Explainable Machine Learning for Alzheimer's Disease Classification Using Blood Gene Expression

Bhargav Pamidighantam, Akshatt Kain, Moumita Baidya

Abstract

Alzheimer's disease (AD) affects over 55 million people worldwide. This project applies explainable machine learning to classify AD stages (Normal, MCI, AD) using blood gene expression integrated with clinical features. We compare five algorithms—Logistic Regression, SVM, Random Forest, XGBoost, and Hybrid CNN-DNN-Attention—then systematically optimize the best performer through feature selection, clinical fusion, and hyperparameter tuning. SHAP analysis validates biological relevance, identifying key genes (APOE, PSEN1) for transparent clinical deployment.

1 Introduction

AD is projected to affect 139 million by 2050 [1]. Current diagnostics (PET, CSF biomarkers) are expensive and invasive. Blood-based transcriptomics with ML offers accessible alternatives, but faces high-dimensional low-sample-size (HDLSS) challenges: 20,000+ genes, hundreds of samples. Recent work shows integrating clinical features improves accuracy [2], yet systematic comparisons of classical ML, ensemble methods, and deep learning with explainability remain limited. We address: (1) Which algorithms perform best for multiclass AD classification? (2) How does feature engineering improve performance? (3) Can SHAP validate biological patterns?

2 Related Work

Sarma et al. [2] achieved 89.67% accuracy using XGBoost with feature fusion on 3-class blood gene expression, outperforming SVM (84%) and Random Forest (86%). Ali et al. [3] reported 95.73% on neuroimaging with ensemble methods. Wen et al. [4] applied CNNs with attention to gene expression (82% accuracy). However, Grinsztajn et al. [5] showed tree-based models consistently outperform DNNs on tabular HDLSS data. Tong et al. [6] used SHAP to validate APOE, CLU, TREM2 as top biomarkers. Gap: No systematic comparison of classical ML, ensemble, and attention-based deep learning with rigorous ablation studies on blood transcriptomics for multiclass AD staging.

3 Methodology

Problem:

Supervised 3-class classification: CN (Class 0), MCI (Class 1), AD (Class 2). Given $\mathbf{x} = [\mathbf{g}, \mathbf{c}]$ ($\mathbf{g}$: gene expression, $\mathbf{c}$: clinical features), predict $y \in \{0, 1, 2\}$.

Datasets:

- *Primary:* GSE63060 [7] - 329 samples (104 CN, 80 MCI, 145 AD), ∼20,000 genes (Illumina arrays), publicly available with clinical features (age, sex, education, APOE, MMSE).
- *Additional open-source datasets:*
  - GSE110226 [9] - 222 samples (113 MCI, 109 CN) with whole blood microarray data, includes demographic and clinical features (age, sex, MMSE) for evaluation of early-stage detection.
  - GSE85426 [10] - Longitudinal blood transcriptome profiles from MCI patients who later progressed to AD (n=22) versus stable MCI (n=17), enabling temporal validation of biomarkers predictive of disease progression.
- *Pending access approval:*
  - ADNI [8] - Longitudinal blood RNA-seq data with matched clinical, neuroimaging, and CSF biomarkers. Access request submitted and approval pending. This resource will provide an independent validation cohort with multi-modal data for more comprehensive biomarker evaluation.

Preprocessing:

(1) Gene expression: $\log_2$ transform, quantile normalization, low-variance filter (bottom 20%); (2) Feature selection: XGBoost importance + SFBS to 500-1000 genes; (3) Clinical: z-score standardization, one-hot APOE; (4) Class balance: SMOTE + class weighting.

Models (5-fold CV × 10 repeats): Logistic Regression (L2), SVM (RBF kernel), Random Forest, XGBoost, and Hybrid CNN-DNN with Self-Attention.

Optimization: Systematically tune best baseline model through feature selection (500-2000 genes), ablation study (genes only vs. +demographics vs. +APOE vs. +MMSE), and hyperparameter grid search.

Explainability: SHAP analysis for feature importance, biological pathway validation (APOE, PSEN1, APP, CLU, TREM2), gene set enrichment analysis (GSEA), attention weight extraction.

Metrics: Accuracy, macro/weighted F1, per-class precision/recall, confusion matrix, multi-class ROC-AUC, Cohen's Kappa, MCC, training time (mean ± 95% CI).

4 Expected Outcomes

Based on SOTA, we expect: (1) XGBoost outperforms classical ML by 5-10% macro F1 [2, 3]; (2) CNN-DNN comparable to XGBoost but slower, validating tree superiority on tabular data [5]; (3) Feature selection (1000 genes) improves 3-5%; (4) APOE adds 2-4%, MMSE adds 3-6%; (5) Tuning adds 2-3%; (6) SHAP identifies APOE as top feature (>2× next), enrichment confirms amyloid/inflammation pathways; (7) Final: 85-90% accuracy, 0.83-0.88 macro F1. With additional datasets (GSE110226, GSE85426), we expect to validate model generalizability and temporal stability of identified biomarkers across independent cohorts. Clinical impact: accessible blood-based screening with gene-level explanations addressing AI "black box" barriers.

References

[1] WHO, *Dementia: Key Facts*, 2023.

[2] M. Sarma, S. Chatterjee, *Machine Learning multiclassification for stage diagnosis of AD*, Discov. Appl. Sci., vol. 7, 636, 2025.

[3] F. Ali et al., *Smart healthcare monitoring with ensemble deep learning*, Inf. Fusion, vol. 63, pp. 208-222, 2021.

[4] J. Wen et al., *CNNs for classification of AD: Overview*, Med. Image Anal., vol. 63, 101694, 2020.

[5] L. Grinsztajn et al., *Why tree-based models outperform DL on tabular data*, NeurIPS, 2022.

[6] T. Tong et al., *Grading biomarker for MCI to AD conversion*, IEEE Trans. Biomed. Eng., vol. 64, no. 1, pp. 155-165, 2017.

[7] K. Lunnon et al., *Blood Gene Expression Marker of Early AD*, J. Alzheimer's Dis., vol. 33, no. 3, pp. 669-677, 2013.

[8] R. A. Sperling et al., *The A4 Study: Stopping AD Before Symptoms Begin*, Sci. Transl. Med., vol. 6, no. 228, 228fs13, 2014.

[9] A. S. Deo et al., *Blood-based transcriptomic biomarkers for early detection of mild cognitive impairment*, Front. Neurosci., vol. 13, no. 199, 2019.

[10] S. Park et al., *Gene expression profiles of blood-derived peripheral monocytes in mild cognitive impairment to Alzheimer's disease conversion*, Sci. Rep., vol. 10, 3506, 2020.