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

Muhammad Abbas
Department of Accounting and Information Systems, Qatar University, Doha, Qatar

Habib Ullah Khan
Department of Accounting and Information Systems, Qatar University, Doha, Qatar

Shah Nazir
Department of Accounting and Information Systems, Qatar University, Doha, Qatar

Abstract:

Diabetes mellitus is among the fastest-growing chronic diseases worldwide, and timely risk stratification is critical for early intervention and cost-effective care; consequently, machine learning has been widely applied to predict diabetes from routine clinical data. Yet the clinical trustworthiness of existing models remains limited, as many high-accuracy results, particularly on the PIMA dataset, are inflated by oversampling-before-split leakage, validated on a single cohort, offered as black-box models, and rarely benchmarked against strong, well-regularized baselines despite the growing adoption of deep and graph neural architectures. To address these gaps, this study aims to develop and rigorously evaluate a leakage-free, self-explaining framework for diabetes risk prediction that benchmarks classical ensemble learning against modern self-attention and graph neural networks while equipping every prediction with clinician-actionable explanations across cohorts of different clinical modality. To achieve this, the framework applies clinically informed preprocessing, recovering implausible zero values as missing values and performing minority class oversampling strictly within each cross-validation fold to prevent leakage and adopts a calibrated soft-voting ensemble of diverse learners (logistic regression, random forest, extremely randomized trees, XGBoost, and LightGBM) with F1-optimal decision thresholds as its primary predictor.

To assess the added value of deep learning, hybrid variants fuse this ensemble with a supervised multi-head self-attention transformer and an inductive patient-similarity graph neural network through a cross-validated meta-learner; interpretability is delivered through complementary views spanning feature attribution (SHAP), interaction analysis (attention maps and graph homophily), a supervised clinical-phenotype manifold for case-based reasoning, and actionable counterfactual "what-if" explanations. Evaluated on the Pima Indians (768 patients, metabolic features) and Early-Stage Diabetes Risk (520 patients, symptom features) cohorts using five-fold cross-validation with bootstrap confidence intervals, the framework achieved a leakage-free AUC of 0.836 (95% CI 0.809–0.863; F1 0.70, recall 0.85) and 0.998 (F1 0.985), respectively. Notably, neither the transformer (0.798) nor the graph network (0.819) surpassed the ensemble, and their fusion was consistently down-weighted and non-improving, demonstrating that well-regularized boosting remains the stronger and more reliable option on small tabular clinical data. The deep components nonetheless yielded

meaningful interpretability, exposing clinically coherent risk phenotypes (manifold silhouette up to 0.74) and strong patient-similarity homophily (0.70–0.90), while counterfactuals produced concrete, actionable targets such as the minimal glucose reduction required to lower predicted risk. By uniting fair, leakage-free evaluation with multi-view explainability, this study delivers a reproducible and clinically transparent decision-support framework and offers a cautionary benchmark for deep learning on small clinical datasets.

Keywords:

Diabetes risk prediction, Explainable AI (XAI), Ensemble learning, Data leakage, Machine learning in healthcare, SHAP, Counterfactual explanations, Self-attention transformers, Graph neural networks, Clinical decision support.