- Oliullah, Khondokar, Rasel, Mahedi Hasan, Islam, Md Manzurul, Islam, Md Reazul, Wadud, Md Anwar Hussen, and Whaiduzzaman, Md (2023): A Stacked Ensemble Machine Learning Approach for the Prediction of Diabetes. *Journal of Diabetes and Metabolic Disorders*, 23 (1): 603-617. <https://doi.org/10.1007/s40200-023-01321-2>.
- Pedregosa, Fabian, et al. (2011): Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825-2830.
- Prokhorenkova, Liudmila, Gusev, Gleb, Vorobev, Aleksandr, Dorogush, Anna Veronika, and Gulin, Andrey (2018): CatBoost: Unbiased Boosting with Categorical Features. *Advances in Neural Information Processing Systems*, 31: 6638-6648.
- Reza, Md Shamim, Amin, Ruhul, Yasmin, Rubia, Kulsum, Woomme, and Ruhi, Sabba (2024): Improving Diabetes Disease Patients Classification Using Stacking Ensemble Method with PIMA and Local Healthcare Data. *Heliyon*, 10 (2): e24536. <https://doi.org/10.1016/j.heliyon.2024.e24536>.
- Simaiya, Sarita, et al. (2022): A Novel Multistage Ensemble Approach for Prediction and Classification of Diabetes. *Frontiers in Physiology*, 13: 1085240. <https://doi.org/10.3389/fphys.2022.1085240>.
- Sisodia, Deepti, and Sisodia, Dilip Singh (2018): Prediction of Diabetes Using Classification Algorithms. *Procedia Computer Science*, 132: 1578-1585. <https://doi.org/10.1016/j.procs.2018.05.122>.
- Smith, Jack W., Everhart, James E., Dickson, W. C., Knowler, William C., and Johannes, Robert S. (1988): Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*: 261-265.

- Van Calster, Ben, McLernon, David J., van Smeden, Maarten, Wynants, Laure, and Steyerberg, Ewout W. (2019): Calibration: The Achilles Heel of Predictive Analytics. *BMC Medicine*, 17: 230. <https://doi.org/10.1186/s12916-019-1466-7>.
- Vickers, Andrew J., and Elkin, Elena B. (2006): Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, 26 (6): 565-574. <https://doi.org/10.1177/0272989X06295361>.
- Wolpert, David H. (1992): Stacked Generalization. *Neural Networks*, 5 (2): 241-259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
- Zhang, Zaiheng, et al. (2024): A Novel Evolutionary Ensemble Prediction Model Using Harmony Search and Stacking for Diabetes Diagnosis. *Journal of King Saud University - Computer and Information Sciences*, 36 (1): 101873. <https://doi.org/10.1016/j.jksuci.2023.101873>.
- Zou, Hui, and Hastie, Trevor (2005): Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B*, 67 (2): 301-320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.
- Al-Batah, Mohammad Subhi, and Mostafa, Hussam Eyad (2026): Fashion Analytics with Feature-Ranked Machine Learning for Accurate Clothing Image Classification. *Social Science Forum*.
- Al-Batah, Mohammad Subhi (2026): Interdisciplinary Machine Learning Approach for COVID-19 Detection from Chest X-Rays Images. *Kronika Journal*, 26 (5): 325-332. <https://doi.org/20.14148/kkj/v26.5223>.
- Al-Batah, Mohammad Subhi (2026): AI-Driven Multiclass Fruit Image Classification for Smart Nutritional Monitoring and Dietary Assessment. *Geoscience Journal*, 7 (5): 1-8. <https://doi.org/10.5281/zenodo.20257017>.
- Al-Batah, Mohammad Subhi (2026): Intelligent Prediction of E-Commerce Order Arrival Times Using Feature-Ranked Machine Learning Models. *Journal of Taiyuan University of Technology*, 8 (5): 86-96. <https://zenodo.org/records/20325076>.
- Al-Batah, Mohammad Subhi (2026): Predicting Term Deposit Subscriptions Using Machine Learning Models: An Evaluation of Performance on the Bank Marketing Dataset. *Liberte Journal*, 14 (5): 900-912. <https://doi.org/20.14118.LRJ.2026.V14I5.485926>.
- Al-Batah, Mohammad Subhi, Al-Eiadeh, Mohammad R., and Alnsour, Yazan (2025): A Novel Approach for Enhancing Plant Leaf Classification with the Binary Cuckoo Search Algorithm. *Applied Computational Intelligence and Soft Computing*, 2025 (1): 7696962. <https://doi.org/10.1155/acis/7696962>.
- Al-Batah, Mohammad Subhi, and Alzboon, Mowafaq Salem (2026): Sea Animal Image Classification Using Machine Learning Algorithms for Accurate and Scalable Prediction. *Discover Artificial Intelligence*, 6: 332. <https://doi.org/10.1007/s44163-026-01005-9>.
- Al-Batah, Mohammad Subhi, Alzboon, Mowafaq Salem, and Zureigat, H. (2026): Predictive Mathematical Modeling and Classification of Retail Sales Orders Using AI Machine Learning Techniques. *Journal of Computer Science*, 22 (3): 878-885. <https://doi.org/10.3844/jcssp.2026.878.885>.

Tables and figures

Table 1. Dataset variables, validity checks and descriptive statistics.

Predictor	Clinical meaning	Median (IQR)	Observed range*	Invalid zeros
Pregnancies	Number of pregnancies	3.00 (1.00-6.00)	0.00-17.00	0
Glucose	Two-hour plasma glucose (mg/dL)	117.00 (99.00-141.00)	44.00-199.00	5
BloodPressure	Diastolic blood pressure (mm Hg)	72.00 (64.00-80.00)	24.00-122.00	35
SkinThickness	Triceps skin-fold thickness (mm)	29.00 (22.00-36.00)	7.00-99.00	227
Insulin	Two-hour serum insulin (microU/mL)	125.00 (76.25-190.00)	14.00-846.00	374
BMI	Body mass index (kg/m2)	32.30 (27.50-36.60)	18.20-67.10	11
DiabetesPedigreeFunction	Family-history-related pedigree score	0.37 (0.24-0.63)	0.08-2.42	0
Age	Age in years	29.00 (24.00-41.00)	21.00-81.00	0

Note. *For variables with physiologically implausible zero values, the reported range excludes zero. Outcome distribution: 500 non-diabetic and 268 diabetic records. Representative raw profiles are described in Section 3.1.

Table 2. Repeated stratified cross-validation performance.

Model	AUROC (95% CI)	PR-AUC	Accuracy	Sensitivity	Specificity	F1	Brier
Proposed CI-Fusion	0.842 (0.825-0.859)	0.730	0.765	0.644	0.829	0.655	0.156
CatBoost	0.838 (0.819-0.856)	0.727	0.763	0.591	0.855	0.634	0.157
Elastic-net LR	0.836 (0.817-0.855)	0.717	0.745	0.711	0.763	0.660	0.168
Extra Trees	0.835 (0.817-0.852)	0.724	0.754	0.700	0.783	0.665	0.161
XGBoost	0.833 (0.815-0.850)	0.719	0.762	0.612	0.842	0.641	0.159
LightGBM	0.823 (0.806-0.841)	0.702	0.758	0.613	0.835	0.638	0.167
RBF SVM	0.823 (0.804-0.841)	0.707	0.742	0.572	0.833	0.607	0.163
HistGradientBoosting	0.819 (0.802-0.836)	0.692	0.752	0.611	0.827	0.631	0.175
MLP	0.547 (0.457-0.637)	0.439	0.673	0.159	0.949	0.188	0.229

Note. Values are means across 15 paired held-out folds. CI-Fusion was significantly different from RBF SVM (Holm p=0.0009), MLP (Holm p=0.0005), HistGradientBoosting (Holm p=0.0018), LightGBM (Holm p=0.0076). Differences from Elastic-net LR, Extra Trees, XGBoost, CatBoost were not significant after Holm correction.

Table 3. Operating thresholds, subgroup performance and ablation analyses.

Panel A. Threshold-dependent performance.

Operating point	Threshold	Accuracy	Sensitivity	Specificity	F1	MCC
Default	0.500	0.757	0.631	0.824	0.644	0.459
Youden	0.377	0.759	0.836	0.718	0.708	0.529
Sensitivity ≥ 0.80	0.395	0.764	0.810	0.740	0.706	0.526

Panel B. Exploratory subgroup performance at threshold 0.50.

Subgroup	n	Prevalence	AUROC	Accuracy	Sensitivity	Specificity
Age 21-29	396	0.212	0.846	0.823	0.440	0.926
Age 30-49	283	0.498	0.779	0.682	0.688	0.676
Age 50+	89	0.483	0.763	0.697	0.814	0.587
BMI <30	285	0.165	0.838	0.835	0.383	0.924
BMI ≥ 30	472	0.464	0.803	0.708	0.685	0.727

Panel C. Three-fold ablation summary.

Configuration	AUROC	Accuracy	Sensitivity	Specificity	Brier
Full pipeline	0.846	0.763	0.668	0.814	0.155
Original features (final)	0.848	0.770	0.690	0.812	0.156
Zeros treated as valid	0.840	0.759	0.653	0.816	0.156

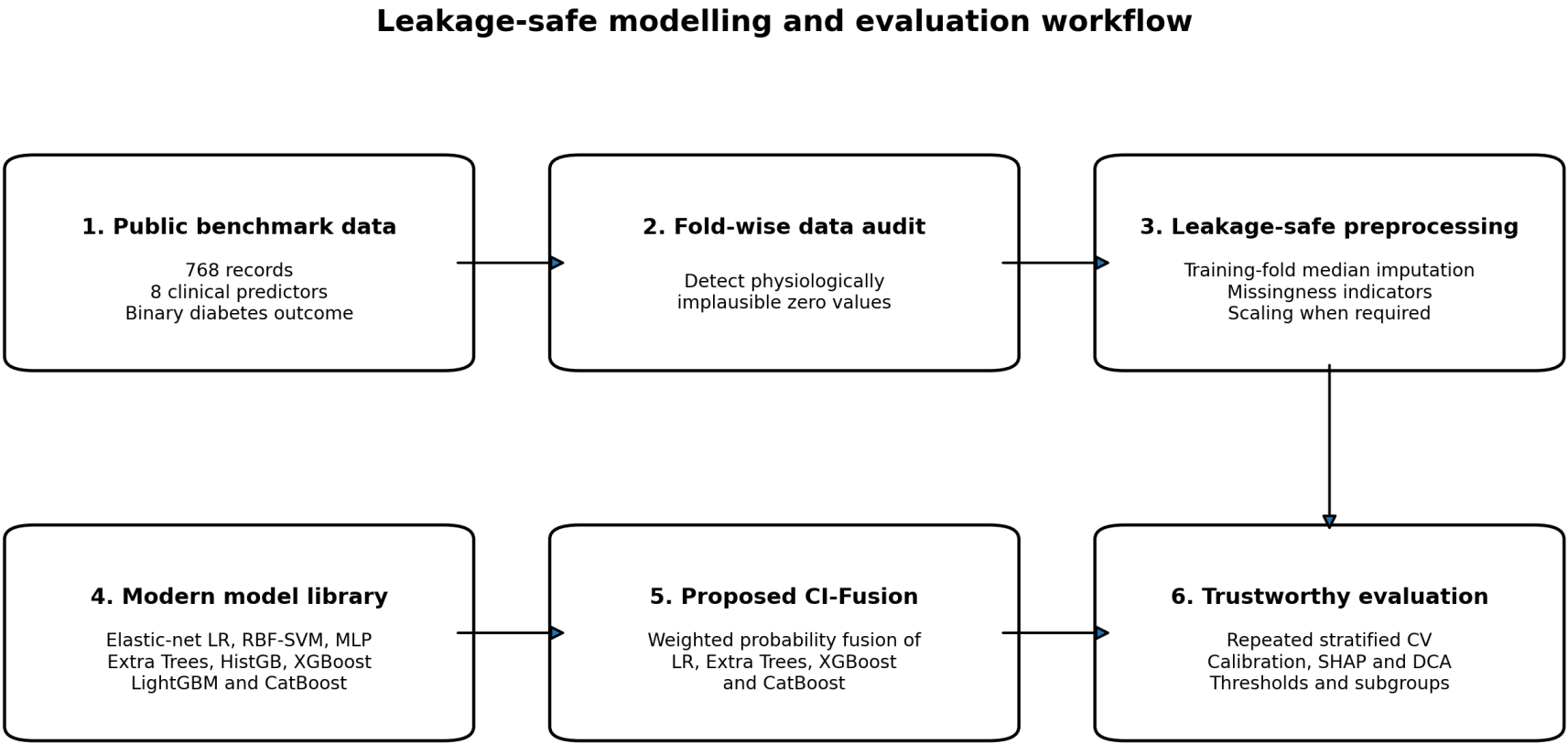
Note. Threshold and subgroup metrics use pooled ten-fold out-of-fold probabilities. Subgroup estimates are exploratory. The ablation used three stratified folds and is intended to assess direction rather than provide definitive inference.

Table 4. Comparison with selected diabetes-prediction studies.

Study	Data / validation	Approach	Reported result	Trustworthiness features	Comparability note
Alzboon et al. (2023)	Pima, 768; comparative evaluation	Classical and ensemble ML	LR and NN strongest across several criteria	Broad model comparison	Closest classical benchmark; protocol differs
Kibria et al. (2022)	Pima; five-fold CV with resampling	Soft voting, SMOTETomek, SHAP	Balanced accuracy about 90%	Explainability	Synthetic resampling changes class distribution
Oliullah et al. (2023)	Pima; grid search and feature engineering	Stacked modern ensembles	Accuracy 92.91%	SHAP	Optimisation and feature engineering limit direct comparison
Reza et al. (2024)	Pima plus simulated and local data	Classical and DNN stacking	Pima CV accuracy reported near 77.1%	Additional local dataset	Pima-only result is more comparable
Zhang et al. (2024)	Pima and CWMD	Harmony search, feature selection, stacking	Pima accuracy 93.09%	Two datasets, optimisation	Extensive optimisation; validation design differs
Hasan et al. (2024)	Pima and two generalisation datasets	AutoGluon AutoML plus XAI	Pima ensemble accuracy 85.01%	Multiple XAI methods and app	AutoML search and external datasets
Present study	Pima, 768; repeated 5-fold x3 plus 10-fold OOF	Leakage-safe CI-Fusion	AUROC 0.842; accuracy 0.765; Brier 0.156	Calibration, DCA, SHAP, thresholds, subgroups, ablation	No resampling or external cohort; conservative estimate

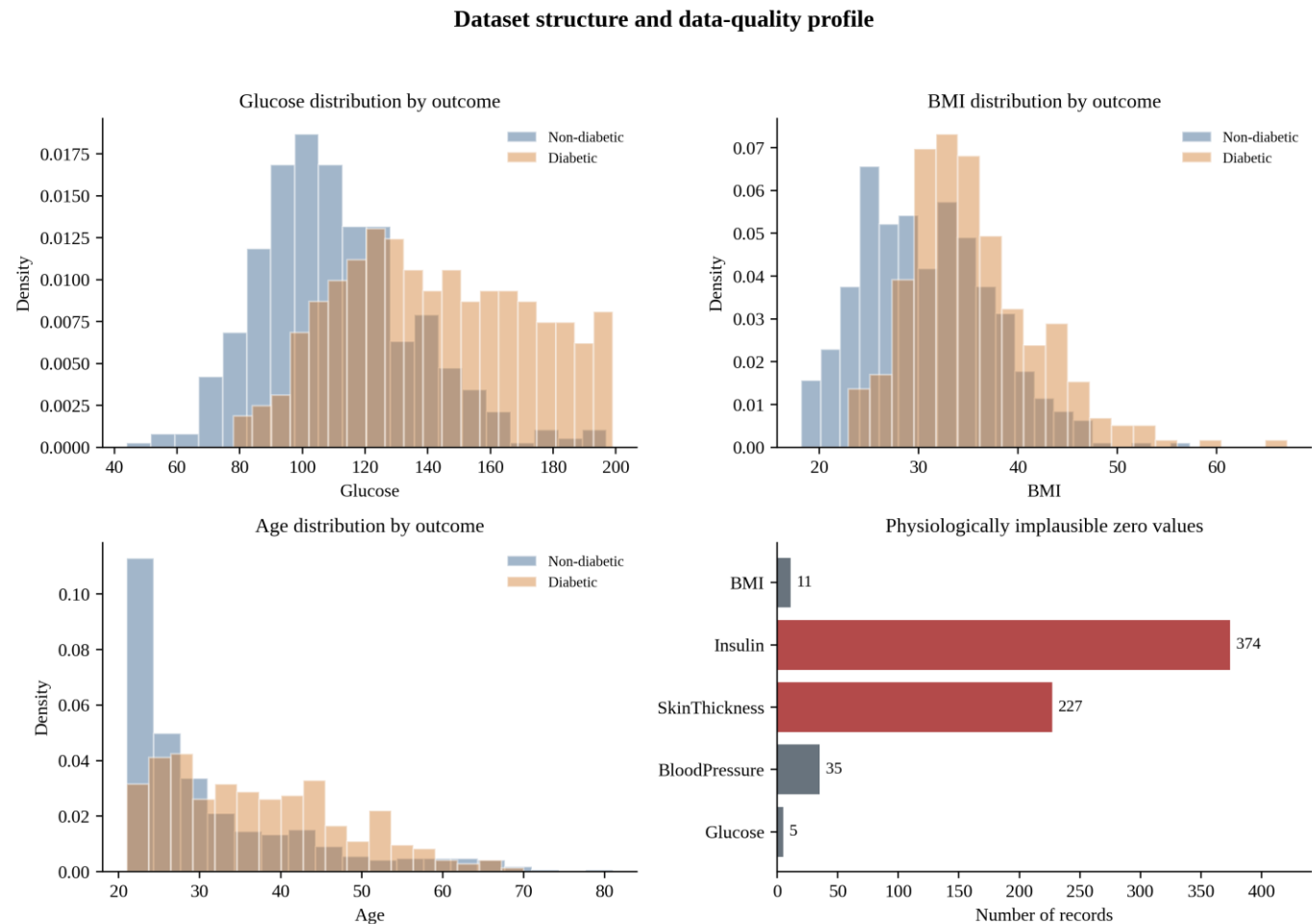
Note. Results are reported as stated by the cited authors and should not be interpreted as a league table. Differences in resampling, optimisation, feature engineering, data sources and validation can materially alter reported performance.

Figure 1. Leakage-safe modelling and evaluation workflow.



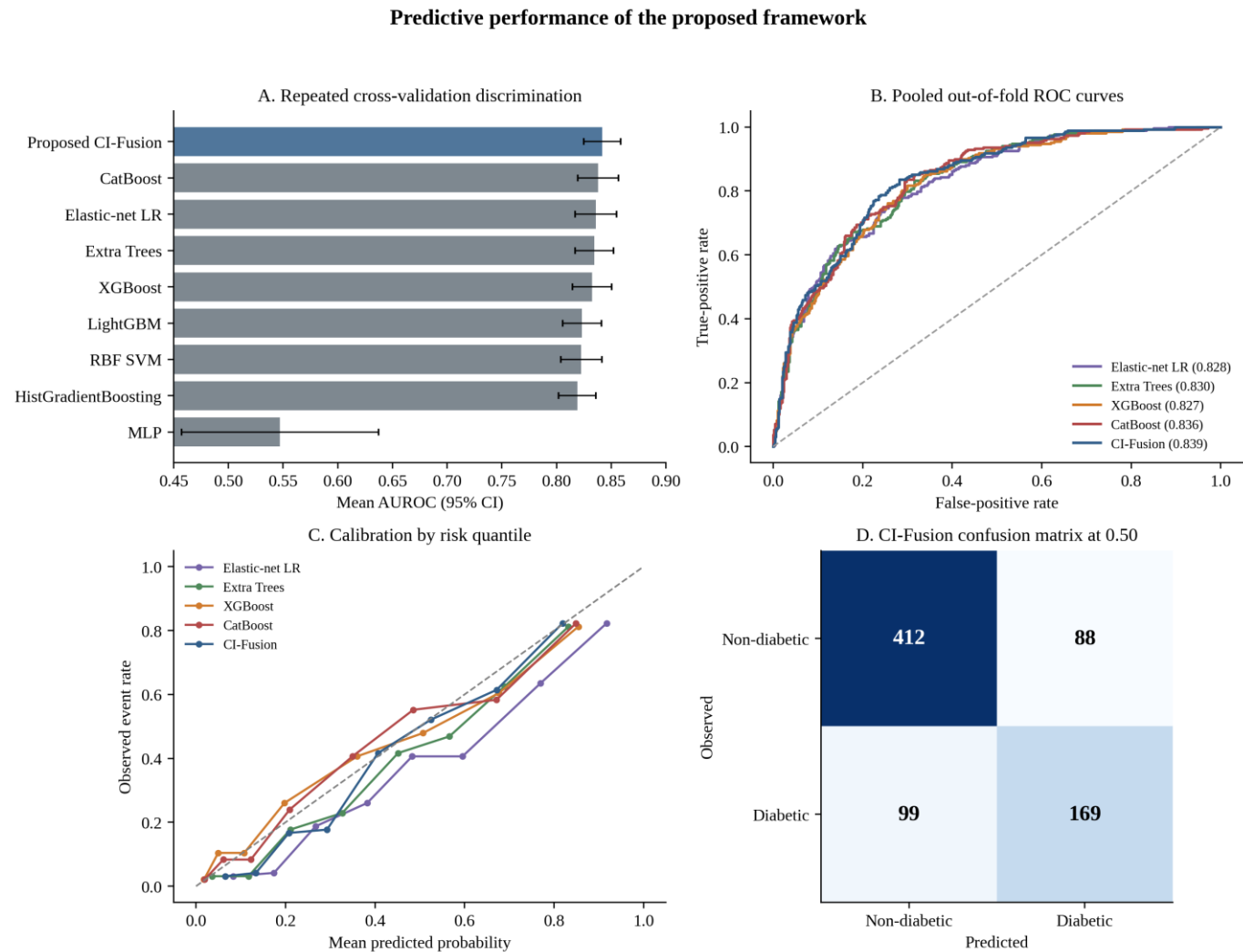
Note. The workflow separates every learned preprocessing operation from the held-out fold and combines discrimination, calibration, clinical-utility, threshold and subgroup analyses.

Figure 2. Dataset structure and data-quality profile.



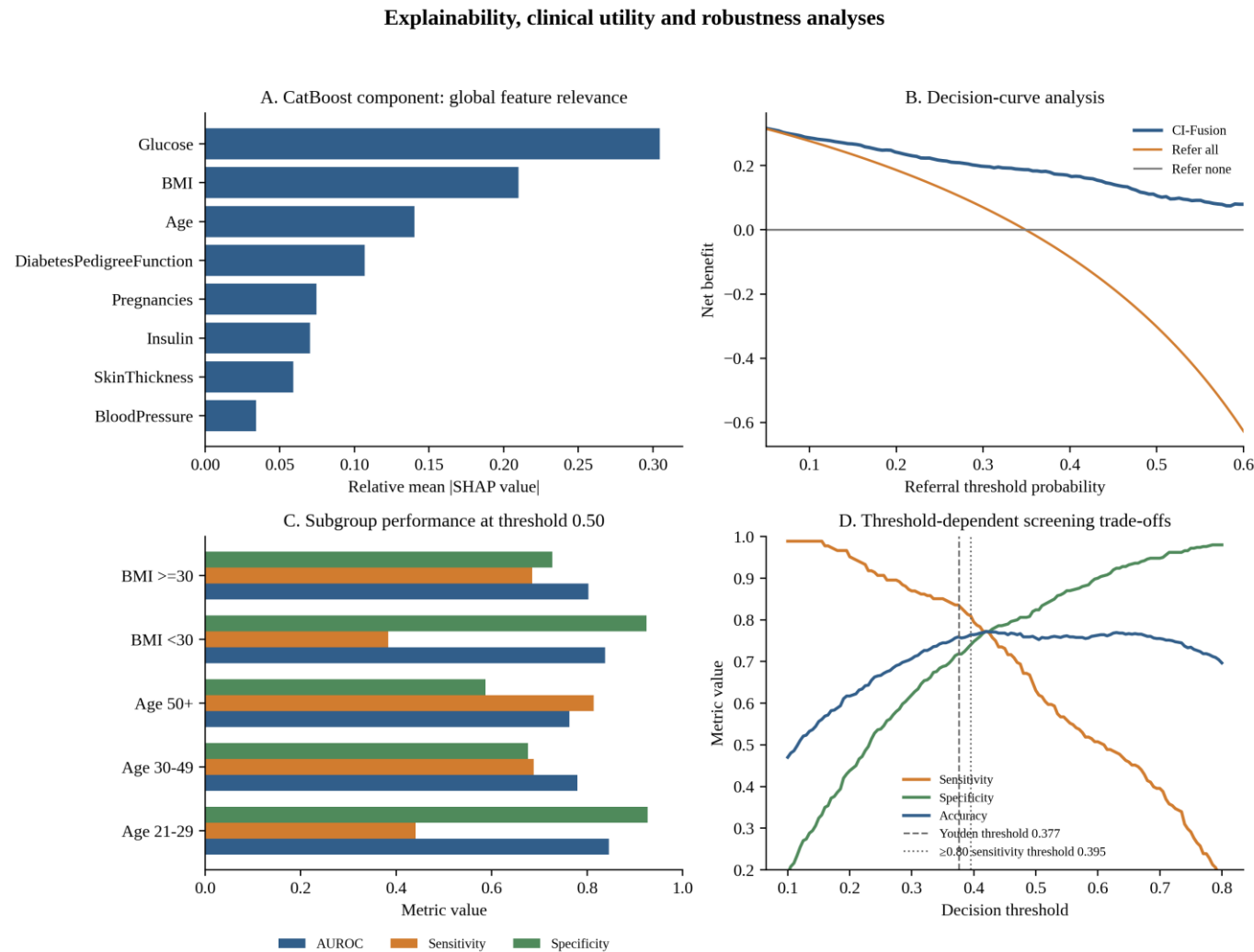
Note. Distributions use non-zero physiologically plausible values for Glucose and BMI. The zero-value panel reports records encoded with values that are not clinically plausible and were therefore treated as missing within each training fold.

Figure 3. Predictive performance of the proposed framework.



Note. Panel A presents mean AUROC with 95% confidence intervals across 15 repeated-cross-validation test folds. Panels B-D use pooled ten-fold out-of-fold predictions.

Figure 4. Explainability, clinical utility and robustness analyses.



Note. SHAP values describe the CatBoost component and are not causal effects. Decision-curve net benefit interprets a positive prediction as referral for confirmatory assessment rather than a definitive diagnosis.