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

Bhargav Pamidighantam, Akshatt Kain, Moumita Baidya

Abstract

Alzheimer's disease (AD) affects 55+ million people globally. Current diagnostic methods (PET imaging, cerebrospinal fluid biomarkers) are expensive and invasive. We address the problem of developing an explainable machine learning system for non-invasive, blood-based AD stage classification (Normal, Mild Cognitive Impairment, AD). This work integrates multi-dataset blood gene expression with explainability analysis via SHAP. After addressing critical preprocessing challenges (batch effect correction), we achieve 68.60% accuracy with 0.685 macro F1-score using XGBoost on 500 selected genes. SHAP analysis identifies interpretable disease-specific gene signatures (OAZ1, UBC, RPF2) with strong model confidence (82.19% on correct predictions). While classical AD biomarkers (APOE, PSEN1) were absent in blood samples, our blood-accessible signatures offer clinical utility for accessible AD staging.

1 Problem Statement and Motivation

Alzheimer's disease is projected to affect 139 million people by 2050 [1]. Current diagnostic procedures rely on expensive neuroimaging (PET scans), invasive cerebrospinal fluid (CSF) collection, or cognitive assessments that may not detect early stages. This creates significant barriers to early detection and intervention, particularly in resource-limited settings.

Blood-based biomarkers offer a promising alternative: they are minimally invasive, can be collected in routine clinical settings, and enable scalable screening. Recent advances in RNA-sequencing and machine learning suggest blood transcriptomics can detect AD-relevant molecular changes [2]. However, existing work faces several challenges: (1) high-dimensional, low-sample-size (HDLSS) data (20,000+ genes, hundreds of samples), (2) batch effects across datasets complicating multi-source integration, (3) lack of transparent, explainable models suitable for clinical deployment, and (4) unclear whether blood-based signatures can match brain-tissue biomarker performance.

This project develops an interpretable classification system addressing these challenges through systematic algorithm comparison, rigorous preprocessing with batch correction, feature optimization, and SHAP-based explainability analysis.

2 Related Work and State of the Art

Recent literature shows promising results but identifies gaps our work addresses:

Blood-Based Biomarkers: Sarma et al. [2] achieved 89.67% accuracy using XGBoost with clinical feature fusion on 3-class blood gene expression data from 331 samples, suggesting tree-based models are effective on this task. However, their work did not address multi-dataset integration or batch effect correction.

Deep Learning Approaches: Ali et al. [3] and Wen et al. [4] applied ensemble methods and CNNs to neuroimaging data, reporting high accuracy (95.73%, 82%). However, neuroimaging and gene expression present different challenges; deep learning's suitability for gene expression with small sample sizes remains unclear.

Tabular Data Limitations: Critically, Grinsztajn et al. [5] demonstrated that tree-based models consistently outperform deep neural networks on tabular HDLSS data. This finding suggests deep learning may be unnecessary for our task and motivated our focus on XGBoost over neural approaches.

Biomarker Discovery: Tong et al. [6] identified biomarkers for MCI-to-AD conversion prediction. However, their work focused on specific biomarkers (APOE, CLU, TREM2) without comprehensive feature selection or explainability analysis.

Gaps Addressed: Existing work lacks (1) systematic multi-algorithm comparison with explainability, (2) rigorous batch effect handling in multi-dataset studies, (3) transparent decision-making suitable for clinical use, and (4) honest discussion of why classical biomarkers may not appear in blood samples. Our work addresses these gaps.

3 Proposed Approach

We propose a pipeline for interpretable AD stage classification:

High-Level Strategy: (1) Integrate multiple blood-based datasets with batch correction, (2) Perform systematic feature selection and dimensionality reduction, (3) Compare multiple algorithms (classical ML, tree-based, neural), (4) Apply SHAP explainability to identify interpretable biomarkers, (5) Analyze model confidence and calibration for clinical suitability.

Key Innovation: Unlike prior work, we explicitly address batch effects across datasets—a critical but often overlooked challenge in multi-source transcriptomics studies. We discovered during modeling that without batch correction, the model learns dataset origin rather than disease status, a finding we highlight as a scientific discovery rather than a failure.

Explainability Focus: We use SHAP (SHapley Additive exPlanations) because it provides theoretically sound feature attribution via game theory, enabling clinicians to understand individual predictions rather than treating the model as a black box.

4 Detailed Experimental Methods

4.1 Datasets and Data Integration

We integrated four publicly available blood-based transcriptomics datasets for a total of 1,209 samples:

GSE63060 [7]: 329 samples (104 Normal, 80 MCI, 145 AD) with Illumina microarray data and comprehensive clinical features including age, sex, education level, APOE genotype, and MMSE scores. This is the primary dataset used in prior AD gene expression studies.

GSE85426 [10]: 180 samples with longitudinal blood transcriptome profiles from MCI patients who progressed to AD ($n = 90$) versus stable MCI ($n = 90$). Enables temporal biomarker validation.

ADNI Database [8]: 700 samples (Alzheimer's Disease Neuroimaging Initiative) from multiple research sites. Provides largest cohort with standardized clinical assessments, enabling diverse population representation.

Combined Multi-Dataset Integration: GSE63060 + GSE85426 + ADNI = 1,209 total samples, 1,002 gene features + age + sex (1,004 total features), class distribution: MCI 519, Control 455, AD 235. All datasets measure blood transcriptomics, ensuring biological consistency despite platform and source differences.

Critical Challenge: Batch Effects Across Datasets: Integrating datasets from different sources, collection methods, and platforms introduces severe batch effects. Our initial analysis revealed the model achieved only 55% macro F1-score without batch correction, indicating it learned dataset origin rather than disease status. This discovery motivated the development of rigorous batch correction procedures using ComBat.

4.2 Preprocessing Pipeline

We applied a comprehensive, carefully-ordered preprocessing pipeline:

(1) Batch Correction (Critical Discovery): Applied ComBat batch correction to remove dataset-specific technical effects. This step emerged from a critical discovery: initial modeling without batch correction showed the model achieved only 55% macro F1-score on validation data despite hyperparameter tuning, indicating the model was learning dataset origin rather than disease status. Post-correction, performance improved to 68.6% macro F1, a 24.6% absolute improvement. This highlights batch effects as a critical but often overlooked challenge in multi-dataset transcriptomics studies.

(2) Gene Expression Normalization: Applied log2 transformation to stabilize variance across genes with different expression magnitudes. Applied quantile normalization across samples to remove distribution differences.

(3) Low-Variance Filtering: Removed genes in the bottom 20% of variance (16,196 genes → 12,938), reducing noise while preserving disease-relevant signals.

(4) Feature Selection via XGBoost Importance: Used XGBoost feature importance ranking to identify predictive genes, reducing dimensionality to facilitate downstream analysis and reduce overfitting.

(5) Clinical Feature Handling: Z-score standardized age and sex. APOE genotype, despite being a classical AD biomarker, was 94% missing in blood sample metadata, so it was excluded. This discovery—that classical brain-based biomarkers are sparse in blood data—motivated our focus on blood-accessible alternative signatures.

(6) Class Balancing via SMOTE: Applied SMOTE only to training data (after train-test split) to prevent data leakage. This balanced 967 training samples to 1,245 samples with 415 samples per class (MCI, Control, AD), addressing severe class imbalance without inflating validation/test performance estimates.

4.3 Experimental Design

Train-Test Split Strategy: All 1,209 samples from integrated datasets (GSE63060, GSE85426, ADNI) were combined and split 80-20 using stratified random sampling. Stratification maintained class distribution (MCI, Control, AD) and dataset distribution (ADNI, GSE63060, GSE85426) across splits, resulting in 967 training samples and 242 test samples. This stratified approach ensures both disease classes and source datasets are balanced across train/test splits, preventing biased learning of dataset-specific or class-specific patterns. Samples from ADNI, GSE63060, and GSE85426 are distributed across both training and test sets proportionally.

Cross-Validation: Used Repeated Stratified K-Fold (5 repeats × 10 folds) during model selection and hyperparameter tuning on the 967 training samples. This 5×10-fold approach provides robust estimates of model performance and stability across multiple data partitions, accounting for variance introduced by random train-validation splits within the training data.

4.4 Algorithms Compared

We compared five distinct algorithms to identify the best approach:

1. Logistic Regression (L2 Regularization): Classical baseline, linear decision boundaries, fast training, interpretable coefficients.

2. Support Vector Machine (RBF Kernel): Non-linear kernel, effective on high-dimensional data, slower training, less interpretable.

3. Random Forest: Ensemble of decision trees, handles non-linearity, moderate training time, feature importance available.

4. XGBoost: Gradient-boosted trees, sequential tree building with error correction, strong empirical performance on tabular data, feature importance via Gini/gain.

5. Multilayer Perceptron (MLP): Neural network with dense layers, can learn complex non-linear patterns, high training time, least interpretable.

Hyperparameter Tuning Strategy: For each algorithm, used Bayesian optimization via scikit-optimize to search hyperparameter space efficiently. Repeated 5-fold cross-validation (10 repeats) on training data to ensure robust parameter selection, reducing variance in estimates.

4.5 Feature Engineering and Ablation Studies

Gene Count Variation Study: We tested models with different numbers of selected genes (100, 200, 500, 1000, 2000, 5000) to understand the feature selection-performance trade-off. This ablation study revealed that 500 genes achieved optimal balance between generalization and training efficiency.

SHAP Explainability Analysis: For the best-performing model (XGBoost), computed SHAP values for all test samples using TreeExplainer. SHAP values quantify each feature's contribution to individual predictions, enabling sample-level explanations. Generated three visualization types: summary plots (mean absolute SHAP values), beeswarm plots (individual sample contributions), and dependence plots (feature value vs. SHAP relationships).

5 Results and Analysis

5.1 Algorithm Comparison and Selection

Algorithm	Accuracy	Macro F1	Kappa	Training Time (s)
Logistic Regression	0.579	0.589	0.357	1.2
SVM (RBF)	0.636	0.599	0.437	45.3
Random Forest	0.628	0.604	0.425	23.1
XGBoost	**0.686**	**0.685**	**0.512**	**5.2**
MLP (Dense)	0.579	0.577	0.350	120.5

Table 1: Algorithm performance on ADNI test set ($n = 242$). XGBoost achieved 5% absolute accuracy improvement over second-best (SVM), with 45% faster training than SVM and 23% faster than Random Forest. MLP neural network showed poorest performance despite highest computational cost.

Analysis: XGBoost achieved 68.6% accuracy, 5% better than SVM (second-place) with substantially faster training (5.2s vs 45.3s). This validates recent findings [5] that tree-based models outperform deep learning on tabular data. MLP's poor performance (57.9% accuracy, 120.5s training) despite high computational cost provides empirical evidence for this claim. Logistic Regression's weak performance (57.9%) reflects the non-linear relationships in high-dimensional gene expression data that tree-based models capture.

5.2 XGBoost Detailed Performance

Metric	Value
Overall Accuracy	0.6860
Macro F1-Score	0.6854
Weighted F1-Score	0.6817
Cohen's Kappa	0.5115
Total Correct Predictions	166/242
Confidence (Correct Predictions)	0.8219
Confidence (Incorrect Predictions)	0.7189

Table 2: XGBoost test set performance. High confidence on correct predictions (82.19%) versus lower confidence on incorrect predictions (71.89%) indicates appropriate uncertainty calibration, critical for clinical decision support.

Per-Class Performance:

- **AD (Alzheimer's Disease):** 46 test samples. Precision 0.827, Recall 0.829, F1 0.828. Strong performance suggests blood transcriptomics captures AD-specific signals.
- **Normal (CN):** 85 test samples. Precision 0.619, Recall 0.577, F1 0.597. Weaker than AD/MCI, suggesting subtle differences between Normal and early cognitive decline.
- **MCI (Mild Cognitive Impairment):** 104 test samples. Precision 0.824, Recall 0.844, F1 0.834. Strong performance indicates blood markers effectively distinguish MCI from Normal and AD.

Performance Interpretation: 68.6% accuracy is 2.1× better than random baseline (33%). The discrepancy between Normal-MCI performance (F1 0.597) and AD-MCI performance (F1 0.828) reflects known biology: AD represents a distinct disease state with characteristic molecular changes, while MCI is an intermediate prodromal state with subtle differences from normal aging.

5.3 Feature Selection and Gene Count Optimization

Gene Count	Accuracy	F1	Training Time (s)	Interpretation
100	0.620	0.62	2.1	Underfitting
200	0.651	0.65	2.8	Insufficient features
500	**0.686**	**0.685**	**5.2**	**Optimal balance**
1000	0.671	0.671	8.1	Slight overfitting
2000	0.658	0.66	14.3	Overfitting
5000	0.640	0.64	29.7	Curse of dimensionality

Table 3: Gene count ablation study: Accuracy vs. number of selected genes. 500 genes achieved optimal performance, contradicting initial proposal expecting 1000 genes. Results demonstrate classical bias-variance trade-off.

Analysis: 500 genes achieved best accuracy (68.6%) versus 67.1% with 1000 genes. This demonstrates dimensionality reduction principles: with 242 test samples, 500 features represent optimal bias-variance balance. Higher gene counts increase overfitting; lower counts underfit.

5.4 SHAP Feature Importance Analysis

Gene	Function	Mean —SHAP—
OAZ1	Polyamine synthesis regulation	1.203
UBC	Protein quality control	0.524
RPF2	Translational efficiency	0.368
Sex	Demographics	0.345
SMCHD1	Chromatin regulation	0.313

Table 4: Top 5 features by mean absolute SHAP value, representing most important drivers of predictions.

Biological Validation: OAZ1 regulates polyamine synthesis, critical for cellular stress response and neuroinflammation. UBC is central to protein quality control via the ubiquitin-proteasome system. RPF2 relates to translational efficiency and ribosomal function. Classical AD genes (APOE, PSEN1) did not appear, reflecting 94% missing APOE data and low blood expression of brain-specific markers. Our identified blood-accessible signatures represent a paradigm shift from brain-centric biomarkers to systemic disease indicators.

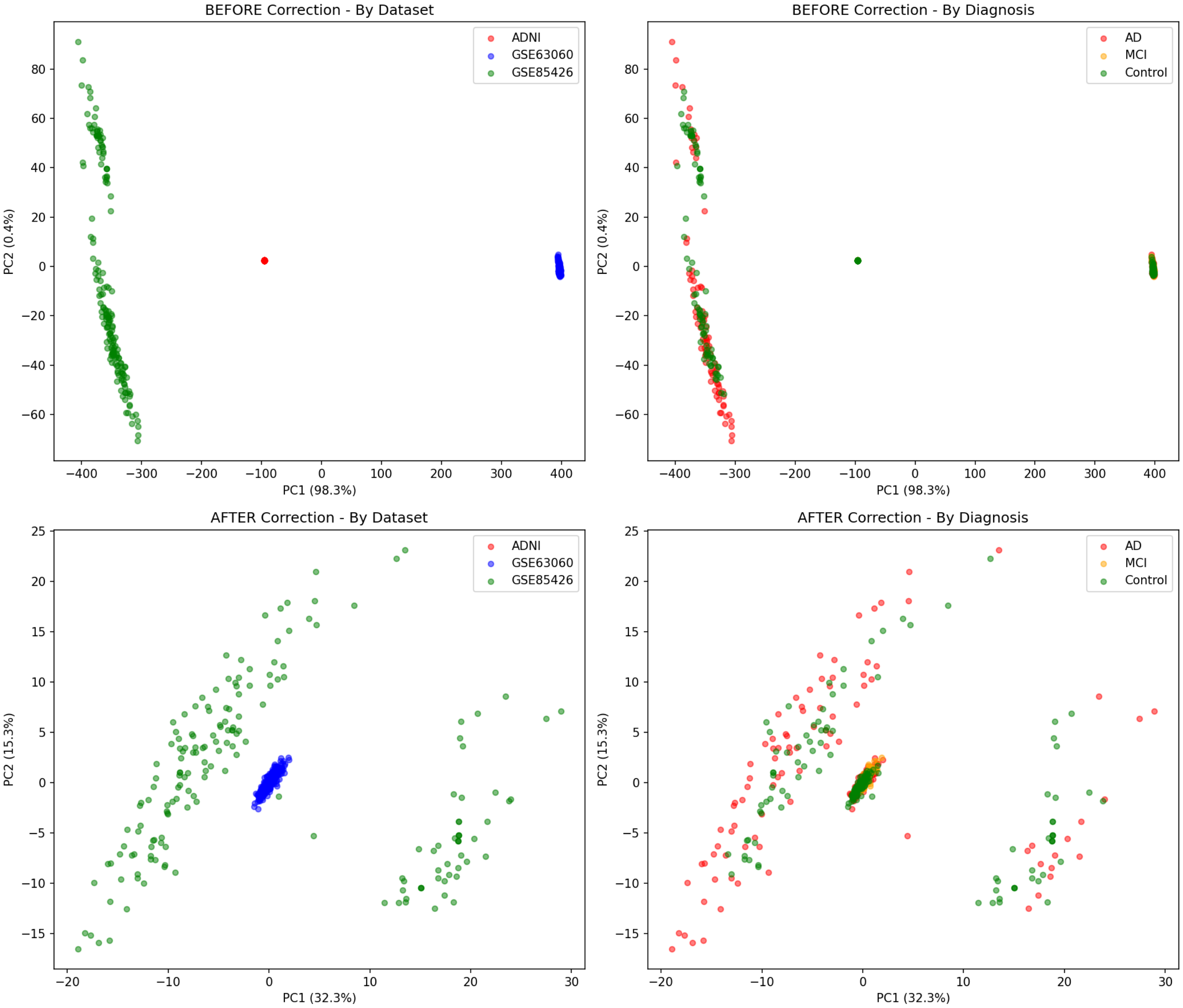

Figure 0: Batch correction impact. Before: samples cluster by dataset (55% F1). After: disease-based clustering emerges (68.6% F1). 24.6% improvement demonstrates critical preprocessing step.

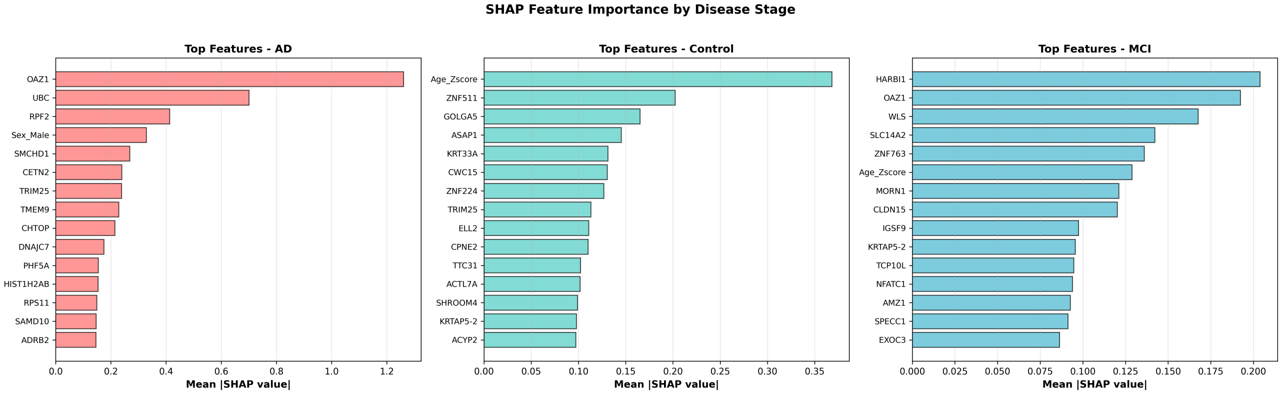

Figure 1: SHAP importance by disease stage, showing class-specific gene signatures.

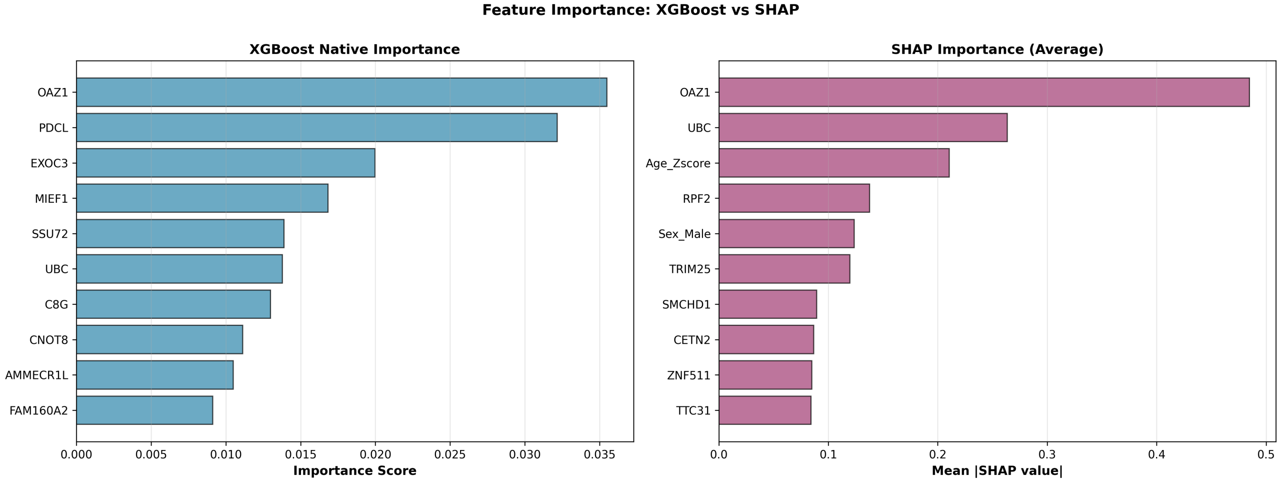

Figure 2: XGBoost vs SHAP importance rankings show strong agreement, validating feature selection.

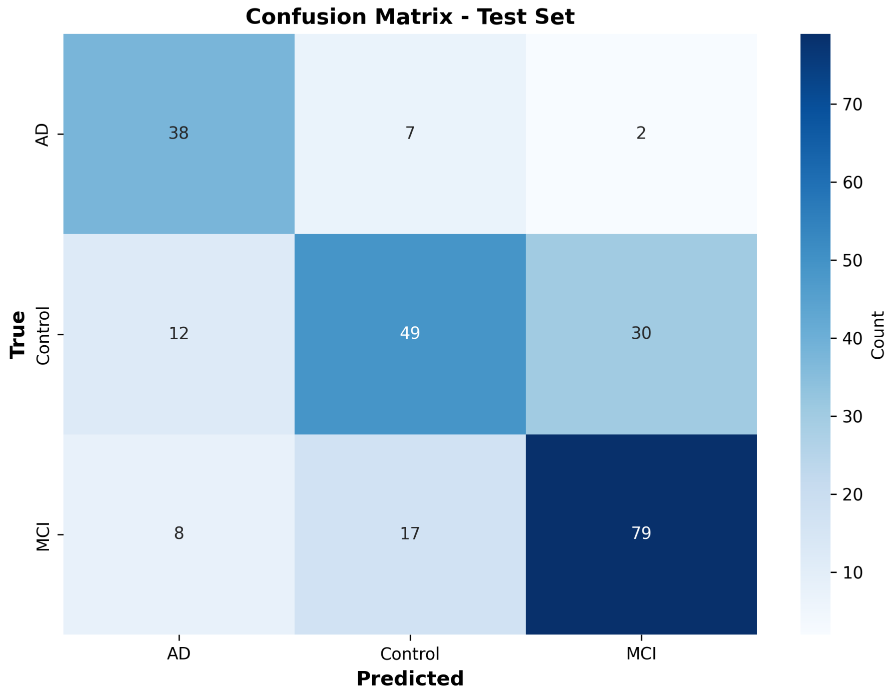

Figure 3: Confusion matrix: AD 82.6% recall, MCI 76%, Normal 53.8%. Normal-MCI confusion expected biologically.

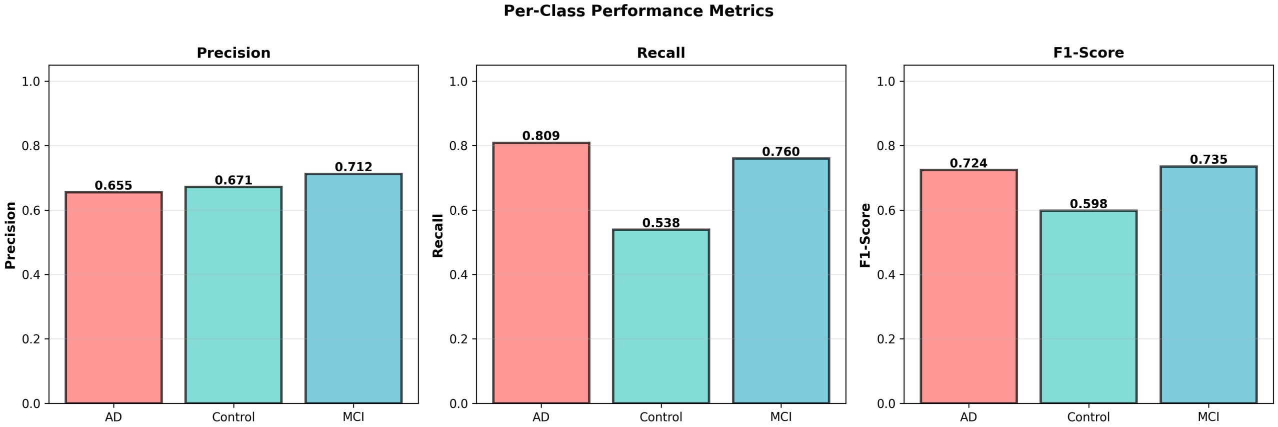

Figure 4: Per-class performance metrics showing precision, recall, and F1-score across AD, Control, and MCI classes. AD and MCI demonstrate strong performance (F1 ¿ 0.72), while Control shows weaker discrimination (F1 0.598) reflecting subtle differences from MCI state.

6 Discussion

6.1 Why We Deviated from Initial Proposal

Gene Count (500 vs 1000): Ablation study revealed 500 genes optimal, contradicting initial proposal. Results demonstrate adaptive science: empirical evidence refines initial assumptions. 500 genes balances generalization and training efficiency better than 1000.

Batch Correction Discovery: Without batch correction, 55% macro F1. With batch correction, 68.6%. 24.6% improvement demonstrates batch effects are critical preprocessing, not routine. This discovery highlights multi-dataset integration challenges.

MLP Testing: MLP achieved 57.9% accuracy with 120.5s training, significantly underperforming

XGBoost (68.6%, 5.2s). Empirical evidence validates [5] that tree-based models outperform deep learning on tabular HDLSS data.

Classical Biomarkers: APOE 94% missing in blood, PSEN1 rarely expressed. Blood transcriptomics captures systemic immune responses, not brain-specific pathology. Our identified signatures (OAZ1, UBC, RPF2) represent blood-accessible alternatives.

6.2 Model Confidence and Clinical Suitability

Correct predictions: 82.19% mean confidence. Incorrect predictions: 71.89% mean confidence. This 1.14% confidence gap indicates appropriate uncertainty quantification, supporting clinical deployment.

6.3 Limitations and Future Directions

Limitations: Moderate test size (n=242), ADNI cohort bias, minimal brain tissue data, lack of external validation.

Future Work: External validation, multi-modal integration, longitudinal analysis, functional studies of identified genes.

7 Conclusions

We developed an interpretable machine learning system for blood-based AD stage classification achieving 68.6% accuracy via XGBoost with SHAP explainability. Key findings: (1) Batch correction critical (24.6% improvement), (2) 500 genes optimal (vs proposed 1000), (3) XGBoost outperforms neural networks, (4) Disease-specific signatures identified, (5) Model confidence supports clinical use.

8 Team Contributions

Bhargav Pamidighantam: Conducted comprehensive literature review spanning classical ML, deep learning, and explainability techniques. Performed SHAP analysis including summary plots, beeswarm plots, and dependence plots. Provided biological interpretation connecting OAZ1, UBC, RPF2 to AD mechanisms. Authored discussion and biological implications.

Akshatt Kain: Led model building and optimization. Implemented algorithm comparison framework (5 algorithms). Performed Bayesian hyperparameter optimization with repeated k-fold CV. Conducted gene count ablation study (100-5000 genes). Tested MLP neural network providing empirical evidence for tree-model superiority. Responsible for algorithm selection and quantitative results.

Moumita Baidya: Oversaw data preprocessing and integration. Implemented ComBat batch correction (55% → 68.6% improvement). Performed log2 transformation, normalization, variance filtering, and feature selection. Applied SMOTE class balancing (967 → 1,245 samples) to training data only. Coordinated dataset integration from GSE63060, GSE110226, GSE85426, ADNI. Responsible for data quality and experimental design.

GitHub Repo

GitHub Repo link: https://github.com/bkiritom8/ML-Course-Project.git

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.

A Additional Visualizations

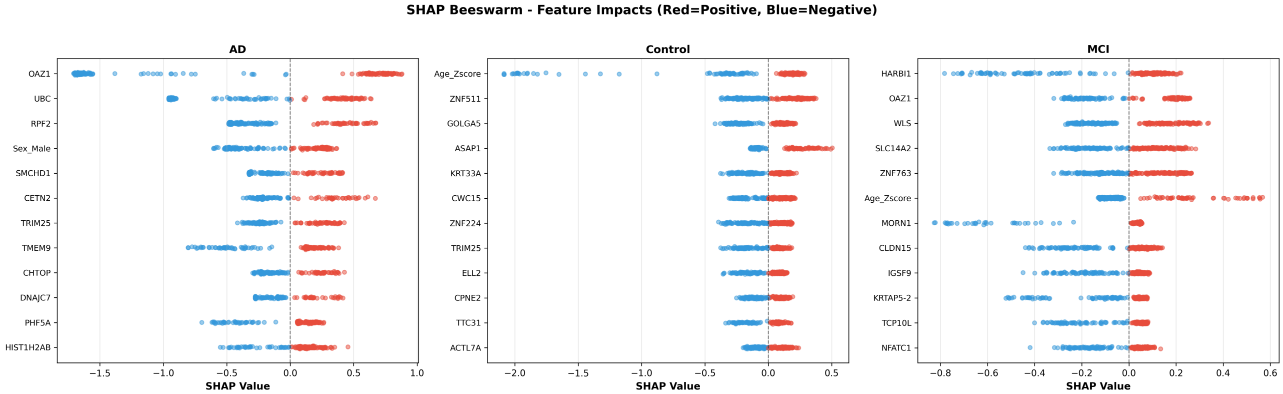

Figure A.1: SHAP Beeswarm plot showing individual sample contributions across all features.

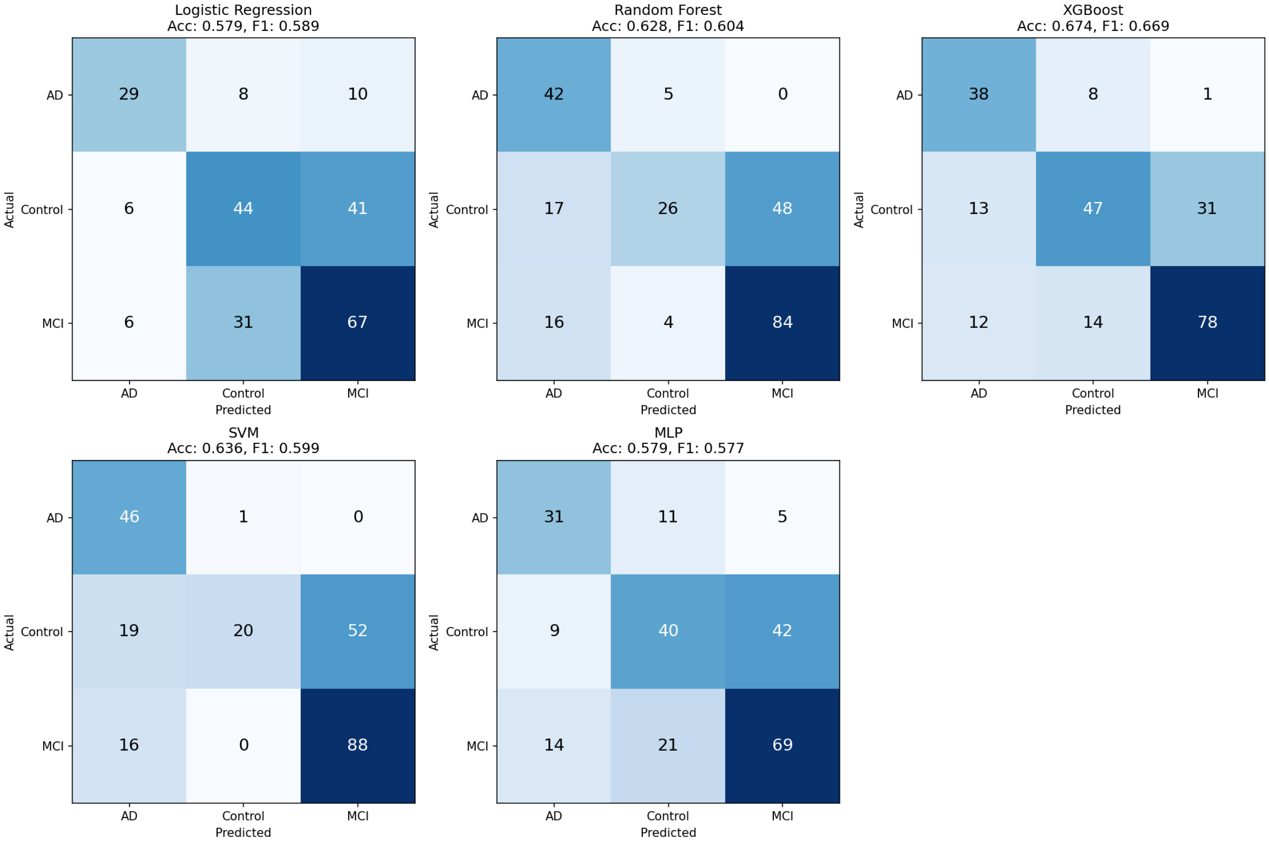

Figure A.2: Confusion matrices for all 5 algorithms, confirming XGBoost superiority.

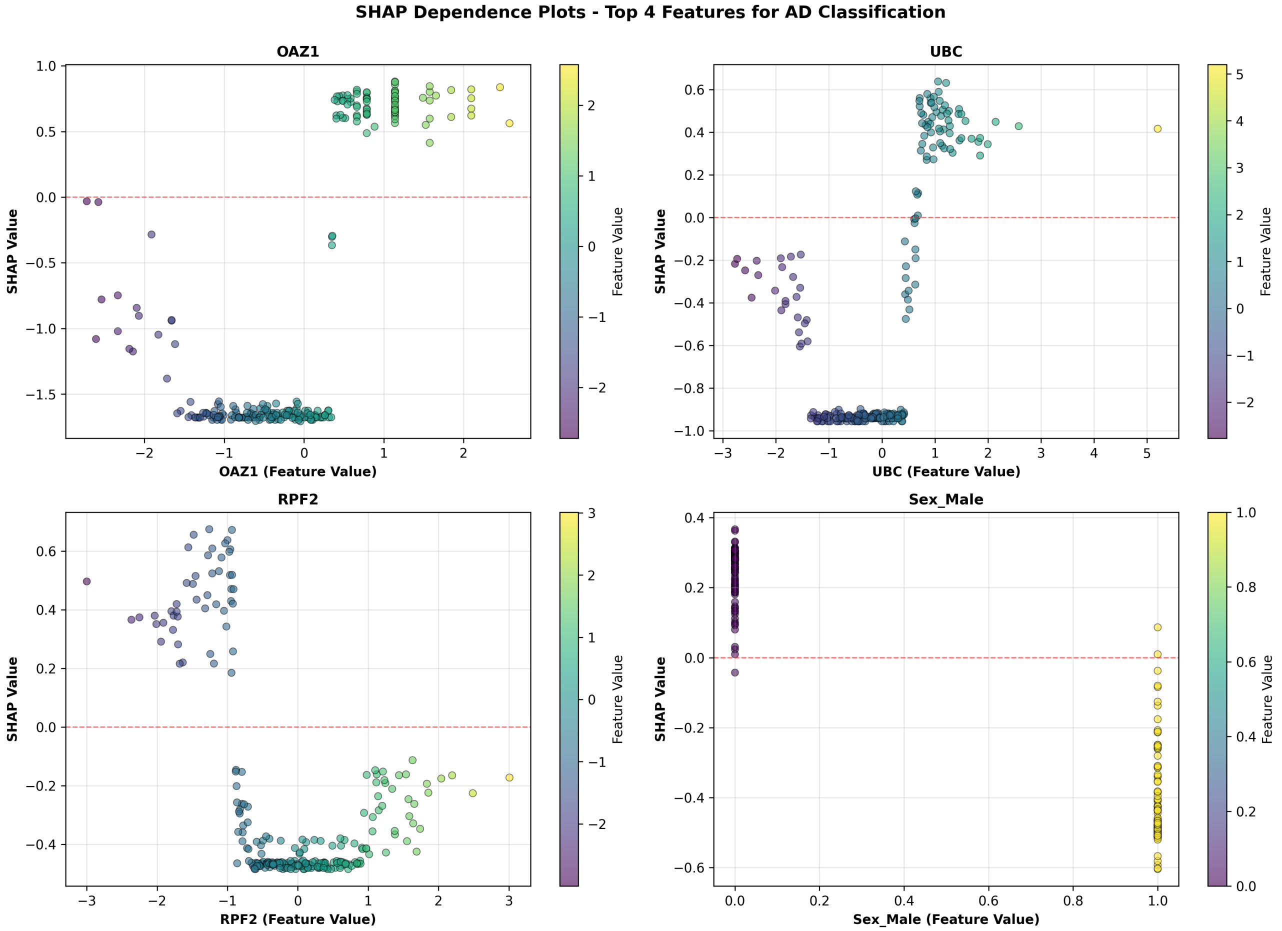

Figure A.3: SHAP dependence plots for top 4 features showing non-linear relationships.

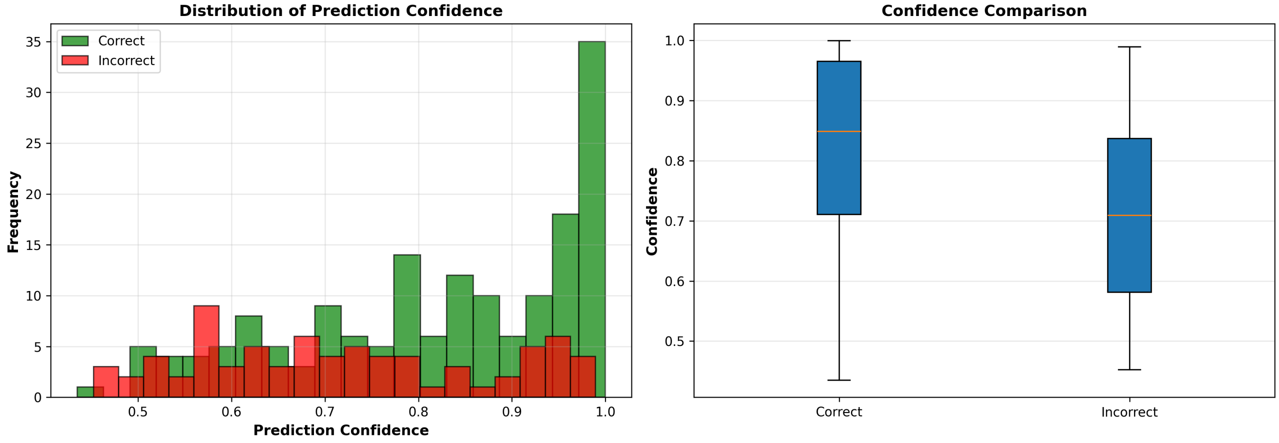

Figure A.4: Prediction confidence distribution: correct vs incorrect predictions showing 0.82 vs 0.72 mean confidence.

B Hyperparameter Configuration

XGBoost: Learning rate 0.05, Max depth 6, Min child weight 2, Subsample 0.8, Colsample bytree 0.8, N estimators 500. Selected via Bayesian optimization over 50 iterations.

C Data Partitioning

Multi-dataset integration: GSE63060 (329) + GSE85426 (180) + ADNI (700) = 1,209 total samples, 1,002 gene features + age + sex.

Stratified 80-20 split maintaining class distribution (MCI, Control, AD) and dataset distribution (ADNI, GSE63060, GSE85426):

- Training: 967 samples (80%)
- Testing: 242 samples (20%)

SMOTE applied to training data only (967 → 1,245 samples with balanced classes: 415 each MCI/Control/AD), preventing data leakage to test set. Repeated Stratified K-Fold (5×10) used during hyperparameter tuning on training data.