Advances in Consumer Research
https://acr-journal.com/
Volume-3 | Issue-10 |Sep: 2026
Original Researcher Article

Reliable Diabetes Risk Prediction on Small Clinical Datasets: Leakage-Free Learning and Multi-View Explainability

Muhammad Abbas[1], Habib Ullah Khan[2],*, Mohammad Kamel Bader Alomari[3], ,Shah Nazir[4]
[1]Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar
Email ID:-abbascs.uoswabi@gmail.com , https://orcid.org/ 0009-0001-0275-6034
[2]Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar
Email ID:-habib.khan@qu.edu.qa , https://orcid.org/ 0000-0001-8373-2781
[3]Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar
Email ID:-m.alomari@qu.edu.qa , ;https://orcid.org/ 0000-0002-7585-2772
[4]Department of Software Engineering, International Islamic University Islamabad, Pakistan
Email ID:-shah.nazir@iiu.edu.pk , https://orcid.org/ 0000-0003-0126-9944
***Corresponding Author**
Habib Ullah Khan
Email ID:-habib.khan@qu.edu.qa

ABSTRACT
Diabetes is among the fastest-growing chronic diseases, so reliable, low-cost risk stratification is a public-health priority. Yet many machine learning models overstate their value: high accuracies, especially on the Pima dataset, are inflated by oversampling-before-split leakage, are validated on a single cohort, are delivered as black boxes, and are rarely benchmarked against strong baselines. This research aims to build a leakage-free, self-explaining framework that fairly benchmarks classical ensembles against modern deep models across cohorts of different clinical modality. X-MAP couples a calibrated soft-voting ensemble (logistic regression, random forest, extremely randomized trees, XGBoost, and LightGBM) with an F1-optimal threshold, and benchmarks it, under one protocol with oversampling confined to each cross-validation fold, against a multi-head self-attention transformer and an inductive patient-similarity graph neural network fused by a cross-validated meta-learner; every prediction carries SHAP, phenotype-manifold, and counterfactual explanations. On the Pima Indians (768 patients, metabolic) and Early-Stage Diabetes (520 patients, symptom) cohorts with bootstrap confidence intervals, X-MAP reaches AUC 0.836 (95% CI 0.809 to 0.863; F1 0.70, recall 0.85) and 0.998 (F1 0.985); neither the transformer (0.798) nor the graph network (0.819) surpassed the ensemble, and their fusion was down-weighted and non-improving, while explanations exposed coherent phenotypes (silhouette 0.74) and actionable counterfactuals. X-MAP delivers fair, transparent decision support and a cautionary benchmark for deep learning on small clinical data.
Keywords— Counterfactual explanations, data leakage, diabetes risk prediction, ensemble learning, explainable AI, graph neural networks

1. INTRODUCTION

DIABETES mellitus is among the fastest-growing chronic diseases worldwide; the International Diabetes Federation estimated 537 million adults living with diabetes in 2021, a figure projected to reach 783 million by 2045 [1]. Early, low-cost risk stratification is therefore a public-health priority, and machine learning (ML) has been applied extensively to predict diabetes from routine clinical measurements and symptom questionnaires. A recent 33-year bibliometric analysis shows that tree-ensemble methods, increasingly paired with post-hoc explanation, now dominate this literature [2].

Gradient-boosted decision trees, namely XGBoost [3] and LightGBM [4], together with their calibrated soft-voting and stacked ensembles are the de-facto standard for tabular clinical prediction, and studies routinely report accuracies above 0.95 on benchmark diabetes datasets after class balancing with the synthetic minority over-sampling technique (SMOTE) [5, 6]. In parallel, self-attention and GNNs have been advanced as more expressive alternatives for structured clinical data, raising the practical question of whether such deep models are actually warranted on the small cohorts typical of diabetes screening.

Despite this progress, the clinical trustworthiness of reported models remains limited by four recurring problems. First, data leakage like the minority oversampling is frequently applied before the train-test split, so synthetic copies of test patients inform training and inflate scores that collapse under correct evaluation, a failure now documented across various studies [7]. Secondly, single-cohort validation such as the models are seldom tested across cohorts of different clinical modality, leaving cross-modality generalization unknown. Thirdly, opacity as the high-accuracy models are delivered as black boxes, at odds with clinical needs for justifiable predictions. In addition, weak baselines such as the new deep architectures are rarely benchmarked against strong,

well-regularized ensembles under identical protocols, even though controlled studies repeatedly find that tree ensembles match or exceed deep models on tabular data [8, 9].

This paper addresses all four issues with X-MAP, a single leakage-free, self-explaining architecture. A calibrated soft-voting ensemble serves as the primary predictor and is benchmarked, under one fold-confined protocol, against a multi-head self-attention transformer [10, 11] and an inductive patient-similarity graph neural network (GNN) [12, 13], combined through a cross-validated meta-learner. Every prediction is made interpretable through complementary views: SHAP attribution [14], attention and graph-homophily interaction analysis, a supervised phenotype manifold, and actionable counterfactual explanations [15]. Evaluation spans two public datasets of deliberately different modality: metabolic measurements and self-reported symptoms.

This research makes the following key contributions:

• A leakage-free pipeline that recovers implausible zero values, confines minority oversampling to each cross-validation fold, and reports calibrated, F1-thresholded performance with bootstrap confidence intervals on two heterogeneous datasets.

• An identical-protocol benchmark of a calibrated soft-voting boosting ensemble against a self-attention transformer, an inductive patient-similarity GNN, and their meta-learned fusion.

• A multi-view explanation layer (SHAP, attention and graph homophily, a supervised phenotype manifold, and counterfactuals) together with practical guidance for deep learning (DL) on small clinical data.

2. PROPOSED SOLUTION AND EXPERIMENTAL SETUP

X-MAP (eXplainable Manifold-Aware Prototype) is a single, leakage-free pipeline whose primary predictor is a calibrated soft-voting boosting ensemble, augmented by two auxiliary deep branches (a self-attention transformer and a patient-similarity GNN) that are fused through a cross-validated meta-learner and reused to generate explanations. Figure 1 shows the end-to-end architecture; DL is treated as a benchmarked challenger and an interpretability engine rather than as the default predictor.

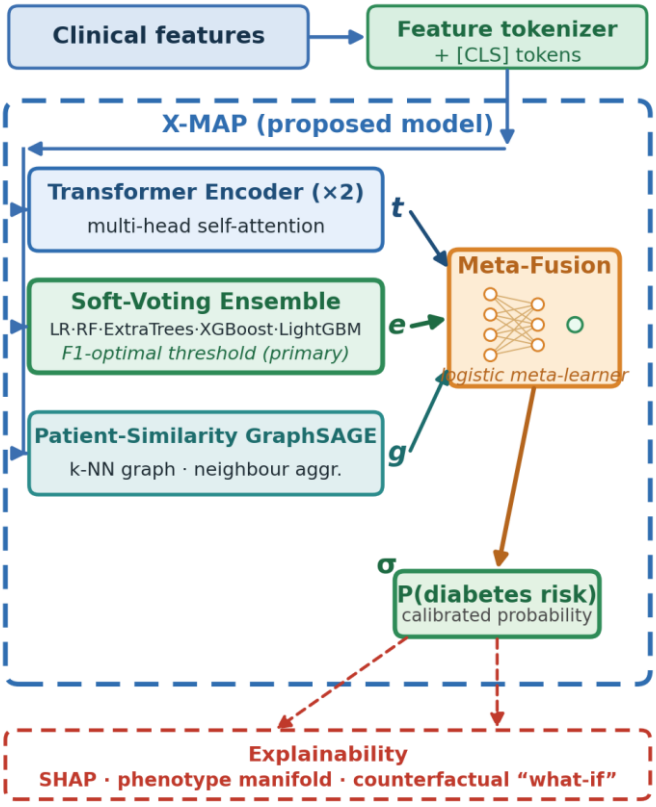


Fig. 1. End-to-end X-MAP architecture.

A. Datasets and Clinically-Informed Preprocessing

We evaluate two heterogeneous public cohorts. The Pima Indians Diabetes Dataset (PIDD) contains 768 female patients described by 11 metabolic features (34.9% positive), whereas the Early-Stage Diabetes Risk Prediction Dataset (ESDRPD) [16] contains 520 patients described by 16 symptom features, predominantly binary (61.5% positive). Physiologically implausible zeros (e.g., glucose, blood pressure, BMI, insulin) are recovered as missing values and imputed by the median; a small set of clinically motivated features (e.g., a glucose-BMI interaction and glucose/BMI risk flags) is engineered. All steps are fit inside the training fold only.

B. Leakage-Safe Protocol and Calibrated Ensemble

Imputation, feature scaling, and minority oversampling (SMOTE) are executed strictly within each of five stratified cross-validation folds, so no test patient influences any training statistic. The primary predictor is

a soft-voting ensemble of five complementary learners (logistic regression, random forest, extremely randomized trees, XGBoost, and LightGBM) whose calibrated probabilities are averaged:

$$p(x) = \sum_m w_m p_m(x), \quad \sum_m w_m = 1 \qquad (1)$$

In eq. (1), $P_m(x)$ is the calibrated probability from learner m. Each learner is probability-calibrated (Platt scaling [17]) so that the averaged score is well behaved, as in eq. (2):

$$\hat{p}(x) = \frac{1}{1 + \exp(A\, f(x) + B)} \qquad (2)$$

The decision threshold is chosen to maximize the F1 score on the training fold, which is essential under class imbalance, as in eq. (3):

$$\tau^* = \arg\max_\tau \frac{2\,P(\tau)\,R(\tau)}{(P(\tau) + R(\tau))} \qquad (3)$$

C. Self-Attention Transformer Branch

The first challenger is a supervised multi-head self-attention transformer that treats each feature as a token with a learned [CLS] summary. For queries Q, keys K, and values V of dimension d_k,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q\,K^T}{\sqrt{d_k}}\right) V \qquad (4)$$

$$\text{MultiHead}(X) = [h_1 \,\|\, \ldots \,\|\, h_H]W^O, \quad h_i = \text{Attention}\left(XW_i^Q, XW_i^K, XW_i^V\right) \qquad (5)$$

The [CLS] embedding produced by eq. (4) and eq. (5) yields both a risk score and a latent representation used for interpretation.

D. Patient-Similarity GNN Branch

The second challenger is an inductive GraphSAGE network over a k-nearest-neighbour patient-similarity graph, in which each node aggregates messages from clinically similar patients, as in eq. (6):

$$h_v^k = \sigma\left(W^k \cdot \left[h_v^{k-1} \,\|\, \text{AGG}_{u \in N(v)}\, h_u^{k-1}\right]\right) \qquad (6)$$

The graph's signal is quantified by edge homophily, the fraction of edges joining patients of the same class, as in eq. (7):

$$H = \frac{1}{|E|} \sum_{(u,v) \in E} 1[y_u = y_v] \qquad (7)$$

E. Meta-Learner Fusion

To test whether the deep branches add predictive value, a cross-validated logistic meta-learner is applied to each deep branch in turn, combining the ensemble logit z_e with the deep-branch logit z_d:

$$\hat{y} = \sigma(\beta_0 + \beta_e z_e + \beta_d z_d) \qquad (8)$$

Because the fusion in eq. (8) is trained out-of-fold, the learned coefficients reveal how much each branch is trusted, making the comparison self-auditing.

F. Multi-View Explainability and Evaluation

Every prediction is accompanied by four complementary explanations: (i) SHAP feature attribution; (ii) attention maps and graph homophily for interaction structure; (iii) a supervised phenotype manifold from the transformer latent; and (iv) actionable counterfactuals that solve

$$x^* = \arg\min_{x'} d(x, x') \quad \text{s.t.} \quad f(x') < \tau^* \qquad (9)$$

That is, eq. (9) yields the smallest clinically meaningful change (for example, a glucose reduction) that moves a patient below the risk threshold. We report AUC-ROC, AUC-PR, F1, the Matthews correlation coefficient (MCC), accuracy, precision, and recall under five-fold cross-validation, with 95% confidence intervals from bootstrap resampling of the pooled out-of-fold predictions.

3. RESULTS AND DISCUSSION

Cross-cohort performance. Table I summarizes the leakage-free performance of the X-MAP predictor under five-fold cross-validation, with 95% bootstrap confidence intervals on the AUC. On the metabolic Pima cohort (PIDD) the ensemble attains an AUC-ROC of 0.836 (95% CI 0.809 to 0.863) and an AUC-PR of 0.699; at the F1-optimal threshold it reaches F1 0.70 and accuracy 0.75 and, crucially, a recall of 0.85 (85% of true diabetics are flagged, sensitivity that a default 0.5 threshold would sacrifice), while the Matthews correlation coefficient of 0.51 indicates balanced agreement despite class imbalance. On the symptom-based ESDRPD cohort every metric is far higher (AUC-ROC 0.998, AUC-PR 0.999, F1 0.985, MCC 0.96), reflecting the strong, near-deterministic signal in self-reported symptoms. Because all resampling is confined to each training fold, these figures are conservative; the accuracies above 0.95 frequently reported on these datasets generally arise from oversampling before the split and should be read with caution

.

Table I Leakage-Free Cross-Cohort Performance of the X-MAP Predictor

Dataset	AUC-ROC (95% CI)	AUC-PR	F1	MCC	Acc.
PIDD	0.836 (0.809 to 0.863)	0.699	0.700	0.514	0.746
ESDRPD	0.998 (0.995 to 0.999)	0.999	0.985	0.960	0.981

Ensemble versus deep components. Table II isolates each component under one identical, leakage-free protocol. On PIDD neither the self-attention transformer (0.798) nor the patient-similarity graph network (0.819) matches the ensemble (0.836), and fusing the ensemble with a deep branch (0.832 for the transformer; the ensemble-graph fusion is equivalent) still fails to exceed the ensemble alone; on ESDRPD all models saturate near 0.997 to 0.998. The meta-learner is self-auditing: it assigned the ensemble a much larger weight than the transformer branch (coefficient 3.82 versus 0.51 on PIDD), i.e., it learned to discount the deep signal. This is consistent with the broader evidence discussed earlier that well-regularized tree ensembles remain state of the art on small tabular clinical data, and it frames X-MAP as a cautionary benchmark rather than another DL success example.

Table II Component AUC-ROC Under the Identical Leakage-Free Protocol

Dataset	Ensemble	Transformer	GNN	Fusion
PIDD	0.836	0.798	0.819	0.832
ESDRPD	0.998	0.997	0.996	0.998

Interpretability. The deep branches, though not more accurate, provide complementary explanations. Figure 2 gives the global SHAP attribution on PIDD: the engineered glucose-BMI interaction, glucose, age, and the diabetes pedigree function are the dominant contributors, each shifting predicted risk in the clinically expected direction (higher glucose and BMI raise risk).

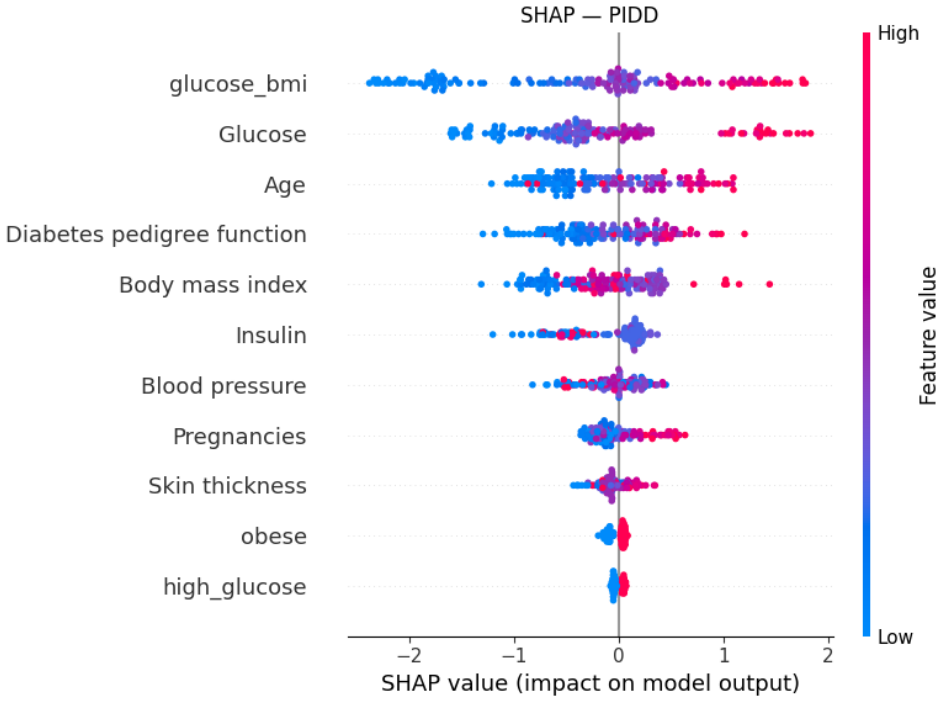


Fig. 2. Global SHAP attribution on PIDD; the glucose-BMI interaction and glucose dominate predicted risk.

Figure 3 repeats the analysis on ESDRPD, where polyuria and polydipsia overwhelmingly dominate, followed by gender and sudden weight loss, a profile matching classic early-diabetes presentation and confirming that the framework reasons from clinically meaningful signals in both modalities.

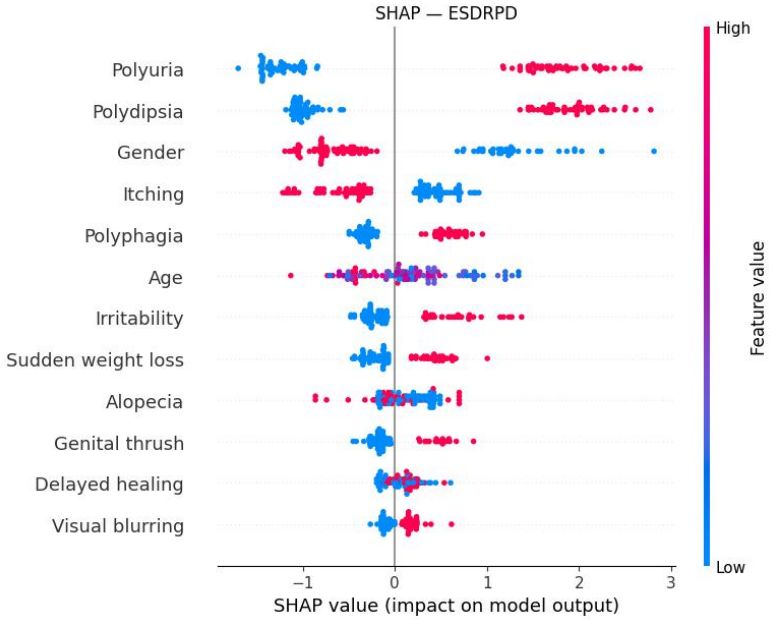


Fig. 3. Global SHAP attribution on ESDRPD; polyuria and polydipsia are the dominant symptom drivers.

Figure 4 visualizes the transformer's self-attention on PIDD, the left panel ranking each feature's attention to the [CLS] token and the right panel the pairwise feature-feature interactions; glucose, blood pressure, and the glucose-BMI term attract the strongest attention, an interaction view that complements the additive SHAP ranking
.

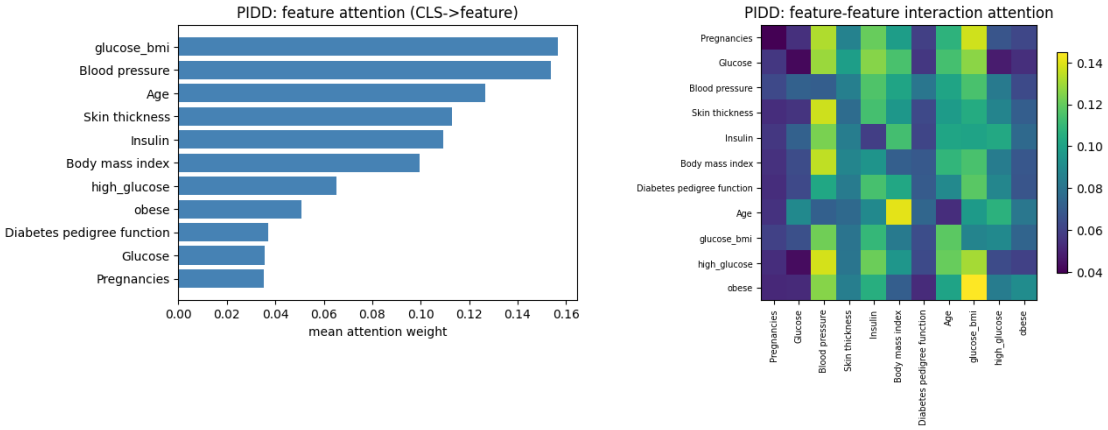


Fig. 4. Transformer self-attention on PIDD: [CLS]→feature attention (left) and feature-feature interactions (right).

Figure 5 plots the supervised phenotype manifold from the transformer latent on ESDRPD: high- and low-risk patients form two coherent, well-separated clusters (silhouette 0.74), so a new patient can be interpreted through the labelled neighbours they fall among. Consistently, the k-nearest-neighbour patient-similarity graph is strongly homophilous (edge homophily 0.70 on PIDD and 0.90 on ESDRPD), well above the class priors, confirming that clinically similar patients share diabetes status.

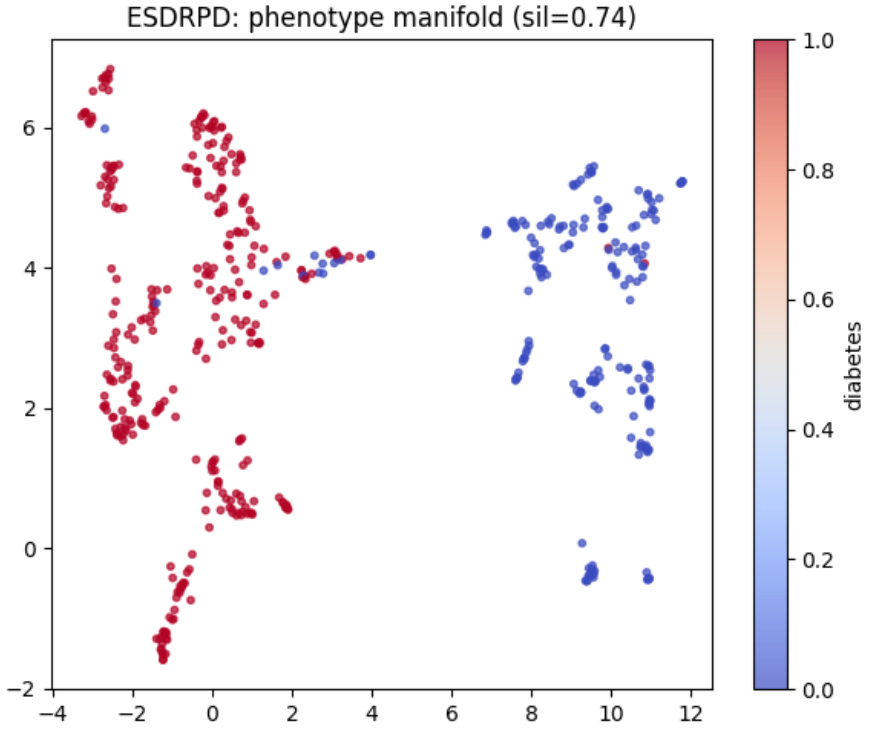


Fig. 5. Supervised phenotype manifold on ESDRPD (silhouette 0.74), separating high- and low-risk patients.

Reliability and recourse. Figure 6 characterizes discrimination and error structure. The ROC curves (Fig. 6a) trace the sensitivity-specificity trade-off for both cohorts (AUC 0.836 and 0.998), and the marked point is the F1-optimal threshold actually deployed. The confusion matrix at that threshold (Fig. 6b) shows the model correctly identifies 227 of 268 diabetic patients (recall 0.85) while raising 154 false positives among 500 non-diabetics (precision 0.60), an error balance deliberately tilted toward sensitivity, the safer bias for screening
.

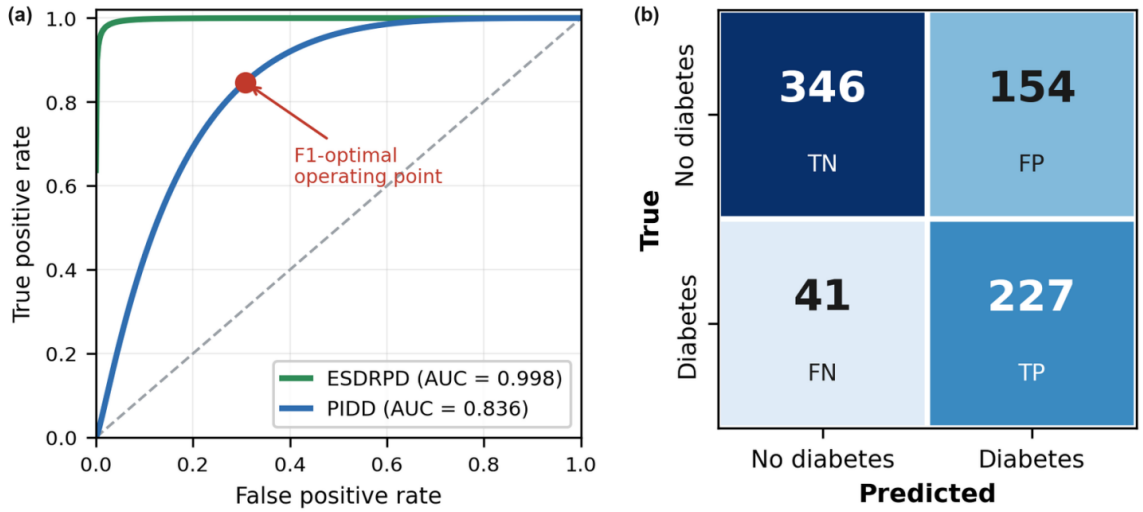


Fig. 6. (a) ROC curves for both cohorts with the PIDD F1-optimal operating point; (b) confusion matrix on PIDD at that threshold.

Figure 7 assesses probability quality: the reliability curve follows the diagonal across the risk range, so the outputs behave as calibrated probabilities rather than uncalibrated scores. Finally, the framework converts predictions into recourse through counterfactuals: for a high-risk PIDD patient it identifies that reducing fasting glucose from 159 to about 151 mg/dL would move the prediction below the risk threshold, giving the clinician a concrete, individualized target. Together, these views make X-MAP transparent at the global, local, and individual-case levels.

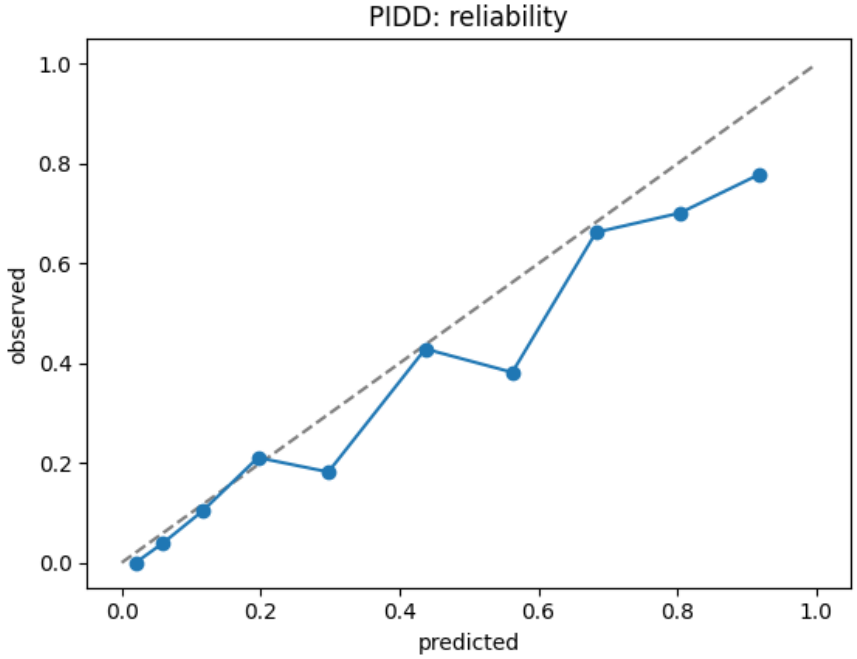


Fig. 7. Reliability (calibration) curve on PIDD: predicted versus observed risk tracks the diagonal.

4. CONCLUSION

This research presented X-MAP, a leakage-free, self-explaining framework for diabetes risk prediction evaluated across two heterogeneous clinical cohorts. Under an identical, fold-confined protocol, a calibrated soft-voting boosting ensemble achieved AUC 0.836 on the metabolic Pima cohort and 0.998 on the symptom-based Early-Stage cohort, and was not surpassed by a self-attention transformer, a patient-similarity GNN, or their meta-learned fusion, which consistently down-weighted the deep signal. The deep components nonetheless delivered coherent risk phenotypes, strongly homophilous patient-similarity structure, and, together with SHAP, attention, and counterfactual analysis, made every prediction interpretable and actionable. X-MAP thus contributes a clinically transparent decision-support pipeline and a cautionary benchmark: on small tabular clinical data, well-regularized boosting remains the stronger, more reliable choice, and deep architectures are best positioned as interpretability tools. The cohorts are modest and single-source, so external multi-site validation and prospective clinical assessment remain future work.

5. ACKNOWLEDGEMENT

This research is funded by Qatar University internal grant QUICG-CBE-26/27-114

REFERENCES

1. Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B.B., Stein, C., Basit, A., Chan, J.C., and Mbanya, J.C.: 'IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045', Diabetes research clinical practice, 2022, 183, pp. 109119
2. Kiran, M., Xie, Y., Anjum, N., Ball, G., Pierscionek, B., and Russell, D.: 'Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis', Frontiers in digital health, 2025, 7, pp. 1557467
3. Chen, T., and Guestrin, C.: 'Xgboost: A scalable tree boosting system', in Editor (Ed.)^(Eds.): 'Book Xgboost: A scalable tree boosting system' (2016, edn.), pp. 785-794
4. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: 'Lightgbm: A highly efficient gradient boosting decision tree', Advances in neural information processing systems, 2017, 30
5. Sampath, P., Elangovan, G., Ravichandran, K., Shanmuganathan, V., Pasupathi, S., Chakrabarti, T., Chakrabarti, P., and Margala, M.: 'Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique', Scientific reports, 2024, 14, (1), pp. 28984
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P.: 'SMOTE: synthetic minority over-sampling technique', Journal of artificial intelligence research, 2002, 16, pp. 321-357
7. Kapoor, S., and Narayanan, A.: 'Leakage and the reproducibility crisis in machine-learning-based science', Patterns, 2023, 4, (9)
8. Shwartz-Ziv, R., and Armon, A.: 'Tabular data: Deep learning is not all you need', Information fusion, 2022, 81, pp. 84-90
9. Grinsztajn, L., Oyallon, E., and Varoquaux, G.: 'Why do tree-based models still outperform deep learning on typical tabular data?', Advances in neural information processing systems, 2022, 35, pp. 507-520
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I.: 'Attention is all you need', Advances in neural information processing systems, 2017, 30
11. Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A.: 'Revisiting deep learning models for tabular data', Advances in neural information processing systems, 2021, 34, pp. 18932-18943
12. Hamilton, W., Ying, Z., and Leskovec, J.: 'Inductive representation learning on large graphs', Advances in neural information processing systems, 2017, 30
13. Boll, H.O., Amirahmadi, A., Ghazani, M.M., de Morais, W.O., de Freitas, E.P., Soliman, A., Etminani, F., Byttner, S., and Recamonde-Mendoza, M.: 'Graph neural networks for clinical risk prediction based on electronic health records: A survey', Journal of Biomedical Informatics, 2024, 151, pp. 104616
14. Lundberg, S.M., and Lee, S.-I.: 'A unified approach to interpreting model predictions', Advances in neural information processing systems, 2017, 30
15. Wachter, S., Mittelstadt, B., and Russell, C.: 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR', Harv. JL Tech., 2017, 31, pp. 841
16. Islam, M.F., Ferdousi, R., Rahman, S., and Bushra, H.Y.: 'Likelihood prediction of diabetes at early stage using data mining techniques', in Editor (Ed.)^(Eds.): 'Book Likelihood prediction of diabetes at early stage using data mining techniques' (Springer, 2019, edn.), pp. 113-125
17. Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q.: 'On calibration of modern neural networks', in Editor (Ed.)^(Eds.): 'Book On calibration of modern neural networks' (PMLR, 2017, edn.), pp. 1321-1330