METABOLOMICS-BASED PREDICTION MODEL FOR DIABETES: A COMPREHENSIVE ANALYSIS OF BIOMARKERS AND MACHINE LEARNING APPROACHES

Doaa Farid, Farhana Saleh, Tareq Mohammed, Atika Nigar, Abderrahmane Maaradji

Aims: To develop a prediction model for diabetes using metabolomics data and to evaluate various machine learning approaches and identify the most effective framework for disease prediction.

Methods: A comprehensive analysis was conducted on the Qatar Biobank dataset comprising metabolomics profiles, instrument measurements, and clinical diagnoses from 450 Qatari nationals. Targeted metabolites were selected based on correlation strength with diabetes status. Five machine learning models (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Neural Network) were evaluated for their predictive performance using metrics including accuracy, precision, recall, F1 score, and ROC AUC.

Results: Among 450 individuals, 9.33 % (n = 42) were diagnosed with diabetes. Correlation analysis identified 140 metabolites significantly associated with diabetes status ($p < 0.05$). The most strongly correlated metabolites included glucose ($r = 0.281$, $p < 0.0001$), mannose ($r = 0.247$, $p < 0.0001$), and 1,5-anhydroglucitol ($r = -0.297$, $p < 0.0001$). Logistic Regression demonstrated superior performance with the highest accuracy (93.3 %), F1 score (0.625), and ROC AUC (0.941) compared to other models.

Conclusion: Metabolomics data can effectively predict diabetes status, with logistic regression providing the optimal balance of performance and interpretability. The identified metabolites offer potential biomarkers for early diabetes detection and monitoring. This model could serve as a valuable tool for clinical risk assessment and personalized preventive interventions.

Keywords: Diabetes mellitus, Metabolomics, Machine learning, Biomarkers, Prediction model, Logistic regression