Explainable Machine Learning for Alzheimer's Disease Stage Classification Using Blood Gene Expression and Clinical Features

Bhargav Pamidighantam, Akshatt Kain, Moumita Baidya Baidyam

Abstract

Alzheimer's disease (AD) affects over 55 million people worldwide, with early detection critical for treatment efficacy. This project applies machine learning to classify AD progression stages (Cognitively Normal, Mild Cognitive Impairment, and AD/Dementia) using blood-based gene expression data integrated with clinical features. We compare five algorithms—Logistic Regression, SVM, Random Forest, XGBoost, and Deep Neural Networks—and systematically fine-tune XGBoost through feature selection, clinical feature fusion including novel early-onset status integration, and hyperparameter optimization. SHAP analysis [3] will provide biological validation and explainability for clinical deployment.

1 Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder affecting over 55 million people worldwide. Current diagnostics rely on expensive neuroimaging and invasive cerebrospinal fluid biomarkers, limiting accessibility. Blood-based biomarkers with machine learning offer a minimally invasive alternative for early detection and stage classification.

The primary challenge is high-dimensional, low-sample-size (HDLSS) data: gene expression datasets contain 20,000+ genes but only hundreds of samples, causing overfitting. Recent research shows integrating clinical features with gene expression improves accuracy, yet early-onset status (age at diagnosis < 65 years) remains unexplored.

This project addresses: (1) Which ML algorithms perform best for multiclass AD stage classification? (2) How does integrating early-onset information affect performance? (3) Can SHAP explainability validate biologically meaningful patterns?

2 Proposed Project

2.1 Problem Formulation

We formulate AD stage classification as a **supervised multiclass classification** problem with three classes: Class 0 (Cognitively Normal), Class 1 (Mild Cognitive Impairment), Class 2 (Alzheimer's Disease/Dementia). The model predicts disease stage from blood gene expression profiles combined with clinical variables.

2.2 Dataset Description

Primary: ADNI [4] - Blood gene expression with comprehensive clinical annotations (age, sex, education, APOE, MMSE, longitudinal diagnostics for onset calculation). **Validation: GSE63060** [5] - 329 samples (104 Normal, 80 MCI, 145 AD), 20,000 genes, clinical features including novel early-onset status (binary: age at onset < 65).

2.3 Preprocessing

Gene expression: log-transformation, quantile normalization, low-variance filtering. Feature selection via XGBoost importance + SFBS to reduce 20,000 genes to 800-1000. Clinical: standardize continuous variables, encode categoricals, create early-onset feature. Class imbalance: SMOTE oversampling and XGBoost class weighting.

2.4 Methodology

We first train five baseline algorithms via 5-fold stratified CV—Logistic Regression (L2), SVM (RBF kernel), Random Forest, XGBoost, and Deep Neural Network (3-layer + dropout)—evaluating accuracy, macro F1, precision, recall, and training time to identify the optimal model for high-dimensional genomic data. We then systematically optimize XGBoost through: (1) feature selection comparing 500, 1000, 2000 genes; (2) ablation study testing genes only, + basic clinical, + MMSE, + early-onset; (3) hyperparameter grid search (learning rate, max depth, estimators, subsample, regularization) via 5-fold CV $\times$ 10 repeats; (4) SHAP analysis identifying top features globally and per-class, with subgroup validation for EOAD markers (PSEN1/APP) and LOAD markers (APOE/CLU/TREM2).

2.5 Expected Outcomes

Based on [2], XGBoost with feature fusion achieved superior multiclass AD classification. [1] found tree-based ensembles consistently outperformed traditional methods on gene expression data. We anticipate XGBoost will emerge as top baseline, with optimization yielding measurable improvements. Ablation will quantify early-onset contribution. SHAP will validate biological pathways (PSEN1/APP for EOAD, APOE/CLU/TREM2 for LOAD).

2.6 Evaluation Metrics

Primary: Accuracy, macro/weighted F1-score, per-class precision/recall. **Secondary:** Confusion matrix, multi-class ROC-AUC (one-vs-rest), Cohen's Kappa, Matthews Correlation Coefficient, training/inference time. All metrics via stratified 5-fold CV $\times$ 10 repeats for robust estimates with confidence intervals.

References

[1] S. Vadapalli, H. Abdelhalim, S. Zeeshan, and Z. Ahmed, *Artificial intelligence and machine learning approaches using gene expression and variant data for person-*

alized medicine, Briefings in Bioinformatics, vol. 23, no. 5, bbac191, 2022. DOI: 10.1093/bib/bbac191

[2] M. Sarma and S. Chatterjee, *'Machine Learning' multiclassification for stage diagnosis of Alzheimer's disease utilizing augmented blood gene expression and feature fusion*, Discover Applied Sciences, vol. 7, 636, 2025. DOI: 10.1007/s42452-025-07237-1

[3] S. M. Lundberg and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*, Advances in Neural Information Processing Systems, 2017.

[4] R. A. Sperling et al., *The A4 Study: Stopping AD Before Symptoms Begin?*, Science Translational Medicine, 2014.

[5] K. Lunnon et al., *A Blood Gene Expression Marker of Early Alzheimer's Disease*, Journal of Alzheimer's Disease, vol. 33, no. 3, pp. 669-677, 2013.