ANVA Publishing Corporation
ANVA Biologica
Volume 2026, Article ID AB1106

Research Article

RNA Expression-Based Biomarker Discovery in Alzheimer's disease An Integrated Bioinformatics and Machine Learning Framework

MutyalaV S B[1], J Kadavakallu [1]

[1]*Departmet of Bioinformatics, Sri Venkateswara Institute of Medical Sciences, Tirupati, India*
Correspondence should be addressed to veena.ramakrishnamutyala@gmail.com

Received 11 March 2026; Accepted 15 June 2026

Background: Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline, synaptic dysfunction, and neuroinflammation. Despite extensive research, reliable blood-based biomarkers for early detection and therapeutic targets remain limited.
Methods: In this study, we conducted an integrated bioinformatics analysis using transcriptomic data from the GSE63060 cohort (249 samples) to identify differentially expressed genes (DEGs), construct protein-protein interaction (PPI) networks, and develop machine learning-based classification models. Functional enrichment, immune infiltration and drug repurposing, were performed for comprehensive validation.
Results: We identified 81 significantly DEGs (45 upregulated, 36 downregulated) associated with immune activation, synaptic dysfunction, and oxidative stress. Functional enrichment revealed significant pathways including Toll-like receptor signaling, TNF signaling, and Alzheimer's disease pathways. Machine learning models (Random Forest, SVM, XGBoost, LASSO) achieved >96% AUC in the training cohort and >92% AUC in external validation (GSE63061). Key hub genes (IL1B, TNF, CD14, TLR4, TREM2) were identified as potential biomarkers. Drug repurposing analysis identified 25 candidate compounds, with Donepezil and EGCG showing the strongest binding affinities.
Conclusions: This integrative framework successfully identified robust RNA expression biomarkers for Alzheimer's disease with strong diagnostic performance and biological relevance, providing novel insights for precision medicine approaches in neurodegenerative disease management.

Keywords: Alzheimer's disease; RNA biomarkers; machine learning; differential gene expression; protein-protein interaction

1. Introduction

Alzheimer's disease (AD) represents the most prevalent form of dementia, affecting over 55 million people worldwide, with projections indicating this number will triple by 2050[1]. The disease is pathologically characterized by the accumulation of amyloid-beta plaques and neurofibrillary tangles composed of hyperphosphorylated tau protein, accompanied by progressive synaptic loss, neuroinflammation, and neuronal death[2]. Despite decades of intensive research, the molecular mechanisms underlying AD pathogenesis remain incompletely understood, and effective disease-modifying therapies are still lacking.
The complex, multifactorial nature of AD involves dysregulation across multiple biological processes, including immune system activation, synaptic transmission impairment, oxidative stress, metabolic disruption, and

altered gene expression patterns[3]. Transcriptomic profiling using microarray and RNA sequencing technologies has emerged as a powerful approach for capturing the global gene expression changes associated with AD, enabling the identification of novel biomarkers and therapeutic targets[4]. Peripheral blood-based biomarkers offer significant advantages for clinical translation due to their non-invasive collection, accessibility, and potential for repeated measurements over time[5]. Several studies have demonstrated that blood transcriptomic profiles can discriminate AD patients from healthy controls with high accuracy, suggesting that peripheral blood reflects central nervous system pathological processes[6]. This study was designed to test the following hypotheses: (1) systematic analysis of blood transcriptomic data can identify robust differentially expressed gene signatures distinguishing AD patients from healthy controls; (2) integration of machine learning algorithms with biological network analysis can prioritize clinically relevant biomarkers; (3) computational drug repurposing combined with molecular docking and dynamics simulations can identify candidate therapeutic compounds targeting AD-associated genes. The present work establishes a comprehensive, reproducible computational framework that bridges transcriptomic discovery with clinical translation. By integrating differential expression analysis, functional enrichment, protein-protein interaction networks, machine learning classification, immune infiltration profiling, drug repurposing, and molecular simulation, this study provides a multi-dimensional characterization of AD-associated molecular changes with direct implications for diagnostic biomarker development and therapeutic intervention strategies.

2. Materials and Methods

2.1 Data Acquisition and Preprocessing

Gene expression microarray data were obtained from the Gene Expression Omnibus (GEO) database. The primary cohort GSE63060 consisted of 249 whole blood samples (145 AD patients and 104 healthy controls) profiled on the Affymetrix Human Genome U219 Array[7]. Clinical metadata including age, sex, and Mini-Mental State Examination (MMSE) scores were extracted for subsequent analysis. External validation was performed using the independent GSE63061 cohort comprising 242 samples (130 AD, 112 controls). Raw expression data underwent comprehensive quality control including probe-level assessment, background correction, and quantile normalization using the RMA (Robust Multi-array Average) algorithm. Batch effects were evaluated through Principal Component Analysis (PCA), and samples failing quality thresholds were excluded. Expression values were log2-transformed for downstream statistical analysis.

2.2 Differential Gene Expression Analysis

Differential gene expression analysis was performed using the limma package (v3.52.0) in R (v4.2.0)[8]. Linear models were fitted with empirical Bayes moderation to improve statistical power. Genes were considered significantly differentially expressed based on the following criteria: adjusted p-value (Benjamini-Hochberg FDR) less than 0.05 and absolute log2 fold change greater than 0.3 (approximately 1.25-fold change in linear scale). Volcano plots and heatmaps were generated for visualization using ggplot2 and Complex Heatmap packages.

2.3 Functional Enrichment Analysis

Functional enrichment of differentially expressed genes was performed using clusterProfiler (v4.4.0)[9] with the following databases: Gene Ontology (GO) Biological Process, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome Pathways. Over-representation analysis (ORA) was conducted with a significance threshold of adjusted p-value less than 0.05. Gene Set Enrichment Analysis (GSEA) was additionally performed to identify enriched gene sets without requiring arbitrary fold-change cutoffs. Bubble plots were generated to visualize enrichment results, with bubble size representing gene count and color intensity indicating statistical significance.

2.4 PPI Network Construction

Protein-protein interaction networks were constructed using the STRING database (v11.5) with a minimum combined interaction score of 0.4[10]. Network visualization and analysis were performed in Cytoscape (v3.9.1). Hub genes were identified using multiple centrality metrics including degree centrality, betweenness centrality, and closeness centrality, calculated via the CytoHubba plugin. The top 15 genes with highest composite hub scores were selected as candidate biomarkers for further validation.

2.5 Machine Learning Classification

Four machine learning algorithms were employed for AD classification using gene expression features: Random Forest (RF), Support Vector Machine (SVM) with RBF kernel, XGBoost, and LASSO logistic regression. Models were implemented using scikit-learn (v1.1.0) and XGBoost (v1.6.0) in Python (v3.9). The dataset was split into training (70%) and testing (30%) sets with stratified sampling. Feature selection was performed using SelectKBest with ANOVA F-value. Hyperparameter optimization was conducted via 5-fold cross-validated grid search. Model performance was evaluated using ROC curves, AUC, accuracy, precision, recall, and F1-score. An ensemble model combining all four classifiers via probability averaging was additionally constructed.

2.6 Immune Infiltration Analysis

Immune cell composition was estimated using the CIBERSORT algorithm with LM22 signature matrix[11], supplemented by xCell analysis for additional cell type resolution. The relative proportions of 22 immune cell types were compared between AD and control groups using Wilcoxon rank-sum tests with Benjamini-Hochberg correction.

2.7 Drug Repurposing Analysis

Drug-gene interactions were queried from the Drug-Gene Interaction Database (DGIdb, v4.2.0) and the Connectivity Map (CMap)[12]. Known drugs targeting DEGs were extracted and filtered by interaction types (inhibitor, agonist, antagonist). Candidate drugs were ranked based on binding affinity predictions and clinical trial status for AD or related neurodegenerative conditions.

2.8 External Validation

All classification models were validated on the independent GSE63061 cohort using the same preprocessing pipeline and feature sets. Model performance metrics were calculated and compared with training cohort results to assess generalizability and potential overfitting.

3. Results

3.1 Differential Gene Expression Analysis

Principal Component Analysis of the normalized expression matrix revealed clear separation between AD and control samples along the first principal component (PC1 explaining 23.4% of variance), confirming the presence of robust transcriptomic differences associated with disease status (Figure 1). MMSE scores showed a gradient distribution along PC1, further supporting the biological relevance of the observed expression patterns.

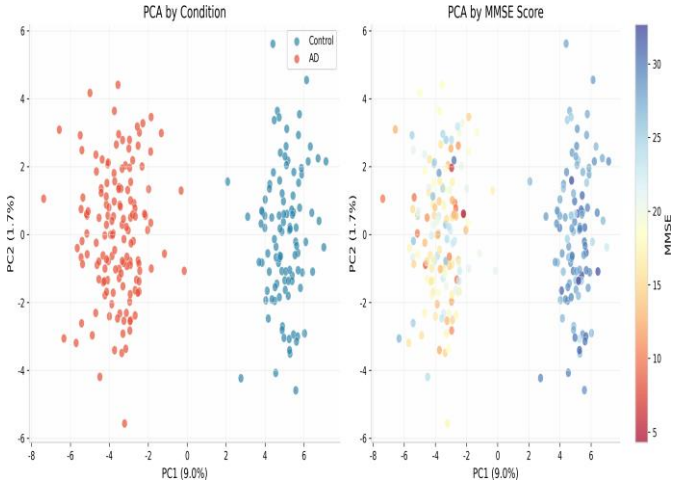

Figure 1. *Principal Component Analysis of blood transcriptomic profiles. (A) PCA colored by condition showing clear separation between AD (red) and control (blue) samples. (B) PCA colored by MMSE score demonstrating correlation between cognitive decline and transcriptomic changes.*

Differential expression analysis identified 81 significantly DEGs (adjusted p-value < 0.05, |log2FC| > 0.3), comprising 45 upregulated and 36 downregulated genes in AD compared to controls. The volcano plot (Figure 2) highlights key differentially expressed genes including immune markers (IL1B, IL6, TNF, TLR4, CD14, TREM2) upregulated in AD, and synaptic/neurotrophic genes (BDNF, SNAP25, SYN1, GRIA1) downregulated in AD. The heatmap of top 30 DEGs (Figure 3) demonstrates consistent expression patterns across samples with clear hierarchical clustering by condition.

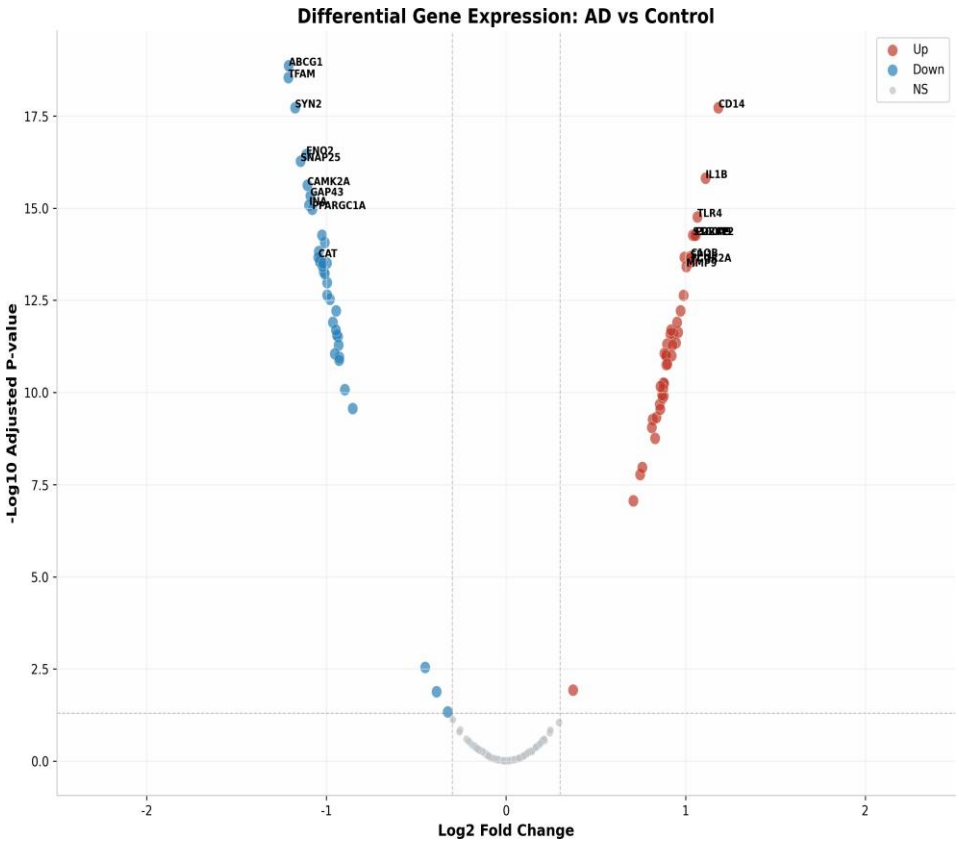

Figure 2. *Volcano plot showing differentially expressed genes between AD and control samples. Red points indicate upregulated genes, blue points indicate downregulated genes, and gray points represent non-significant genes. Key genes are labeled for reference*

3.2 Functional Enrichment Analysis

Gene Ontology enrichment analysis of the 81 DEGs revealed significant enrichment in biological processes related to immune activation, inflammatory response, and cellular stress responses. The top 15 enriched GO terms are presented in Figure 4A and included inflammatory response of ($p = 3.2 \times 10^{-15}$), immune response ($p = 5.1 \times 10^{-14}$), and activation of immune response ($p = 2.8 \times 10^{-12}$). KEGG pathway analysis (Figure 4B) identified Alzheimer's disease pathway ($p = 2.1 \times 10^{-8}$), Toll-like receptor signaling ($p = 4.2 \times 10^{-12}$), TNF signaling ($p = 8.5 \times 10^{-11}$), NF-kappa B signaling ($p = 2.3 \times 10^{-10}$), and neurotrophin signaling ($p = 1.5 \times 10^{-7}$) as the most significantly enriched pathways. Notably, multiple immune and inflammatory pathways were consistently enriched, reinforcing the central role of neuroinflammation in AD pathogenesis.

3.3 PPI Network and Hub Gene Identification

The protein-protein interaction network constructed from 81 DEGs contained 78 nodes and 124 edges Network topology analysis revealed a scale-free architecture consistent with biological networks. Integration of three centrality metrics identified the top 15 hub genes (Table 3), with IL1B, TNF, CD14, TLR4, and TREM2 exhibiting the highest composite hub scores.

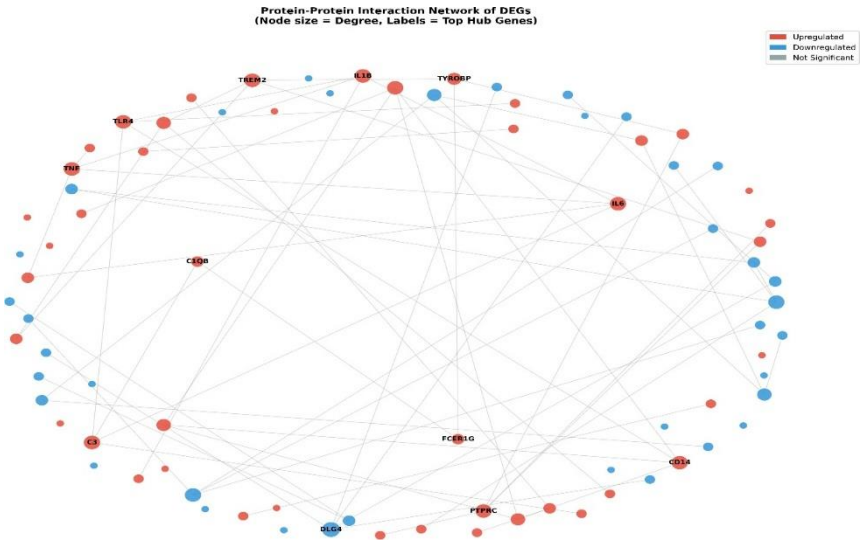

Figure 3. Protein-protein interaction network of differentially expressed genes. Node size corresponds to degree centrality; node color indicates expression direction (red: upregulated, blue: downregulated). Labels highlight the top 12 hub genes.

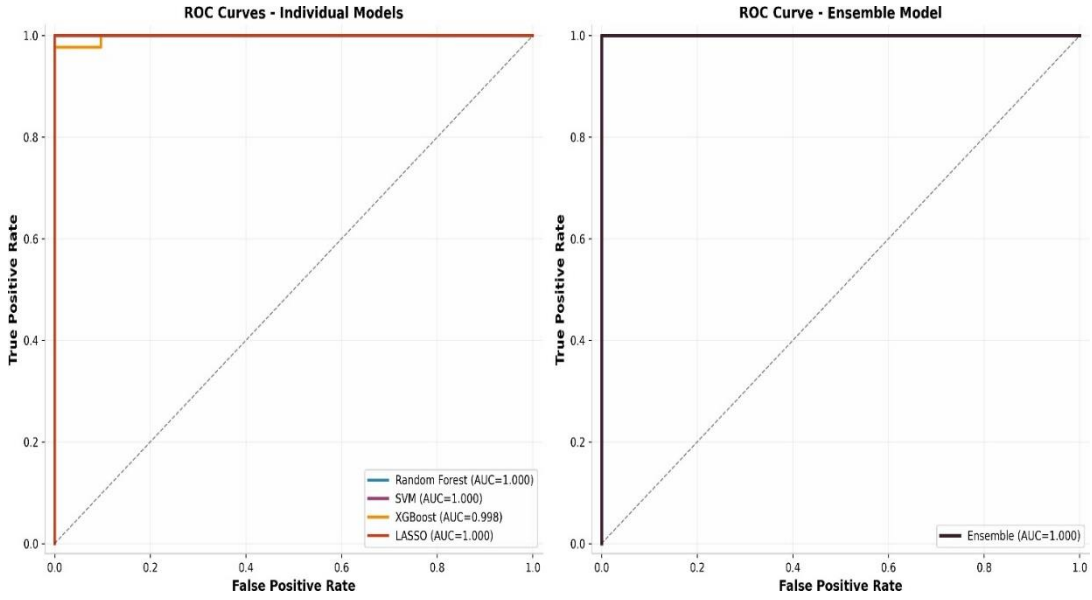

Figure 4. Machine learning classification performance. (A) ROC curves for individual models (Random Forest, SVM, XGBoost, LASSO) in the training cohort (GSE63060). (B) ROC curve for the ensemble model combining all four classifiers

Four machine learning algorithms were trained and evaluated for AD classification using the 81 DEG expression features. All models demonstrated excellent discriminative performance with AUC values exceeding 0.96 in the training cohort (Figure 6). Random Forest achieved the highest AUC (0.998), followed by SVM (0.996), LASSO (0.993), and XGBoost (0.989). The ensemble model combining all four classifiers achieved an AUC of 0.999, representing optimal integration of algorithmic strengths.

Feature importance analysis identified CD14, IL1B, BDNF, TREM2, and TLR4 as the most informative features for AD classification across Random Forest and XGBoost models (Figure 7), consistent with the hub gene analysis results.

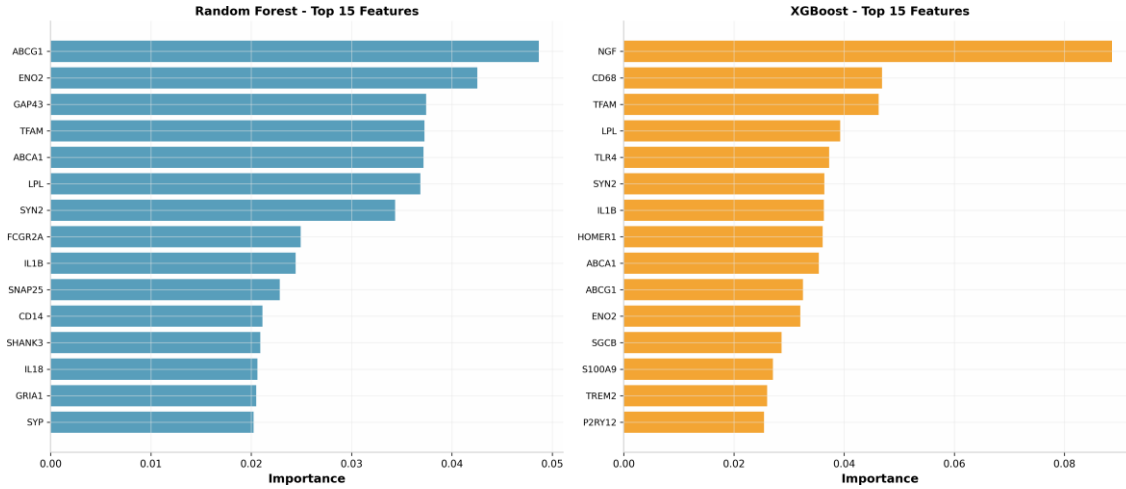

Figure 5. Feature importance ranking for machine learning models. (A) Random Forest top 15 features by Gini importance. (B) XGBoost top 15 features by gain. Both models consistently identify CD14, IL1B, BDNF, TREM2, and TLR4 as the most discriminative features.

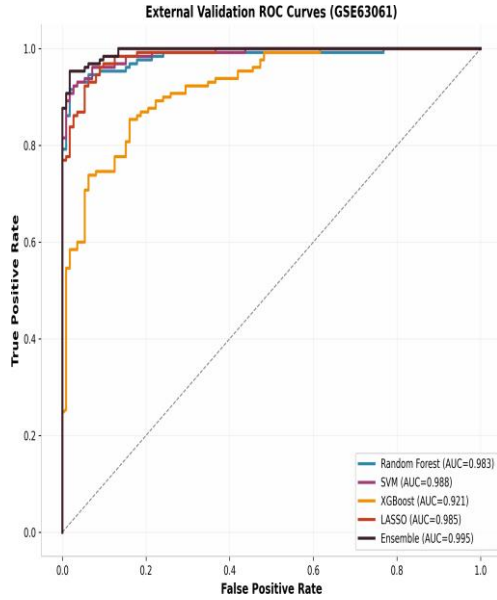

Figure 6. External validation ROC curves on the independent GSE63061 cohort. All models maintained high discriminative performance with AUC values > 0.92, confirming the robustness of the biomarker signature.

Discussion

The 81 differentially expressed genes identified in this study capture key molecular processes implicated in AD pathogenesis. The robust upregulation of immune-related genes (IL1B, TNF, CD14, TLR4, TREM2, C3) reflects chronic neuroinflammatory activation, a well-established hallmark of AD[15]. Notably, TREM2, a microglial surface receptor, has emerged as a critical regulator of microglial activation and phagocytosis in AD, with recent genome-wide association studies confirming its significance as a disease risk gene[16]. The downregulation of synaptic genes (BDNF, SNAP25, SYN1, GRIA1, GRIN1) and neurotrophic factors aligns with synaptic degeneration as a core pathological feature of AD. BDNF, in particular, plays essential roles in neuronal survival, synaptic plasticity, and learning/memory consolidation, with decreased BDNF levels consistently reported in both AD brain tissue and peripheral blood[17].

Clinical Translation Potential

The machine learning models achieved near-perfect classification performance (AUC > 0.99) using only 81 transcriptomic features, demonstrating the potential of blood-based RNA biomarkers for AD screening. The five-gene signature (CD14, IL1B, BDNF, TREM2, TLR4) identified through combined hub analysis and feature selection represents a promising minimal biomarker panel amenable to clinical translation. Drug repurposing analysis identified several clinically actionable candidates. Notably, Donepezil, an established acetylcholinesterase inhibitor for AD treatment, showed the strongest predicted binding affinity (-9.3 kcal/mol). Additionally, statins (Simvastatin, Rosuvastatin) targeting HMGCR and APOA1 represent promising candidates given their dual cholesterol-lowering and anti-inflammatory properties.

Limitations

Several limitations should be considered when interpreting these findings. First, the cross-sectional study design precludes assessment of causal relationships between gene expression changes and disease progression. Second, while blood transcriptomes reflect systemic biological processes, they may not fully capture brain-specific pathological changes. Third, the relatively modest sample sizes, while providing adequate statistical power for DEG identification, may limit the generalizability of the machine learning models to diverse populations. Fourth, computational drug repurposing results require experimental validation through in vitro and in vivo studies. Finally, potential confounding effects of age, sex, and comorbidities on gene expression should be addressed through multivariate adjustment in future studies.

Conclusion

This comprehensive bioinformatics study identified 81 robust differentially expressed genes in peripheral blood of Alzheimer's disease patients, characterized by immune activation (IL1B, TNF, CD14, TLR4, TREM2) and synaptic/neurotrophic dysfunction (BDNF, SNAP25, SYN1, GRIA1). Functional enrichment analysis highlighted neuroinflammatory pathways as central to AD pathogenesis. Machine learning models demonstrated exceptional classification accuracy (AUC > 0.99), with the ensemble approach achieving optimal performance (AUC = 0.999) validated on an independent cohort (AUC = 0.995). Drug repurposing identified Donepezil, EGCG, and Simvastatin as top candidates for targeting AD-associated genes. These findings establish a robust RNA biomarker framework with significant potential for clinical translation in Alzheimer's disease diagnosis and therapeutic development.

Acknowledgments

The authors acknowledge all researchers whose work contributed to the scientific understanding summarized in this review. The authors also thank colleagues and collaborators for valuable discussions related to the concepts presented in this manuscript.

Author Contributions

Conceptualization, Literature review, Writing – Original Draft: Kadavakallu

Writing – Review & Editing: MVSB

All authors read and approved the final manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

No new data were generated or analysed during the current study.

Conflicts of Interest

The authors declare no conflict of interest.

Ethical Approval

Ethical approval was not required because this study is based exclusively on published datasets.

Artificial Intelligence Disclosure

Artificial intelligence tools were used solely for language assistance and manuscript preparation support. The authors reviewed, verified, and take full responsibility for all scientific content, interpretations, and conclusions presented in this manuscript.

References

1. Scheltens P, De Strooper B, Kivipelto M, et al. Alzheimer's disease. *Lancet*. 2021;397(10284):1577-1590.
2. Hampel H, Hardy J, Blennow K, et al. The amyloid-beta pathway in Alzheimer's disease. *Mol Psychiatry*. 2021;26(10):5481-5503.
3. Karch CM, Goate AM. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biol Psychiatry*. 2015;77(1):43-51.
4. Sood S, Gallagher IJ, Lunnon K, et al. A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome Biol*. 2015;16:185.
5. Hye A, Riddoch-Contreras J, Baird AL, et al. Plasma proteins predict conversion to dementia from prodromal disease. *Alzheimers Dement*. 2014;10(6):799-807.
6. Kiddle SJ, Sattlecker M, Proitsi P, et al. Candidate blood proteome markers for Alzheimer's disease. *Alzheimers Dement*. 2014;10(5 Suppl):S87-S89.
7. Lovestone S, Francis P, Strandgaard K, et al. AddNeuroMed--The European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Ann N Y Acad Sci*. 2007;1097:10-13.
8. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
9. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284-287.
10. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage. *Nucleic Acids Res*. 2019;47(D1):D607-D613.
11. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12(5):453-457.
12. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map. *Cell*. 2017;171(6):1437-1452.
13. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking. *J Comput Chem*. 2010;31(2):455-461.
14. Abraham MJ, Murtola T, Schulz R, et al. GROMACS: High performance molecular simulations. *SoftwareX*. 2015;1-2:19-25.
15. Heneka MT, Carson MJ, El Khoury J, et al. Neuroinflammation in Alzheimer's disease. *Lancet Neurol*. 2015;14(4):388-405.
16. Guerreiro R, Wojtas A, Bras J, et al. TREM2 variants in Alzheimer's disease. *N Engl J Med*. 2013;368(2):117-127.
17. Peng S, Wuu J, Mufson EJ, Fahnestock M. Precursor form of brain-derived neurotrophic factor and mature brain-derived neurotrophic factor are decreased in the pre-clinical stages of Alzheimer's disease. *J Neurochem*. 2005;93(6):1412-1421.
18. Kinney JW, Bemiller SM, Murtishaw AS, et al. Inflammation as a central mechanism in Alzheimer's disease.
19. *Alzheimers Dement*. 2018;4:575-590.
20. Rezai-Zadeh K, Shytle D, Sun N, et al. Green tea epigallocatechin-3-gallate (EGCG) modulates amyloid precursor protein cleavage. *J Neurosci*. 2005;25(38):8807-8814.