[BIOINFO 2025]

Classifying Dementia Risk in Mild Cognitive Impairment Using SNP-Based Machine Learning with Longitudinal Validation

Myeongji Cho[1], Jeong Won Hong[1], Byung Soo Park[1], Hye Ryeong Nam[1] and Sang Cheol Kim[1,*]

[1]*Division of Healthcare and Artificial Intelligence, Department of Precision Medicine, National Institute of Health, Korea Disease Control and Prevention Agency, Cheong-Ju, Republic of Korea.*
**Corresponding author: sckim.knih@korea.kr*

Alzheimer's disease (AD) is a progressive neurodegenerative disorder, and mild cognitive impairment (MCI) represents its prodromal stage, with many individuals progressing to dementia annually. Genetic factors, particularly single nucleotide polymorphisms (SNPs), play a key role in AD pathogenesis. We aimed to develop predictive models for classifying MCI patients into high- and low-risk dementia groups using SNP chip data and machine learning (ML) algorithms.

Using data from the Biobank Innovations for Chronic Cerebrovascular Disease with Alzheimer's Disease Study, we performed a genome-wide association study (GWAS) in a Korean population-based cohort. Dementia-associated SNPs were selected to train four ML algorithms—random forest (RF), k-nearest neighbor (KNN), artificial neural network (ANN), and support vector machine (SVM). Three predictive models were constructed using distinct SNP subsets: Model 1 (38 SNPs, subjective memory impairment [SMI] vs. AD + vascular dementia), Model 2 (44 SNPs, SMI vs. AD), and Model 3 (68 SNPs, combined SNPs from Models 1 and 2).

All models showed high training accuracy, with Model 3 achieving the strongest performance (AUC = 0.818). In two-year follow-up validation, RF and ANN performed best in predicting MCI progression to AD. Despite these promising results, false-positive rates remained high, underscoring the need for refined feature selection and integration of additional biomarkers. External validation using an independent dataset of 80 participants (normal vs AD) further supported the generalizability of our framework.

These findings highlight the promise of integrating genetic data with ML-based approaches for personalized dementia risk assessment. Early identification of high-risk MCI patients may facilitate timely interventions and improved outcomes.