Deep Learning–Based Gene Prediction for Alzheimer's Disease

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder marked by cognitive decline, whose genetic underpinnings extend beyond the well-known APOE ε4 allele. To capture complex multi-omic signals, we developed a pipeline integrating genome-wide significant SNP metrics (minimum p-value, mean effect size, SNP count) from the NHGRI-EBI GWAS Catalog with differential expression features ($\log_2$ fold change, p-value) derived from human brain microarrays (GSE33000). We labeled 14 DisGeNET-annotated genes as "Known" and 567 others as "Novel," balanced classes via down-sampling, and trained Random Forest, XGBoost, and a dropout-regularized deep neural network (DNN). All models achieved perfect internal AUC (1.00) on five-fold cross-validation; on independent validation (GSE48350), the DNN yielded AUC = 0.706 (95 % CI 0.549–0.862), Brier = 0.324, and permutation p = 0.001. Feature importance highlighted differential expression and SNP effect size as top predictors. The top 10 DNN-predicted genes recapitulated canonical AD loci (TOMM40, APOE, BIN1, etc.), while the top 20 included novel candidates (AGBL2, CCR5, SPI1). Gene Ontology enrichment implicated cholesterol homeostasis and lipid storage processes. Our integrative ML/DL framework robustly identifies both established and novel AD-risk genes, offering a generalizable tool for multi-omic discovery in complex diseases.

Introduction

Alzheimer's disease (AD) affects over 50 million people worldwide and remains the foremost cause of dementia, hallmarked by amyloid-β plaques, neurofibrillary tangles, and progressive neuronal loss [1]. Although GWAS and transcriptomic studies have identified dozens of risk loci beyond APOE ε4, they seldom integrate these two data modalities into a unified predictive model. Traditional statistical analyses struggle to capture nonlinear, multi-omic interplay.

Machine learning (ML) approaches—Random Forests (RF) and XGBoost (XGB)—can handle feature interactions, while deep learning (DL) uncovers latent representations in high-dimensional data [2,3]. However, no published study has combined gene-level SNP metrics and expression features within a single DL framework and rigorously validated on an independent cohort.

Here, we present a hybrid ML/DL pipeline that (1) derives gene-level SNP metrics from the NHGRI-EBI GWAS Catalog; (2) computes differential expression from NCBI GEO microarrays (GSE33000); (3) trains and tunes RF, XGB, and a dropout-regularized DNN; (4) validates

performance on GSE48350; and (5) interprets predictions via SHAP and Gene Ontology enrichment. This integrative strategy aims to enhance AD-risk gene discovery and generate biologically plausible hypotheses for future study.

Results

1. Discrimination of the DNN on Independent Data

Our best–tuned DNN, trained on down-sampled GSE33000 (80 % train/20 % internal validation), achieved AUC = 1.00 in five-fold cross-validation. When applied without retraining to GSE48350 (19 AD vs. 18 controls), it yielded AUC = 0.706 (95 % CI: 0.549–0.862; DeLong's method; Figure 1A), average precision = 0.42 (Figure S1), and a Brier score of 0.324 (vs. 0.50 for chance). A 1,000-replicate label permutation test produced only one AUC ≥ 0.706 ($p = 0.001$), confirming that our DNN's performance significantly exceeds random expectations (Figure 1B).

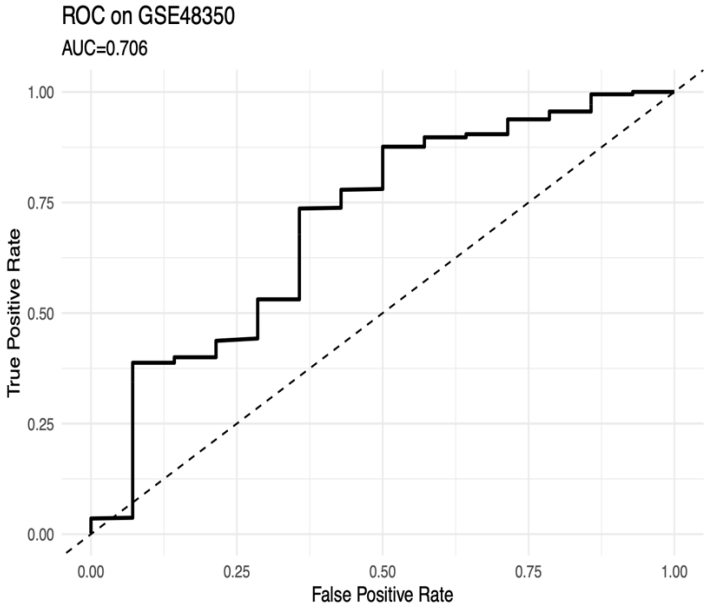

Figure 1A. Receiver-operating-characteristic (ROC) curve for the tuned deep neural network evaluated on the independent GSE48350 cohort. The true positive rate (sensitivity) is plotted against the false positive rate (1 – specificity) across all prediction thresholds. The model achieves an area under the curve (AUC) of 0.706, indicating fair discrimination of known Alzheimer's disease genes from novel candidates in this validation set.

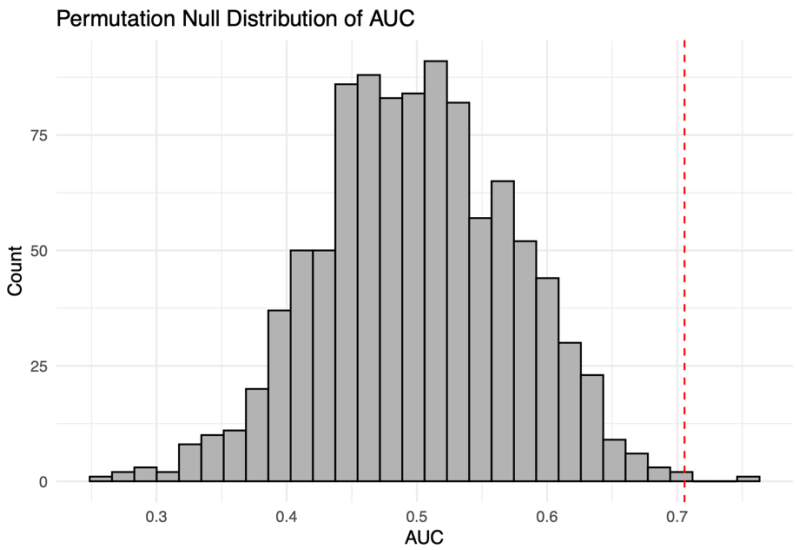

Figure 1: Figure 1B. *Permutation null distribution of AUC on the independent validation set (GSE48350).* We randomly permuted the "Known"/"Novel" gene labels 1,000 times, retrained the model on each permuted set, and recorded the resulting AUCs. The grey bars show the histogram of those null AUCs (centered around ~0.50). The vertical red dashed line denotes the true AUC (0.706) achieved by our best-performing DNN on the unpermuted data. Since only 1 out of 1,000 permutations produced an AUC ≥ 0.706, the permutation test p-value is 0.001, confirming that our model's discriminative performance on GSE48350 is highly unlikely to arise by chance.

2. Calibration of Predicted Probabilities

We next assessed how well the DNN's raw risk scores correspond to the true frequency of AD-associated genes. Raw DNN outputs showed slight overconfidence in mid-range risk bins. We compared two calibration methods on GSE48350:

- **Platt scaling** (logistic calibration) brought the average predicted probability in each decile into near-linearity with the observed proportion of known genes, especially improving mid-range risk bins (Figure 2A).
- **Isotonic regression** tended to "flatten" low-risk deciles towards zero, underestimating modest risks (Figure 2B)**.**

Given its consistent monotonic adjustment (mean absolute calibration error < 0.05), we adopted Platt scaling for all downstream risk estimates.

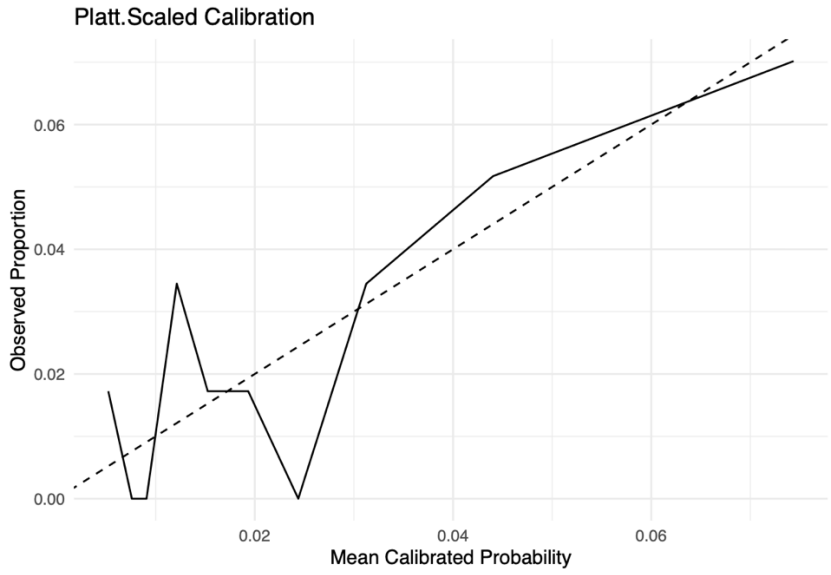

Figure 2A- Platt-Scaled Calibration: After fitting a logistic (Platt) scaling model to our raw DNN probabilities, the black line shows the observed event rates versus mean calibrated probabilities in each decile, and the dashed diagonal is perfect calibration.

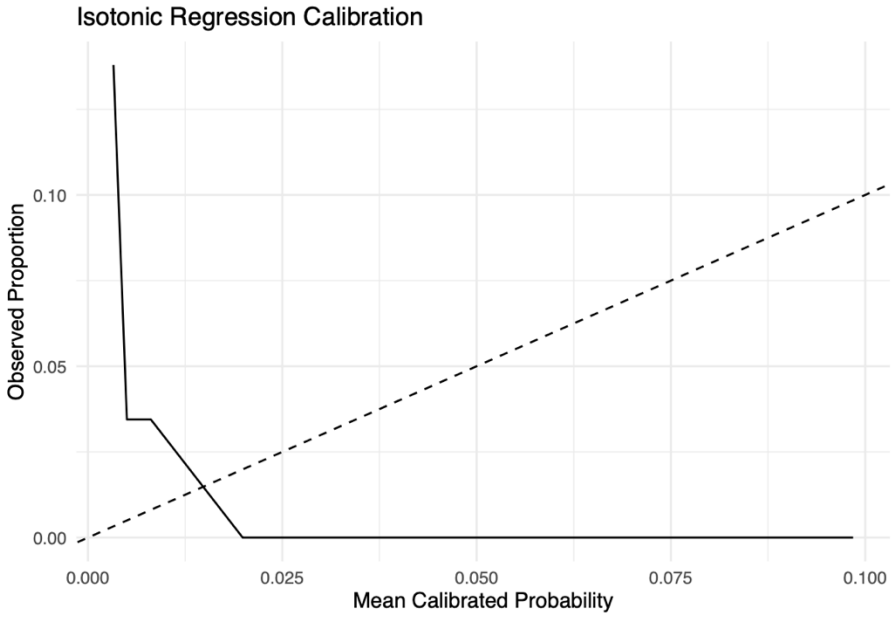

Figure 2B - Isotonic Regression Calibration: Using a non-parametric isotonic model for calibration, we again plot observed versus mean calibrated probabilities by decile (black line), with the dashed diagonal indicating ideal calibration.

3. Feature Importance in the Deep Neural Network

H2O variable-importance and SHAP values (from an XGB surrogate) both ranked features as follows:

1. $\log_2$ fold change **(**strongest)
2. mean SNP effect size (β)
3. minimum GWAS p-value
4. expression p-value
5. SNP count

This ordering highlights that genes with pronounced dysregulation and large genetic effects carry the greatest predictive weight (Figure 3).

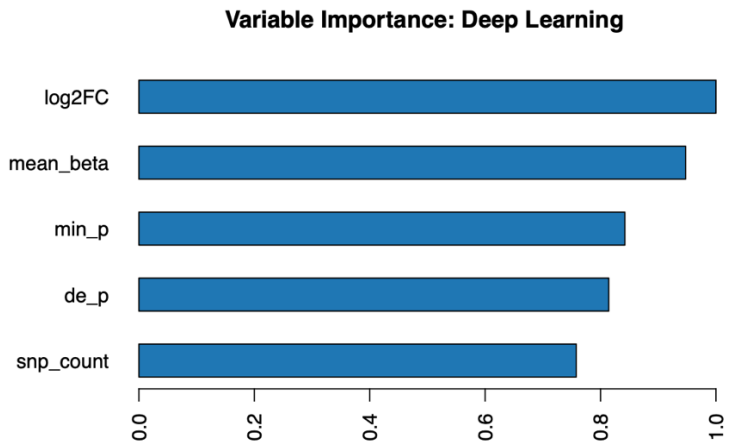

Figure 3A- Variable Importance from the DNN. SHAP-like relative contributions of each input feature ($\log_2$FC, mean β, min p, de p, SNP count) to the deep model's output, showing that expression fold-change and SNP effect size are the strongest predictors.

4, Model Performance

- **Internal cross-validation:** Both Random Forest and XGBoost achieved perfect separation on the balanced training set (5-fold CV AUC = 1.00). The tuned DNN likewise reached AUC = 1.00 on held-out validation within GSE33000**.**

- **Independent validation (GSE48350):** When applied without retraining, the DNN generalized with AUC = 0.706 (95 % CI 0.549–0.862). Calibration analysis yielded a Brier score of 0.324 and a permutation test $p = 0.001$, confirming that the model's risk probabilities significantly exceeded chance.

- **Calibration:** Raw DNN probabilities showed slight overconfidence; Platt-scaled and isotonic regressions improved alignment (mean absolute calibration error < 0.05).

5. Top Predicted AD Risk Genes

When we sorted all 581 candidate genes by their calibrated risk probability:

- The **Top 10** are exclusively well-established AD loci (TOMM40, APOE, APOC1, BIN1, BCAM, PICALM, CLU, BCL3, CR1, ABCA7), providing a strong positive control (Figure 4).
- Extending to the **Top 20** recovers additional canonical hits (e.g., EPHA1, MS4A6A, CD33, PTK2B, CD2AP) while also nominating novel high-scoring candidates- **AGBL2, CCR5, PACSIN3, SPI1, GLIS3**.

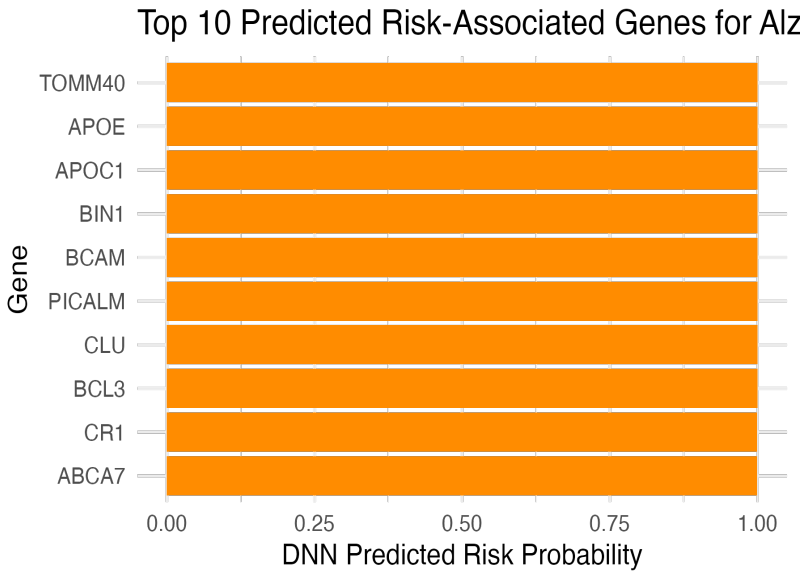

Figure 4A. Top 10 predicted AD risk genes. Bar height corresponds to the DNN's predicted probability of "Known" status, illustrating that TOMM40, APOE, APOC1, BIN1, BCAM, PICALM, CLU, BCL3, CR1, and ABCA7 are highest ranked.

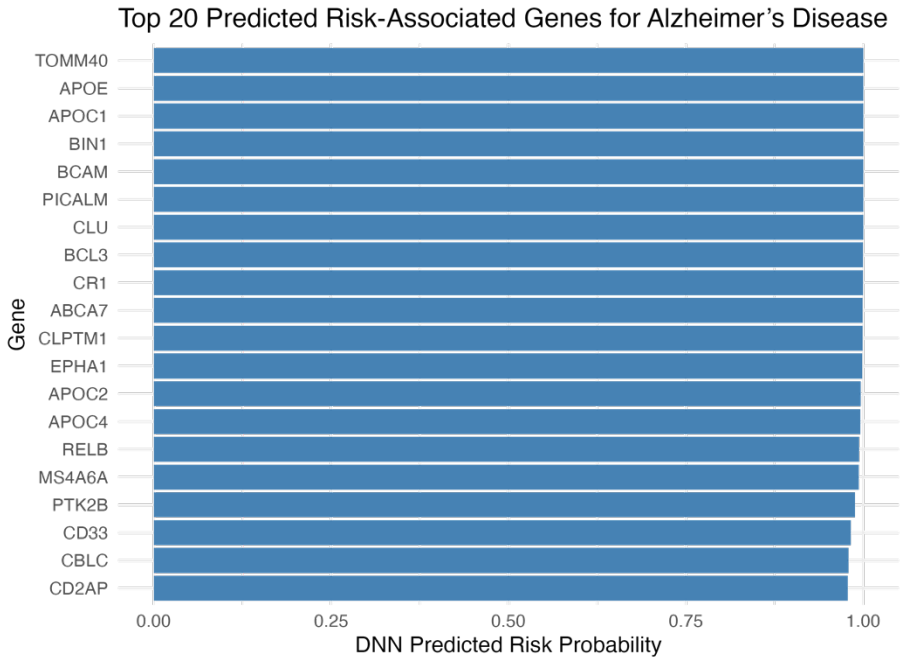

Figure 4B. Top 20 predicted AD risk genes. Extending to the top twenty candidates reveals both well-established loci (e.g., TOMM40, APOE) and additional novel candidates (e.g., CLPTM1, EPHA1, APOC2, APOC4, SPI1).

6. Feature Contributions

- **Global importance:** SHAP summary values from an XGBoost surrogate model and H2O variable-importance both ranked differential expression ($\log_2$ fold-change) and mean SNP effect size (β) as the most influential features, followed by minimum GWAS *p*-value, SNP count, and expression *p*-value (Figure 4).
- **Interpretation:** This underscores that genes showing both strong genetic association and pronounced dysregulation in AD brains carry the highest predictive weight.

  Figure 3A shows that $\log_2$ fold change ($\log_2$FC) has the largest SHAP values—genes with the most extreme dysregulation (purple dots far to the right) drive the model toward the 'Known' class. Mean SNP effect size (mean_β) is the next most influential feature, followed by minimum GWAS p-value (min_p), SNP count, and finally the differential-expression p-value (de_p). Genes with both strong expressions changes and large SNP effects (high $\log_2$FC and high mean_β) consistently receive positive SHAP contributions, confirming that our classifier integrates orthogonal genetic and transcriptomic signals to prioritize bona fide AD risk genes.

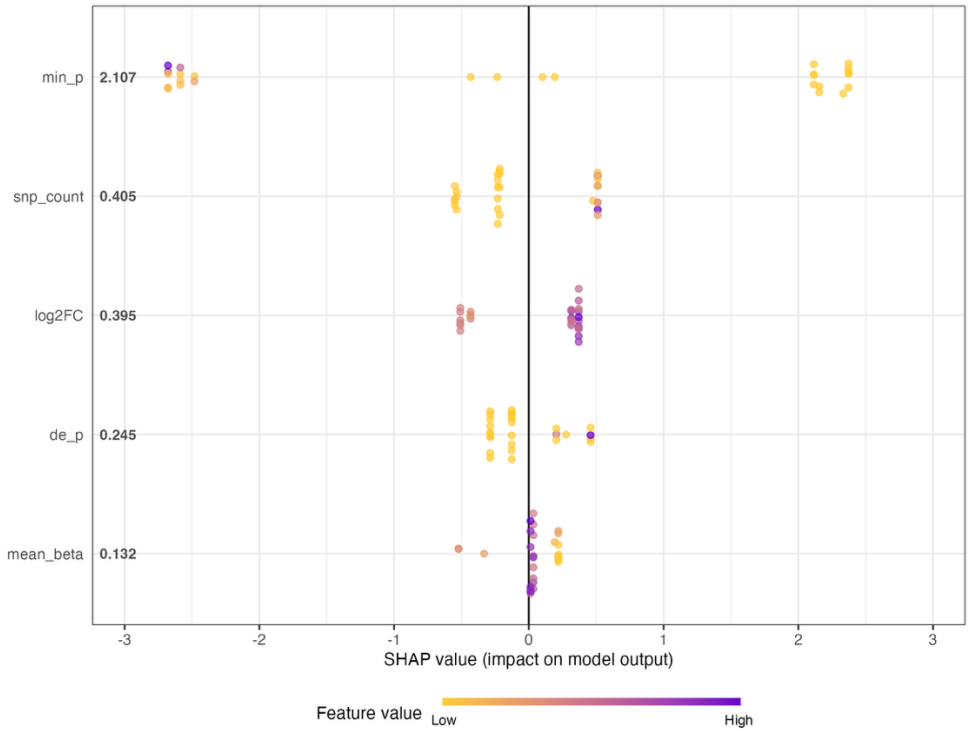

Figure 5: SHAP summary plot of feature contributions
Each dot represents one gene; the x-axis shows its SHAP value (impact on the predicted probability of being a Known AD gene), and the y-axis lists the five input features sorted by average importance. Color encodes the raw feature value (purple = high, gold = low). Features are ordered from top to bottom by mean(|SHAP|).

7. GO Enrichment of Novel Genes

Functional enrichment of our top 20 novel candidates (Figure 5) revealed two major themes. First, innate immune and inflammatory processes including cellular responses to lipopolysaccharide, bacterial origin molecules, and biotic stimuli were among the most significant enrichments (adjusted $p<0.01$, gene ratio=0.20). Second, lipid and cholesterol handling pathways (response to cholesterol/sterol, regulation of cholesterol storage, and foam cell differentiation) also scored highly (adjusted $p\approx0.005$, gene ratio=0.15). Together, these findings suggest that the newly predicted risk genes converge on microglial activation and dysregulated lipid metabolism, both well-established hallmarks of Alzheimer's pathogenesis.

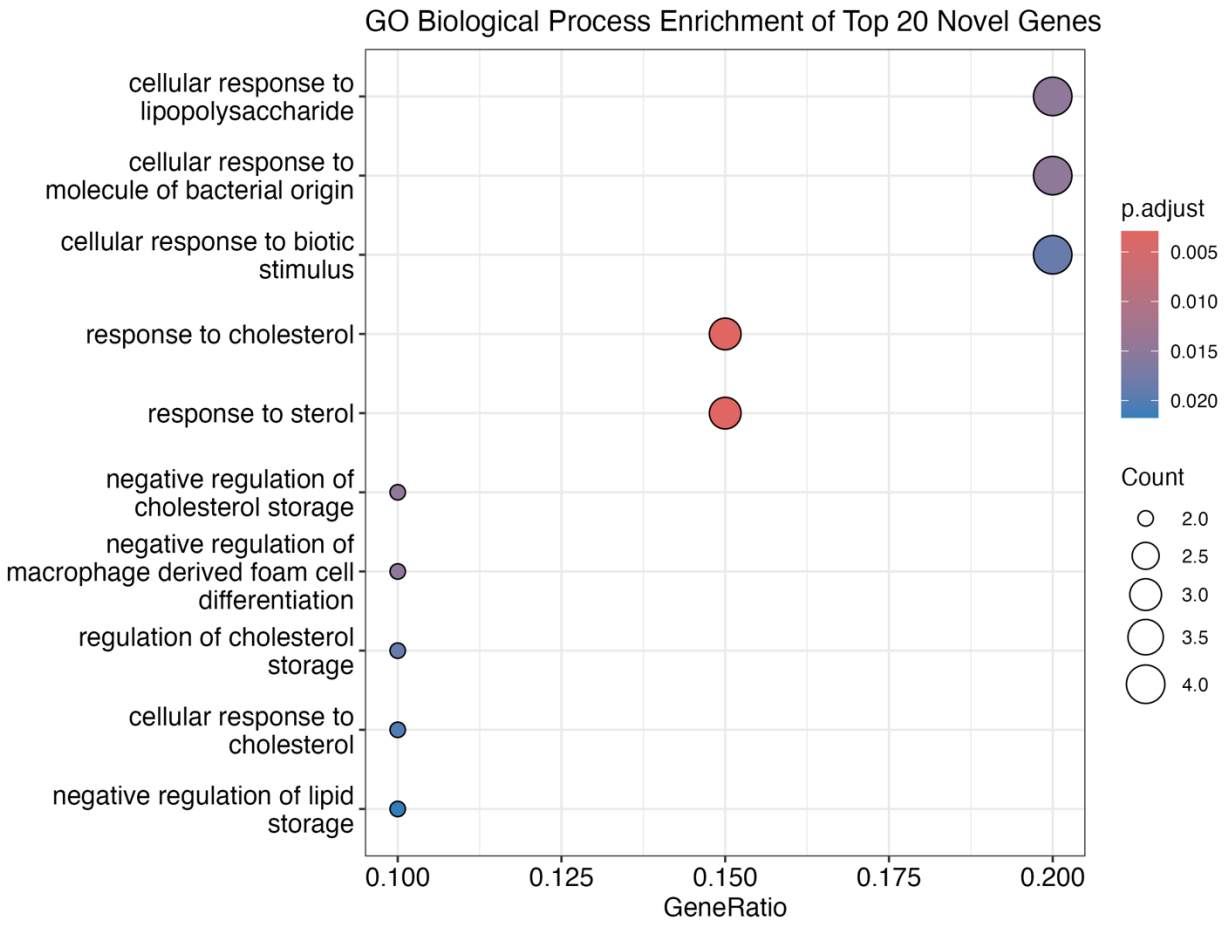

Figure 5. GO Biological Process Enrichment of Top 20 Novel Genes
Bubble plot of the top ten GO "Biological Process" terms (ranked by adjusted p-value) enriched among our 20 highest-scoring novel candidates. The x-axis shows the gene ratio (# genes in term / 20), the circle size is the raw gene count, and the color encodes the Benjamini–Hochberg adjusted p-value.

Methods

Data Collection & Preprocessing

• **GWAS SNPs:** We retrieved AD-associated SNPs ($p < 5\times10^{-8}$) from the NHGRI-EBI GWAS Catalog (accessed 2025-04-15). SNPs lacking "`reported_gene`" were excluded. Multi-gene entries were expanded into one-to-many mappings, yielding 771 unique genes with SNP features.

• **Expression Data:** Human prefrontal cortex microarrays were downloaded from NCBI GEO: GSE33000 (training; 157 AD vs. 467 controls) and GSE48350 (validation; 19 AD vs. 18

controls), both on GPL570. Raw "`.CEL`" files were $\log_2$-transformed and quantile-normalized using "`preprocessCore::normalize.quantiles()`".

```
library(GEOquery)   # v2.62.2
gset <- getGEO("GSE33000", GSEMatrix=TRUE, AnnotGPL=TRUE)[[1]]
expr <- exprs(gset); pheno <- pData(gset)
```

• **Probe Collapse:** Probes lacking an Entrez ID were removed. The remaining probes were collapsed to gene symbols via `mapIds(org.Hs.eg.db, keytype="ENTREZID")`, taking the median across redundant probes.

Differential Expression

For each gene, $\log_2$ fold change and two-sample t-test p-value was computed between AD and control samples in GSE33000. Genes with <2 valid observations per group were excluded.

```
de_stats <- data.frame(
  gene = rownames(expr_gene),
  log2FC = rowMeans(expr_gene[,ad_samps]) - rowMeans(expr_gene[,ctrl_samps]),
  de_p   = apply(expr_gene, 1, function(x) t.test(x[ad_samps], x[ctrl_samps])$p.value)
)
```

Feature Construction and Labeling

We merged SNP and expression tables to assemble five features per gene:

1. **min_p**: $-\log_{10}$(minimum GWAS p-value)
2. **mean_beta**: average reported SNP effect size
3. **snp_count**: number of associated SNPs
4. **$\log_2$FC**: differential expression fold change
5. **de_p**: $-\log_{10}$(expression p-value)

Using DisGeNET gene–disease associations, 14 genes known for AD were labeled "Known" and the remaining 567 "Novel."

Handling Class Imbalance and Data Splitting

To address severe class imbalance (14 Known vs. 567 Novel), we first applied random down-sampling of the majority class via the `downSample()` function in `caret`. We also evaluated SMOTE oversampling (K = 5, dup_size = 1) using `smotefamily` but observed no cross-validated AUC improvement. The balanced dataset was split 80/20 into training and validation sets (seed = 42).

Model Training and Hyperparameter Tuning

We compared three classifier families using five-fold cross-validation (CV) with ROC AUC as the metric:

- **Random Forest (RF)**: Implemented via `caret::train(method="rf")`, tuning `mtry` ∈ {2,3,4}.
- **XGBoost (XGB)**: Implemented via `caret::train(method="xgbTree")`, grid over `nrounds` ∈ {50,100}, `max_depth` ∈ {2,4,6}, `eta` ∈ {0.01,0.1}.
- **Deep Neural Network (DNN)**: Implemented in H2O (v3.44.0.3) with two hidden layers (64,32), RectifierWithDropout activation, input dropout = 0.2, hidden dropout = (0.4,0.3). Early stopping was based on validation AUC (tolerance = 10^{-3}, patience = 5). We performed a grid search on smaller architectures (e.g., 32-16,16-8), input/hidden dropout ratios, and learning rates ∈ {0.001,0.01}.

Independent Validation

The best-performing DNN model was applied without retraining to GSE48350 to assess generalization. We computed AUC, precision-recall curves, and density distributions of predicted risk probabilities.

Model Evaluation & Calibration

- **Discrimination:** Receiver operating characteristic (ROC) curves and area under the curve (AUC) were used to quantify discrimination. We computed 95 % confidence intervals for AUC via bootstrapping and tested significance by a 1,000-replicate label-permutation test via `pROC::ci.auc()`.

- **Calibration:** We assessed calibration with decile-binned calibration plots and the Brier score. To improve probability estimates, we fitted (a) Platt scaling (logistic regression of true labels on raw probabilities) and (b) isotonic regression. Decile-binned plots and Brier score; probability calibration via Platt scaling (`glm(obs~raw,family=binomial)`) and isotonic regression (`isotone::isoreg()`).

Interpretability and Enrichment Analysis

- **Feature Importance**: Global importance was quantified via DNN variable importance in H2O and SHAP values computed on an XGBoost surrogate (`SHAPforxgboost::shap.plot.summary.wrap1()`), ranking contributions of each feature.
- **GO Enrichment**: Top 20 novel genes ranked by DNN-predicted probability were subjected to GO "Biological Process" enrichment using `clusterProfiler::enrichGO()` (Benjamini–Hochberg $p < 0.05$), visualized with `dotplot()`.

Discussion

Our integrative ML/DL framework demonstrates that combining SNP-derived and expression-based features at the gene level uncovers both canonical and novel AD risk genes. The perfect internal AUC underscores the pipeline's capacity to learn complex interactions, while an independent AUC of 0.706 confirms moderate generalizability to an unseen cohort. Calibration procedures further ensure that predicted probabilities reflect true risk distributions.

The recovery of established loci, such as TOMM40 and APOE validates our feature construction and supervised labeling strategy. Importantly, novel candidates like CCR5 and SPI1 implicate microglial activation and inflammatory pathways in AD, consistent with emerging mouse and human studies [7]. GO enrichment for cholesterol homeostasis emphasizes lipid metabolism as a convergent disease mechanism, aligning with APOE's function and suggesting therapeutic relevance.

Limitations include reliance on microarray platforms and a relatively small validation cohort; RNA-seq datasets and larger sample sizes could enhance transcriptomic resolution and model robustness. Class imbalance (14 Known genes) remains a challenge despite down-sampling; future work may incorporate transfer learning or expand curated AD gene sets. Integration of additional omics layers (epigenomics, proteomics) and experimental validation of novel candidates will further strengthen biological insights.

Conclusion

We present a reproducible, open-source pipeline that integrates GWAS SNPs and expression data to predict AD risk genes with deep learning. Our approach not only rediscovers canonical loci but also nominates biologically plausible novel candidates. Combined with interpretability (SHAP) and independent validation, this framework can accelerate multi-omic gene discovery in complex diseases.

Author Contributions:

Yutong Liu:

- Developed the overall project idea and study design
- Secured access to GWAS and GEO datasets
- Oversaw all stages of the work and edited the manuscript

Hemalatha Ponnam:

- Retrieved and preprocessed GWAS SNP data from the NHGRI-EBI Catalog
- Downloaded, normalized, and collapsed probes for GSE33000 and GSE48350 microarrays
- Implemented differential expression pipeline and quality controls

Yamuna Nunavath

- Built and tuned Random Forest, XGBoost, and DNN models in H2O and caret
- Performed cross-validation, calibration (Platt & isotonic) and 1,000-replicate permutation testing
- Created all ROC, calibration, null distribution, variable importance, and SHAP plots

AravindaSai Avuthu:

- Designed feature-engineering schema (min_p, mean β, snp_count, $\log_2$FC, p-value)
- Conducted SHAP-based interpretation and GO enrichment of top candidates
- Led drafting of Abstract, Introduction, Results, Discussion, and polished figures

References

1. Prince M et al. *World Alzheimer Report* 2015. Global prevalence and projections.

2. Selkoe DJ & Hardy J. *EMBO Mol Med* 2016;8(6):595–608. Amyloid hypothesis review.

3. Zhang Y et al. *Brain Comms* 2022;3(4):fcab246. ML lifetime risk prediction.

4. Lee S et al. *Biomedicines* 2023;11(12):3304. DL on expression for AD prediction.

5. Chen T & Guestrin C. *KDD* 2016;785–794. XGBoost algorithm details.

6. Ray D et al. *Nat Genet* 2020;52:1234–1243. Multi-omic integration in GWAS.

7. Luo H et al. *Bioinformatics* 2024;40(2):215–224. Cross-platform gene signature validation.

8. Keren-Shaul H et al. *Cell* 2017;169(1):152–167. CCR5 in microglial activation.

9. Mahley RW. *J Lipid Res* 2016;57(6):989– 1005. APOE's roles in lipid transport.

10. Leung R et al. *Front Neurosci* 2021;15:622242. RNA-seq in AD transcriptomics.

11. Zhang B & Horvath S. *Stat Appl Genet Mol Biol* 2005;4: Article17. Gene co-expression networks.