Title: A Healthcare-Oriented Machine Learning Framework for Early Detection of Type 2 Diabetes

Author names: Md Ashraful Alam; Jahintaqi Chisty; Masud Rana Nayeem; Shomitro Kumar Ghosh; Anik Biswas; Md Fokhrul Alam

Abastact: Type 2 Diabetes Mellitus (T2DM) is a major health issue of critical importance on the international level, and Bangladesh has the highest incidence rate of 13.2% of the disease among adults. Although machine learning (ML) has been developed to predict T2DM, current research has several drawbacks, such as small samples, a lack of clinical factors, extreme levels of class imbalance, etc. This research constructs an optimally implemented ML model to predict T2DM in Bangladesh based on an enhanced dataset of 485 patients, including a complete set of demographic and biochemical characteristics. We evaluated eight ML models, including Logistic Regression, K-Nearest Neighbors, Support Vector Classifier, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and XGBoost, with rigorous GridSearchCV hyperparameter tuning. To address class imbalance, we applied five balancing techniques: SMOTE, Random Oversampling, Random Undersampling, ADASYN, and SMOTETomek. Performance was measured using Accuracy, Precision, Recall, F1 Score, Specificity, Matthews Correlation Coefficient, Kappa, Confusion Matrix, and ROC-AUC curves. After hyperparameter tuning, Random Forest achieved the best performance with 98.21%±0.0791 accuracy, 98.47%±0.0724 precision, and 98.43%±0.0787 F1 score, with an AUC of 0.98. Among balancing techniques, Random Oversampling with Random Forest produced a superior performance with 99.50% accuracy, and 99.50% F1 score. Our approach effectively surpasses the previous limitations by integrating balanced datasets with optimized models and comprehensive consideration that can offer a solid practice in predicting T2DM at the early stage and implementing clinical intervention in Bangladesh.

DOI: 10.1109/ICMCSI67283.2026.11412560