Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning Approaches Enhance Polygenic Risk Score in Alzheimer's Disease

Maria Ilaria Curci
National Research Council
Alessandro Orro
alessandro.orro@cnr.it

National Research Council

Article

Keywords:

Posted Date: September 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10829763/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine Learning Approaches Enhance Polygenic Risk Score in Alzheimer's Disease

Maria Ilaria Curci [1] and Alessandro Orro [1,*], the Alzheimer's Disease Neuroimaging Initiative

[1] Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate

* Corresponding author

alessandro.orro@cnr.it

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide, with late-onset AD (LOAD) accounting for most cases. Although the APOE ε4 allele represents the strongest known genetic risk factor, it explains only a small fraction of disease variability, while heritability estimates suggest a substantially larger genetic contribution. This highlights the importance of polygenic architectures and complex interactions among genetic variants. Polygenic risk scores (PRS), derived from genome-wide association studies, are widely used to estimate individual genetic susceptibility; however, their reliance on linear assumptions limits their ability to capture non-linear effects.

In this study, we systematically compared state-of-the-art PRS approaches with multiple machine learning (ML) models to improve genetic risk prediction in AD.

Our results demonstrate that ML-based approaches outperform traditional PRS methods in predictive accuracy, highlighting their ability to model complex genetic

relationships. Furthermore, explainable AI techniques enabled the identification of key genetic variants driving model predictions, providing insights into potential novel candidate genes involved in AD pathogenesis. These findings support the integration of ML methods into genetic risk assessment frameworks for improved early prediction of Alzheimer's disease.

Introduction

Alzheimer's disease (AD), the most common cause of dementia, is a progressive neurodegenerative disease, with late-onset Alzheimer's disease (LOAD) representing the vast majority of cases [Andrews2023]. It is a complex and multifactorial disease, associated with different risk factors, including both environmental and genetic factors [Andrade-Guerrero2023], with a genetic architecture is supposed to be polygenic [Zhou2021], with numerous genetic variants contributing to the susceptibility of the disease [Gunter2024]. Among these, the Apolipoprotein E (APOE) ε4 allele represents the strongest genetic risk factor, with different dosage conferring different levels of risk and affecting brain function decades before clinical onset [DiBattista2016]. However, APOE ε4 accounts for only approximately 6% of the total phenotypic variance [Ridge2013], indicating that it's insufficient to determine individual genetic risk alone, and, indeed, genome-wide association studies (GWAS) identified additional SNPs with small individual effects that may be correlated with AD [Baker2020]. This limited explanatory power of APOE contrasts with heritability estimates from twin studies, which place the overall genetic contribution in the range 60%-80% [Andrews2023]. Even when considering all currently known common variants identified via genome-wide association studies (GWAS), they collectively explain only about 30% of the total genetic variance [Ridge2016]. This significant gap, often referred to as "missing

heritability," suggests that much of the genetic risk resides in complex polygenic architectures and epistatic interactions among numerous variants with small individual effects [Baker2020]. In order to take into account these small effects, Polygenic Risk Scores (PRS) have become the conventional approach [Choi2020] [Duncan2019] [Lewis2020] [Mistry2018]. By combining risk estimates from GWAS into a single quantitative score, PRS provides a measure of individual genetic susceptibility [Bellou2025]. PRS methods have been extensively applied in different genetic diseases including Alzheimer [Bellou2025] [Escott-Price2017].

Different PRS methods, each based on distinct prior distribution for SNP effect sizes, were proposed in literature. LDpred-inf assumes an infinitesimal genetic architecture, modelling posterior SNP effect sizes under a Gaussian distribution, without a point mass at zero [Pham2022] [Clark2022]. PRS-CS applies continuous shrinkage (CS) priori on the SNP effect size, inversely proportional with the strength of the GWAS signal [Ge2019]. MegaPRS [Zhao2024] and QuickPRS [Zhang2021] employ a flexible Bayesian PRS framework that can employ multiple different prior specifications. Finally, BayesR models SNP effect sizes as a mixture of a point mass at zero and three Gaussian distributions [Shan2022]. However, most traditional PRS methods rely on linear, additive models and often assume SNP independence. This limits their ability to capture the non-linear complexities and gene-gene interactions that likely characterize the genetic landscape of Alzheimer’s disease [Zhou2023].

Therefore, given the central role of genetic predisposition in AD, identifying genetic markers could be crucial for risk prediction and consequently for early diagnosis

[Krishnamurthy2025], that may help identify individuals at high-risk, allow initiating targeted prevention and delay disease onset [Rasmussen2019].

An alternative approach for risk prediction based on genomic data relies on machine learning (ML) techniques, a branch of artificial intelligence (AI), that develop algorithms and models capable of learning from data and make predictions, especially Explainable AI (XAI) [Lundberg2017]. A key advantage of XAI ML methods is their ability to analyse complex multimodal data, such as medical or genomics data, to identify predictive patterns and provide information about which features contributed most to the model's decision [Soladoye2025] [ArnalSegura2025].

In this study, we evaluate multiple approaches for estimating individual genetic risk, comparing state-of-the-art polygenic risk score (PRS) methods with a range of machine learning (ML) models to determine which framework achieves the highest predictive accuracy.

Finally, we applied model interpretability approaches to identify the genetic variants driving the predictions of the best-performing ML model, which could reveal new potential candidate target genes.

Materials and Methods

Cohorts

The discovery cohort consist of a set of variant effect sizes taken from a meta-analysis of GWAS summary statistics from the European Alzheimer & Dementia Biobank (EADB) [Bellenguez2022] and the UK Biobank (UKBB) [Bahcall2018], which include 39,106 clinically diagnosed Alzheimer's disease (AD) cases, 46,828 proxy-AD cases, and 401,577 population-based controls [Bellenguez2022].

The target cohort used for this study is the Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal study, aiming at identifying biological markers (SNPs) and neuroimaging measures associated with cognitive decline and functional changes in both cognitively normal elderly individuals and in mild cognitive impairment (MCI) and AD patients [Mueller2005].

Participants, aged 55 to 90 years, were followed for 2 to 3 years with repeated imaging scans and psychometric assessments (ADNI1), and the study was subsequently expanded with the inclusion of new participants (ADNI-GO, ADNI2, ADNI3, and ADNI4).

The ADNI dataset includes genomic data, neuroimaging measures, cognitive assessments, and cerebrospinal fluid (CSF) biomarkers.

In this study, genetic data were used for risk stratification, while CSF biomarkers (amyloid-β42, tau, and phosphorylated tau) and demographic variables were incorporated in downstream analyses for model interpretation and explainability.

Genetic data were acquired from different genotyping platforms: Human610-Quad BeadChip for ADNI1 (757 participants), Illumina HumanOmniExpress BeadChip for ADNI-GO/ADNI2 (432) and Illumina Infinium Global Screening Array v2 Chip for ADNI3 (327). Standard quality control was independently applied to each dataset. First, individuals genetically distant from the European ancestry cluster (PCA analysis of whole genome) in the 1000 Genomes Project [Auton2015] were excluded. Then, SNPs with MAF<0.01 or Hardy-Weinberg Equilibrium (HWE), with p-values ≤ 10E-6, or with missing data rate >5% were removed using PLINK [Chang2015] and the resulting data were normalized and imputed using TOPMed [Das2016] and the relative reference panel. After imputation, the three datasets were merged and QCed, removing variants

with INFO < 0.7, MAF < 0.05 and missingness < 0.05. Finally, individuals were checked for relatedness (Pi^hat >0.2), yielding a final dataset (referred to as ADNI123 hereafter) of 5,511,868 SNPs and 989 individuals, of which 477 cases, 512 controls. See **Table 1** for details.

Dataset	Subjects (cases/controls)	SNPs
ADNI1	757	620,901
ADNI-GO/ADNI2	432	730,525
ADNI3	327	759,993
ADNI123-imputed	1000	28,724,720
ADNI123-topmed	989 (477/512)	5.511.868

Table 1. Description of the ADNI raw datasets used in this study (ADNI1, ADNI-GO/ADNI2 and ADNI3) and the ones obtained during the harmonization process. Sample sizes and SNP counts are reported for the individual study phases alongside the post-imputation merged steps. ADNI123-imputed is the initial merged assembly before quality control, imputation and phenotypic filtering. The ADNI123-topmed is the final dataset.

Methods for PRS analysis

In order to compare Machine Learning with state-of-art PRS, we selected three PRS programs, PRS-CS (using two distinct LD reference panels: 1000G and UKB), LDpred-inf and QuickPRS, and used the same setup in terms of number of subjects.

Methods for Machine Learning

We applied a range of machine learning methods to predict disease status from genomic data, evaluating models with different underlying assumptions to identify the approach that most accurately distinguishes patients from controls. Specifically, we

tested logistic regression, Naive Bayes, Random Forest, XGBoost, LightGBM, and CatBoost. The simplest statistical methods, Naive Bayes (NB) and Logistic Regression (LR), assume independence among predictors. All other methods are ensemble decision tree models, able to go beyond linearity and model complex relationships between data [Uddin2024]. Random Forest (RF) [Breiman2001] is an ensemble model based on bagging, where training data are randomly resamples and records excluded from each bootstrap are used as a test set, generating independent predictive models, whose aggregated outputs enhance performance and robustness [Nguyen2020]. In contrast, XGBoost (XGB) [Chen2016], LightGBM (LGBM) [Dorogush2018], and CatBoost [Ke2017] are ensemble models based on boosting, where predictive models are not independent, but they are iteratively aggregated so that each one compensates for the errors from the previous one, thus yielding a high-performance and preventing overfitting problems [Khan2024].

To minimize the risk of algorithmic overfitting to the training cohort and to provide a fair assessment of the predictive performance of the selected variants, hyperparameter optimization was deliberately omitted. Unless otherwise specified, all models were trained using their default settings, with only a limited number of manually specified parameters to ensure reproducibility and stable model fitting. Specifically, fixed random seeds (42, 128, 2056) were used for all models, n_estimators was set to 200 for the Random Forest model, and max_iter was set to 500 for Logistic Regression.

Data Pre-processing and Feature Selection

Before model training, we use the same setup described in [Bracher-Smith2025], where SNPs were LD-pruned through clumping (R^2 = 0.75, window = 1 Mb) using GWAS

summary statistics from the European Alzheimer & Dementia Biobank (EADB), retaining relatively independent variants and reducing redundancy due to linkage disequilibrium [Privé2019]. This procedure resulted in a final set of 786,509 variants. The APOE-defining SNPs (rs429358 and rs7412) were subsequently manually reintroduced, as they were not included in the EADB GWAS summary statistics and were therefore excluded during LD-based dataset reduction.

Subjects were partitioned using stratified 10-fold cross-validation ensuring that each one was included exactly once in the test set across iterations. Within each iteration, data from the remaining folds were used for training, while the held-out fold was used for testing. The same experiment is replicated three time.

Within each training set, an additional SNP reduction step was performed by ranking variants according to their chi-square association statistics with case-control status using PLINK. Only variants selected from the training data were used for model construction, preventing any data leakage.

To determine the optimal number of variants for modelling, different feature set sizes (ranging from 50 to 10,000 variants) were empirically evaluated. The best predictive performance and stability across replicate folds were obtained with 5,000 SNPs, which were therefore used in the final models.

Genotypes were encoded as numbers (0=missing, 1=A1A1, 2=A1A2, 3=A2A2). To ensure a fair comparison between ML-based approaches and PRS methods, we used identical train/test splits and the same set of SNP variants across all analyses.

In the ML framework, the APOE locus was explicitly retained because of its well-established predictive relevance for AD. In contrast, when using state-of-the-art PRS approaches based on EADB GWAS, APOE variants were not present after variant harmonization and filtering, leading to a potentially unbalanced comparison.

To improve comparability between the two frameworks, PRS-CS was therefore recalculated using GWAS summary statistics generated directly from the training set (via –logistic regression), while maintaining the same SNP subset, including the APOE-defining SNPs, used in the ML models (5,000 SNPs). The resulting PRS models were then evaluated on the independent test sets.

This strategy ensured that both ML and PRS approaches were trained and evaluated under equivalent genetic information and data partitioning conditions.

Explainability Analysis

In order to understand what are the features that mostly contribute to the prediction, we employed two feature importance methods: Permutation feature importance (PFI) [Breiman2001] and SHapley Additive exPlanations (SHAP) [Lundberg2017] [Janzing2020].

PFI calculates how much the model's prediction errors increase when permuting the values of a feature, breaking the relationship between the feature and the true outcome. If a feature is important, permuting its values increases the model's error, because the model relies on that feature to generate the prediction.

SHAP analysis is an alternative feature-based interpretability method that computes feature importance by training a classification model on all the features and then computing the SHAP values for each one.

The Shapley values indicate the magnitude and the direction of each feature's contribution to the model’s output.

SHAP and PFI analyses were performed on the best-performing model, including cerebrospinal fluid (CSF) biomarkers as additional covariates in order to contextualize model predictions within well-established Alzheimer’s disease biological markers. SNP-level results were subsequently aggregated by retaining only variants consistently identified across cross-validation folds and independent seed replicates.

Specifically, we selected for each fold the best 100 ranked SNPs, and we analysed only the SNPs present in at least 8 folds and in at least 2 replicate seeds.

As AD cases without APOE ε4 are present in the dataset, we evaluated which variants emerged independently of the APOE ε4 effect, training a separate model restricted to individuals not carrying any APOE ε4 allele. Due to the reduced sample size (530 subjects), this analysis was performed using 5-fold cross-validation instead of 10-fold. In this setting, retained SNPs were required to appear in at least 3 out of 5 folds and across at least 2 independent seed replicates. Furthermore, among the retained SNPs we also investigated variable interactions using SHAP interaction values.

Statistical analysis

Descriptive methods

To describe the sample's characteristics, we summarized them using descriptive statistics. In results, continuous variables are presented as mean, while categorical variables are shown as frequency (%). Differences between cases and controls were

assessed using t-tests for continuous variables and chi-squared tests for categorical variables.

Performance evaluation

PRS and ML methods were compared with the Area Under the ROC curve (AUC-ROC) using both the whole genome (M-whole) and excluding variants physically present in the APOE region (CHR19: 43.9–46.0 Mb (GRCh38/hg38), M-noAPOE). Comparison of AUC-ROC values of PRS methods were computed using logistic regression, adjusting for age, sex, and eight principal components (PCs), and the resulting PRS were standardized.

Machine learning performance was also assessed using additional common metrics

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{accuracy}$$

$$P = \frac{TP}{TP+FP} \quad \text{precision}$$

$$R = \frac{TP}{TP+FN} \quad \text{recall}$$

$$F1 = \frac{2 \cdot P \cdot R}{P+R} \quad \text{F1 score}$$

where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives.

Comparison of PRS and ML

We performed pairwise comparisons of individual scores across M-whole PRS models and, separately, across M-whole ML models, computing Pearson's correlations for each set of comparisons.

We also evaluated the extent to which both PRS- and ML-based models can identify individuals at high and low risk of AD. We defined the extremes as the lowest and highest 2% of the prediction distribution and examined the proportion of real cases and controls within each group.

Finally, we compared AUC ROC between methods using DeLong's two-sided test [DeLong1988]. Unadjusted p-values were reported for pairwise comparisons of AUC-ROC values.

Results

Demographic and clinical characteristics of dataset

ADNI dataset includes a large number of demographic and clinical features, including age, gender, education level, cognitive scores and biomarkers. This study included 989 participants, comprising 535 males and 454 females, with a mean age of 73.78 years (range: 55–90.9).

Several significant differences emerged between cases and controls. Sex distribution differed, with a higher proportion of females among controls (51.4%) than cases (40.0%, $p < 0.001$). Cases were slightly older than controls (72.64 vs. 75.00 years, $p < 0.001$). Regarding APOE genotype, the proportion of ε2 carriers was lower in AD cases compared to controls (4.6% vs. 12.7%, $p = 1.23 \times 10^{-5}$), whereas ε4 carriers were more frequent among cases (65.0% vs. 29.1%, $p < 0.001$).

Cognitive performance differed markedly between groups: cases showed worse performance across multiple measures, with higher ADAS-Cog 11/13 and CDR-SB scores, and lower MMSE and MoCA scores, consistent with impaired cognition. In

addition, AD cases had lower educational attainment compared to cognitively normal participants.

Significant differences were also observed in CSF biomarkers. Compared with controls, AD cases showed markedly lower Aβ42 levels (668.63 vs. 1312.03, $p = 2.85 \times 10^{-51}$), together with higher total tau (355.23 vs. 244.29, $p = 1.67 \times 10^{-30}$) and phosphorylated tau levels (34.49 vs. 22.14, $p = 3.29 \times 10^{-32}$), reflecting the characteristic amyloid and tau pathology associated with Alzheimer's disease.

See **Table 2** for details.

Variable	Control Mean (number)	Case Mean (number)	p-value	num. of missing
ABETA42	1312.03 (374)	668.63 (317)	2.85e-51	298
TAU	244.29 (374)	355.23 (316)	1.67e-30	299
PTAU	22.14 (371)	34.49 (315)	3.29e-32	303
TOTSCORE (ADAS-Cog11)	6.87 (419)	27.05 (202)	1.56e-55	368
TOTAL13	11.00 (415)	38.47 (197)	4.07e-65	377
CDRSB	0.63 (418)	8.62 (213)	2.17e-70	358
MMSCORE	28.54 (416)	18.68 (213)	3.79e-58	360
MOCA	25.73 (311)	15.08 (52)	4.65e-17	626
PTEDUCAT	16.65 (512)	15.51 (477)	6.65e-11	0
AGE	72.64 (512)	75.00 (477)	1.61e-07	0
Male	48.6% (512)	60.0% (477)	0.000452	0
e2_carrier	12.7% (512)	4.6% (477)	1.23e-05	0

Table 2. Description of demographic and clinical data of the dataset

PRS state-of-art results

Given that PRS methods rely on different prior distributions, we benchmarked their performance in discriminating cases from controls to determine the most accurate approach. **Figure 1** summarizes the prediction accuracies of all tested PRS models: PRS-CS (1000G), PRS-CS (UKB), LDpred-inf and QuickPRS.

The best-performing model was PRS-CS (1000G), with an AUC-ROC of 0.593, closely followed by PRS-CS (UKB), which achieved a similar AUC-ROC of 0.587. QuickPRS ranked third with an AUC-ROC of 58.0%, while LDpred-inf showed the lowest performance, with an AUC-ROC of 56.5%.

When excluding the APOE region (M-noAPOE model), predictive performance showed a slight reduction and became highly similar across methods, with AUC-ROC values ranging from 56.5% to 56.8%. PRS-CS (UKB) achieved the highest performance (AUC-ROC = 0.568), followed by PRS-CS (1000G) (56.7%), QuickPRS (56.6%), and LDpred-inf (56.5%), which remained largely unchanged compared to M-whole.

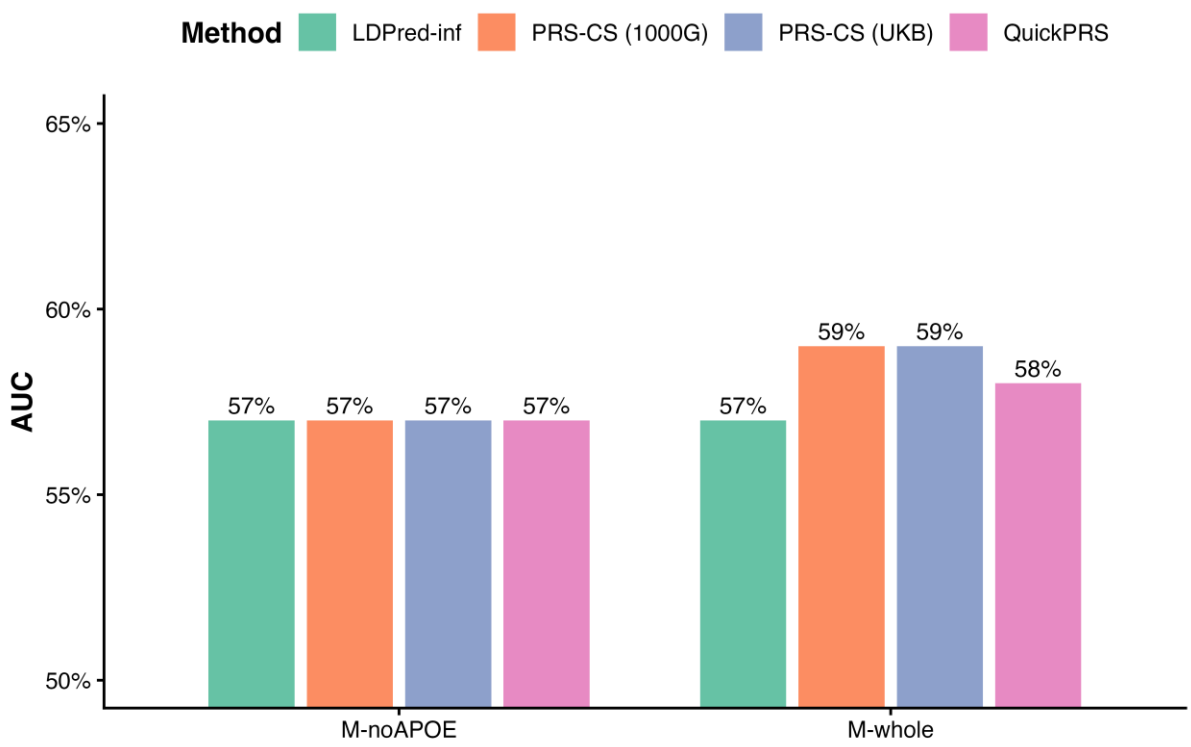


Figure 1. Comparison of PRS state-of-art methods. All the analysis were restricted to variants with a pT < 0.1, as suggested in literature [Bellou2025] [Leonenko2021].

Further, when a PRS method relied on pre-computed LD reference panels, the analysis was restricted to the corresponding set of SNPs.

Comparison of PRS individual's genetic scores

Supplementary Figure S1 represents the pairwise comparison of individual's score calculated as M-whole between each method.

The highest correlation is achieved between PRS-CS (1000G) and PRS-CS (UKB) (Pearson correlation 0.982). The lowest correlation is between QuickPRS and PRS-CS (UKB) (0.784).

Supplementary Table S1 summarizes the performance of all methods in the top and bottom 2% extremes, evaluating the ability of PRS to identify individuals at high and low risk. PRS-CS (both configurations) and LDpred-inf perform similarly, with LDpred-inf identifying one additional case in the highest-risk tail compared to PRS-CS. QuickPRS shows the worst performance, with a tendency to classify controls rather than cases.

Machine Learning results

We further benchmarked Machine Learning performance (**Figure 2**). Among all tested models, CatBoost achieved the highest predictive performance, with a mean AUC of 69%. It was closely followed by LightGBM (AUC = 68%) and XGBoost (AUC = 67%).

Logistic Regression and Naive Bayes showed substantially lower performance (AUC = 56% and 57%, respectively), while Random Forest performed modestly (AUC = 59%).

The PRS-CS (ML setup) model alone achieved lower discriminative ability (AUC = 54%) compared to all machine learning approaches.

When excluding the APOE region, predictive performance in tree-based models showed a significant decrease, becoming highly similar, if not worse, to LR, NB and PRS-CS, with AUC-ROC values ranging from 0.53 to 0.56 (**Figure 3**). For further details see **Supplementary Table S2.**

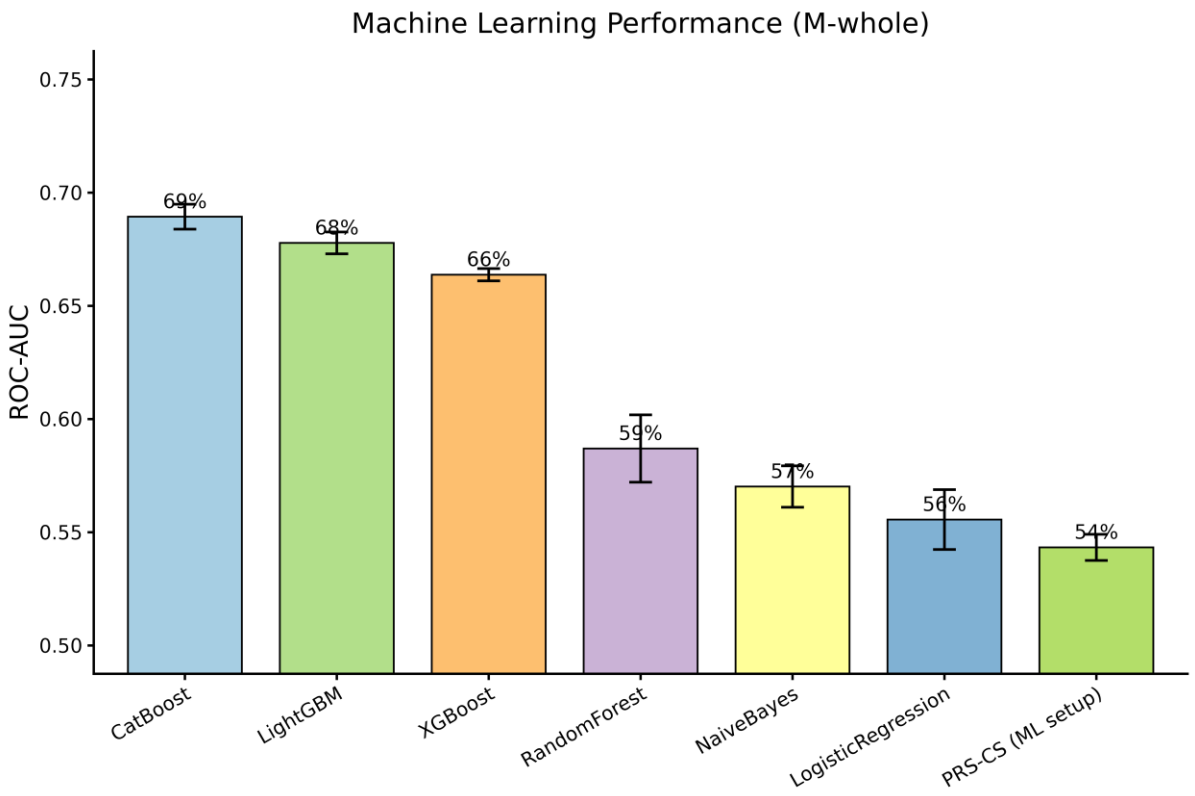


Figure 2. Predictive performance of six machine learning models (CatBoost, LightGBM, XGBoost, Random Forest, Naive Bayes, and Logistic Regression) and the PRS-CS method, with the APOE region included in the analysis. All methods were evaluated using the same training/test split and the same set of genetic variants.

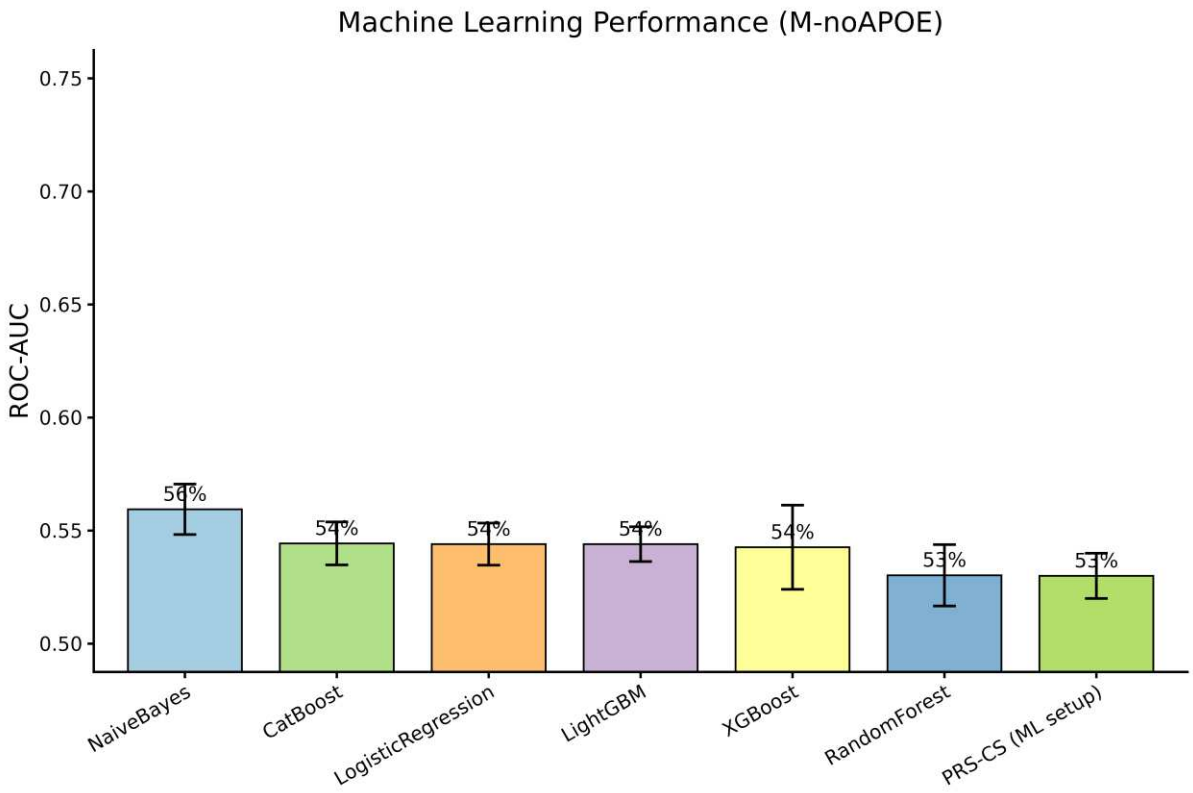


Figure 3. Predictive performance of six machine learning models (CatBoost, LightGBM, XGBoost, Random Forest, Naive Bayes, and Logistic Regression) and the PRS-CS method, excluding the APOE region in the analysis. All methods were evaluated using the same training/test split and the same set of genetic variants.

Comparison of ML individual's risk scores

Even for ML models, we evaluated pairwise comparisons of individual risk scores, as shown in Supplementary **Figure S2**.

Pairwise comparisons of individual risk scores revealed high correlations among gradient boosting models, with CatBoost, LightGBM and XGBoost showing correlations ranging from r = 0.90 to 0.93. Random Forest displayed moderate correlations with boosting models (r = 0.44–0.51), while Logistic Regression and Naive Bayes showed

higher concordance with each other (r = 0.74). The PRS showed negligible correlations with all machine learning approaches (r = −0.005 to 0.045).

We also evaluated model performance in the top and bottom 2% of the risk distribution. In this extreme-stratification setting, LGBM showed the highest precision in identifying individuals at high risk, outperforming the other ML methods (**Supplementary Table S1**).

Direct comparison of ML vs PRS

When directly comparing PRS state-of-the-art approaches with machine learning methods in the presence of the APOE region, tree-based ensemble methods showed the strongest discriminative performance. In particular, XGBoost, LightGBM and Catboost significantly outperformed Random Forest, PRS-CS (1000G), Logistic Regression and Naive Bayes according to the DeLong test. Conversely, Random Forest, Logistic Regression and Naive Bayes did not significantly differ from PRS-CS. Overall, these results indicate that ensemble boosting methods achieve superior predictive performance when APOE information is included (Supplementary Material, **Table S3**).

In contrast, after excluding the APOE region, the differences between methods were markedly reduced, with no substantial statistically significant superiority among models. In this setting, state-of-the-art PRS approaches achieved the highest AUC-ROC values overall, although performances remained broadly comparable across methods.

When considering risk stratification in the full model at the extreme 2% of the predicted risk distribution, we evaluated model performance based on the number of cases in the lowest-risk tail and the number of controls in the highest-risk tail, with lower values indicating better discrimination. In the lowest-risk extreme, RF, LR, and NB performed

worse than state-of-the-art PRS methods, whereas XGB achieved equivalent results to MegaPRS and superior to the other PRS approaches. LGBM showed the best overall performance.

Similarly, in the highest-risk extreme, RF and XGB performed worse than most PRS methods but remained comparable to MegaPRS. Catboost outperformed all state-of-the-art PRS methods except LDpred-inf, with which it achieved similar performance, while LGBM again showed the best discrimination overall. LR performed similarly to most PRS methods but worse than MegaPRS, whereas NB showed the poorest performance (Supplementary **Table S1**).

XAI results

CatBoost achieved the highest predictive performance and was therefore selected for the interpretability analyses. To provide biological context for the identified genetic variants, CatBoost was retrained by including cerebrospinal fluid (CSF) biomarkers as additional covariates. Missing values in the CSF biomarkers were retained without imputation, as CatBoost can supports missing data. As expected, the inclusion of CSF biomarkers increased the predictive performance of the model (AUC-ROC = 95% ± 0.015). Because this extended model was developed exclusively to support feature interpretability, it was not included in the comparative performance analyses of the SNP-only models. Variants were retained if they were consistently prioritized across CatBoost replicates according to SHAP values, permutation importance, or both, resulting in a final set of seven variants.

These variants were functionally annotated with genes mapped from GTEx database [gtex2013] when available, otherwise using genes mapped from the Genomics

Database of The National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) [Kuzma2025].

As shown by the global SHAP analysis (Supplementary Figure S3) rs429358 had the largest average contribution to model predictions. Of the remaining variants, two currently lack functional or disease-related annotation, whereas the others showed evidence of association with Alzheimer's disease-related traits, like $A\beta_{42}$ and Neuritic plaques (**Table 3** and Supplementary **Figure S3**).

Variant (consequence)	**Gene**	**Selection method**	**p-value (Pubmed ID)**	**AD-related traits**
rs429358 (missense)	**APOE**	SHAP + PI	2e-303 (31473137)	$A\beta_{42}$* Neuritic plaques**
rs8077247 (intron)	ASIC2	SHAP + PI	0.3152 (35379992)	Neuritic plaques ** Lewy body disease**
rs7094120	NA	SHAP only	0.3149 (35379992)	NA
rs1903535	NA	SHAP only	0.3145 (35379992)	NA
rs11946632 (upstream gene).	**PABPC4L**	SHAP only	0.0000355 (34112972)	Cerebral amyloid angiopathy ** Neurofibrillary tangles **

rs6116375 (upstream gene)	AL121781.1 (lncRNA)	SHAP only	0.1516 (35379992)	Neuritic plaques **
rs55844294 (upstream gene)	LINC00661 (lncRNA)	PI only	0.6139 (35379992)	Neuritic plaques **

Table 3: Seven variants consistently identified across CatBoost replicates. The selection method indicates whether each variant was identified as important by SHAP, permutation importance (PI), or both methods (SHAP + PI). Variants shown in bold reached nominal significance ($P < 0.05$). AD-related traits refer to gene-level associations. * indicates accession NG00055 and ** indicates accession NG00041.

Since Catboost may capture non-linear SNP–SNP interactions, we additionally computed SHAP interaction values among these prioritized genes. Interestingly, the two previously uncharacterized variants displayed the APOE variant as their second strongest interaction partner after $A\beta_{42}$, whereas APOE variant ranked lower for the remaining variants. This pattern could suggest that these variants may encode APOE-related information through alternative genetic signals (Supplementary **Figure S4**).

We also investigated the genetic signals emerging beyond the APOE ε4 effect. After including cerebrospinal fluid biomarkers as additional covariates, the extended CatBoost model achieved a mean AUC-ROC of 93% ± 0.022. As in the primary analysis, this extended model was used exclusively for interpretability and was not included in the comparison of predictive performance across SNP-only models. In this setup variants prioritized by Catboost were defined as those consistently selected across at

least 3 out of 5 cross-validation folds and in at least 2 out of 3 random seeds, yielding a final set of seven variants (**Table 4** and Supplementary **Figure S5**).

Variant (consequence)	**Gene**	**Selection method**	**p-value (PubMed ID)**	**AD-related traits**
rs9389691 (intron)	**AL592429.2 (lncRNA)**	SHAP + PI	0.000256 (23571587)	Hippocampal sclerosis of the elderly ** Neurofibrillary tangles ** Cerebral amyloid angiopathy ** Neurofibrillary tangles ** $A\beta_{42}$ *
rs12320891 (regulatory region)	**ETFBKMT**	SHAP only	0.01498 (35379992)	$A\beta_{42}$ * Tau *
rs72849471 (intron)	LINC01482 (lncRNA)	SHAP only	0.369 (35379992)	ptau * tau * Lewy body disease **
rs871049 (intron)	ITIH5	SHAP only	0.6026 (35379992)	Neuritic plaques ** Neurofibrillary tangles **

rs7858803	AL591644.1 (processed pseudogene)	SHAP only	0.4495 (35379992)	$A\beta_{42}$ *
rs2623589 (intron)	CSMD1	PI only	0.07186 (35379992)	Neuritic plaques ** Neurofibrillary tangles **
rs2938513 (intron)	AC092100.1 (lncRNA)	PI only	0.7114 (35379992)	Hippocampal sclerosis of the elderly ** Lewy body disease **

Table 4: Seven variants consistently identified across CatBoost replicates. The selection method indicates whether each variant was identified as important by SHAP, permutation importance (PI), or both methods (SHAP + PI). Variants shown in bold reached nominal significance ($P < 0.05$). AD-related traits refer to gene-level associations. * indicates accession NG00055 and ** indicates accession NG00041.

Discussions & conclusions

Identifying individuals at genetic risk for Alzheimer's disease is essential for enabling early risk stratification and preventive interventions [Abdelnour2022]. Further, predictive models can uncover patterns among genetic and biological markers associated with disease susceptibility, providing insight into the mechanisms underlying disease development [Winchester2023].

The aim of this study was first to evaluate and compare different methodologies for polygenic risk score (PRS) construction, and subsequently to assess whether machine

learning approaches could improve polygenic trait prediction beyond standard linear PRS models. In particular, we investigated models capable of modelling nonlinear relationships and interaction effects between SNPs and then compared their predictive performance with that of conventional linear approaches.

Overall, state-of-the-art PRS approaches showed remarkably similar performance to each other, suggesting that increasingly complex PRS optimization strategies may yield only limited gains in predictive accuracy in AD.

In contrast, machine learning approaches exhibited greater variability in performance, with boosting-based models generally outperforming simpler linear and probabilistic models. Pairwise correlations between individual risk scores revealed a high degree of concordance among gradient boosting approaches, consistently with the shared methodological assumptions of these algorithms. Conversely, RF showed lower correlation with boosting methods, likely reflecting differences between bagging-based and boosting-based learning strategies, despite both approaches being tree-based ensemble methods. Similarly, LR and NB displayed higher concordance compared with their correlation with tree-based models, consistent with their simpler modelling assumptions. The low correlation between the PRS and all machine learning models suggests that ML-derived risk scores capture complementary aspects of genetic risk beyond those represented by a conventional additive polygenic model.

Our findings also highlight the dominant role of APOE ε4 in genetic risk prediction in ADNI dataset. The marked reduction in predictive performance observed after excluding APOE variants suggests that APOE accounts for a substantial proportion of both the overall genetic signal and the complex patterns leveraged by machine learning

methods in this dataset. In the absence of APOE, the remaining genetic signal, likely distributed across many weakly associated SNPs, appears to be largely additive, which may explain the reduced performance of machine learning approaches and suggests that PRS, which are optimized for additive genetic effects, may be a more suitable modeling strategy in this context.

A different pattern of performance is observed when evaluating the ability to discriminate individuals at extreme genetic risk. LightGBM demonstrated the strongest and most consistent performance across both extreme-risk groups (High: 18 cases / 2 controls; Low: 4 cases / 16 controls), outperforming all other methods, including both alternative machine learning approaches and PRS-based models. In contrast, XGBoost and CatBoost showed minor discriminatory performance. Notably, unlike the AUC-ROC results obtained across the entire cohort, complex machine learning models did not consistently outperform simpler linear approaches in stratifying individuals at the extremes of genetic risk. In several cases, PRS-based and linear models achieved comparable or superior classification performance, indicating that greater model complexity does not necessarily improve extreme-risk stratification in this context.

The second aim of this study was to explore the variants prioritized by the predictive models to identify potential risk loci in both APOE ε4 carriers and non-carriers, that may offer valuable insights into subgroup-specific genetic contributions and help inform future validation studies.

Among the top-ranked variants identified by the machine learning model, variants mapped to PABPC4L in APOE ε4 carriers and AL592429.2 (lncRNA) and ETFBKMT in APOE ε4 non-carriers showed nominal evidence of association and may represent

biologically relevant candidates for Alzheimer's disease. PABPC4L (Poly(A) Binding Protein Cytoplasmic 4 Like) has been suggested to be involved in the formation of stress granules (SGs) [Aslam2019]. Stress granules are transient RNA- and protein-containing cytoplasmic structures that help protect neurons under stress by suppressing global mRNA translation and temporarily storing RNAs and proteins. However, under chronic stress conditions, SGs can become persistent, and long-lasting SGs have been associated with neurodegenerative diseases, where they may promote the aggregation of pathological proteins such as phosphorylated tau. Pathological SGs may also irreversibly sequester RNAs, thereby disrupting neuronal RNA metabolism and contributing to neuronal dysfunction, cell loss, and cognitive impairment [Sato2023].

Given its involvement in cellular stress responses, PABPC4L may therefore be implicated in neurodegenerative processes, and further investigation of this gene could help uncover novel mechanisms underlying neurodegeneration, as previously suggested in the literature [Aslam2019]. In line with this hypothesis, recent transcriptomic evidence has also reported increased expression of PABPC4L in Alzheimer’s disease cases compared with controls [Vastrad2024].

AL592429.2 was identified as a gene mapped to a GWAS locus associated with PhenoAge acceleration, a measure of biological aging derived from clinical biomarkers [Kuo2021]. While the functional role of this lncRNA remains uncharacterized, the locus could be implicated in aging processes, which are known major risk factors for Alzheimer’s disease. Therefore, AL592429.2 may reflect aging-related regulatory mechanisms potentially relevant to neurodegeneration.

ETFBKMT encodes a mitochondrial lysine methyltransferase that regulates the electron transfer flavoprotein complex and acts as a negative regulator of fatty acid β-oxidation [Henriques2021]. Given the growing evidence linking mitochondrial dysfunction, altered lipid metabolism, and chronic neuroinflammation to Alzheimer's disease [Song2021], dysregulation of ETFBKMT may influence disease-relevant metabolic pathways, particularly in microglia and other innate immune cells. The prominence of ETFBKMT in the predictive model may therefore reflect its involvement in mitochondrial fatty acid metabolism, a process increasingly implicated in microglial dysfunction and neuroinflammatory responses associated with Alzheimer's disease.

Interestingly, the model also assigned high importance to variants located in genes previously implicated in AD pathogenesis. While these variants did not achieve nominal statistical significance in the present analysis, their prioritization by the model further supports the biological relevance of the identified predictive features.

ASIC2 (acid sensing ion channel subunit 2) is a subunit of the acid-sensing ion channels (ASICs), a group of voltage-independent proton-gated cation channels belonging to the degenerin/epithelial sodium channel (DEG/ENaC) superfamily [Sun2020].

ASIC1a, ASIC2a, and ASIC2b are the major subunits, which form acid-sensing ion channels that are activated by transient decreases in pH, particularly within the synaptic cleft during neurotransmission [Vessey2005]. They modulate neurotransmitter release, postsynaptic responses, and synaptic plasticity [Zha2013]. It has been proposed that brain acidification observed in AD may be linked to ASIC function. ASIC1a can increase hippocampal neuron excitability in the presence of Aβ and metabotropic glutamate receptor activation, suggesting a possible role in AD-related

synaptic dysfunction. The AD drug memantine has been reported to inhibit ASIC1a in addition to NMDA receptors, which may contribute to its effects. However, the specific contribution of ASIC2 is still not well defined, and the overall role of ASIC channels in neurodegenerative processes remains to be fully understood.

CSMD1 (Cub and Sushi Multiple Domains 1) acts as a modulator of the complement cascade, a key element of innate immunity. It is strongly expressed in the central nervous system, where the complement pathway has emerging roles in regulating neural development and synaptic function [Werren2024]. In mouse models of Alzheimer's disease, complement-mediated synaptic clearance by microglia has been shown to be impaired [Hong2016].

Overall, this study provides evidence for the contribution of nonlinear genetic effects to complex phenotypes, with the model prioritizing genes that span distinct but converging biological layers of Alzheimer's disease risk, including RNA stress responses, systemic aging processes, and microglial immunometabolic reprogramming.

This study has a number of limitations. First, it is limited by moderate sample size. Future work will focus on model validation in independent cohorts to assess robustness and generalizability. Second, the selection and identification of top SNPs depend on the method, as different approaches rely on different assumptions. No standard consensus exists on how to define or prioritize the most relevant variants for model prediction.

Acknowledgements

T3-AN-03 ”REte di Genomica Integrata per Nuove Applicazioni in medicina di precisione – REGINA” CUP B53C22002520006 founded by Italian Ministry of Health.

Data availability

[Bellenguez2022] summary statistics has been deposited to European Bioinformatics Institute GWAS Catalog (https://www.ebi.ac.uk/gwas/) under accession no. GCST90027158. Initiative (ADNI) used in this study were available at the database (http://adni.loni.usc.edu/), upon registration and compliance with the data usage agreement.

References

[Andrews2023] Andrews SJ, Renton AE, Fulton-Howard B, Podlesny-Drabiniok A, Marcora E, Goate AM. The complex genetic architecture of Alzheimer’s disease: novel insights and future directions. EBioMedicine. 2023;90: 104511.

[Andrade-Guerrero2023] Andrade-Guerrero J, Santiago-Balmaseda A, Jeronimo-Aguilar P, Vargas-Rodríguez I, Cadena-Suárez AR, Sánchez-Garibay C, et al. Alzheimer’s Disease: An Updated Overview of Its Genetics. International Journal of Molecular Sciences. 2023;24: 3754.

[Zhou2021] Zhou X, Li YYT, Fu AKY, Ip NY. Polygenic Score Models for Alzheimer’s Disease: From Research to Clinical Applications. Front Neurosci. 2021;15: 650220.

[Gunter2024] Gunter NB, Gebre RK, Graff-Radford J, Heckman MG, Jack CR Jr, Lowe VJ, et al. Machine Learning Models of Polygenic Risk for Enhanced Prediction of Alzheimer Disease Endophenotypes. Neurol Genet. 2024;10: e200120.

[DiBattista2016] M. Di Battista, A., M. Heinsinger, N. and William Rebeck, G., 2016. Alzheimer's disease genetic risk factor APOE-ε4 also affects normal brain function. Current Alzheimer Research, 13(11), pp.1200-1207.

[Ridge2013] Ridge, P.G., Mukherjee, S., Crane, P.K., Kauwe, J.S. and Alzheimer's Disease Genetics Consortium, 2013. Alzheimer's disease: analyzing the missing heritability. PloS one, 8(11), p.e79771.

[Baker2020] Baker E, Escott-Price V. Polygenic Risk Scores in Alzheimer's Disease: Current Applications and Future Directions. Front Digit Health. 2020;2: 14.

[Choi2020] Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15: 2759–2772.

[Duncan2019] Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R. and Domingue, B., 2019. Analysis of polygenic risk score usage and performance in diverse human populations. Nature communications, 10(1), p.3328.

[Lewis2020] Lewis, C.M. and Vassos, E., 2020. Polygenic risk scores: from research tools to clinical instruments. Genome medicine, 12(1), p.44.

[Mistry2018] Mistry, S., Harrison, J.R., Smith, D.J., Escott-Price, V. and Zammit, S., 2018. The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: Systematic review. Schizophrenia research, 197, pp.2-8.

[Bellou2025] Bellou E, Kim W, Leonenko G, Tao F, Simmonds E, Wu Y, et al. Benchmarking Alzheimer's disease prediction: personalised risk assessment using polygenic risk scores across various methodologies and genome-wide studies. Alzheimer's Research & Therapy. 2025;17: 1–11.

[Escott-Price2017] Escott-Price, V., Myers, A.J., Huentelman, M. and Hardy, J., 2017. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. Annals of neurology, 82(2), pp.311-314.

[Pham2022] Pham, D., Truong, B., Tran, K., Ni, G., Nguyen, D., Tran, T.T., Tran, M.H., Nguyen Thuy, D., Vo, N.S. and Nguyen, Q., 2022. Assessing polygenic risk score models for applications in populations with under-represented genomics data: an example of Vietnam. Briefings in Bioinformatics, 23(6), p.bbac459.

[Clark2022] Clark, K., Leung, Y.Y., Lee, W.P., Voight, B. and Wang, L.S., 2022. Polygenic risk scores in Alzheimer's disease genetics: methodology, applications, inclusion, and diversity. Journal of Alzheimer's Disease, 89(1), pp.1-12.

[Ge2019] Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A. and Smoller, J.W., 2019. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature communications, 10(1), p.1776.

[Zhao2024] Zhao, Z., Gruenloh, T., Yan, M., Wu, Y., Sun, Z., Miao, J., Wu, Y., Song, J. and Lu, Q., 2024. Optimizing and benchmarking polygenic risk scores with GWAS summary statistics. Genome Biology, 25(1), p.260.

[Zhang2021] Zhang, Q., Privé, F., Vilhjálmsson, B. and Speed, D., 2021. Improved genetic prediction of complex traits from individual-level data or summary statistics. Nature communications, 12(1), p.4192.

[Shan2022] Shan, Y. and Weeks, D.E., 2022. BayesRB: a markov chain Monte Carlo-based polygenic genetic risk score algorithm for dichotomous traits. bioRxiv, pp.2022-02.

[Zhou2023] Zhou X, Chen Y, Ip FCF, Jiang Y, Cao H, Lv G, et al. Deep learning-based polygenic risk analysis for Alzheimer’s disease prediction. Communications Medicine. 2023;3: 1–20.

[Krishnamurthy2025] Krishnamurthy HK, Jayaraman V, Krishna K, Wang T, Bei K, Changalath C, et al. An overview of the genes and biomarkers in Alzheimer’s disease. Ageing Res Rev. 2025;104: 102599.

[Rasmussen2019] Rasmussen, J. and Langerman, H., 2019. Alzheimer’s disease–why we need early diagnosis. Degenerative neurological and neuromuscular disease, pp.123-130.

[Lundberg2017] Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[Soladoye2025] Soladoye, A.A., Aderinto, N., Omodunbi, B.A., Esan, A.O., Adeyanju, I.A. and Olawade, D.B., 2025. Enhancing Alzheimer's disease prediction using random forest: A novel framework combining backward feature elimination and ant colony optimization. Current Research in Translational Medicine, 73(4), p.103526.

[ArnalSegura2025] Arnal Segura M, Bini G, Krithara A, Paliouras G, Tartaglia GG. Machine Learning Methods for Classifying Multiple Sclerosis and Alzheimer’s Disease Using Genomic Data. Int J Mol Sci. 2025;26. doi:10.3390/ijms26052085.

[Bellenguez2022] Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer’s disease and related dementias. Nat Genet. 2022;54: 412–436.

[Bahcall2018] Bahcall, O.G., 2018. UK Biobank--a new era in genomic medicine. Nature reviews genetics, 19(12), pp.737-738.

[Mueller2005] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin N Am. 2005;15: 869–77, xi–xii.

[Auton2015] 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature, 526(7571), p.68.

[Chang2015] Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 4(1), pp.s13742-015.

[Das2016] Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze S, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. Next-generation genotype imputation service and methods. Nature Genetics 48, 1284–1287 (2016).

[Uddin2024] Uddin S, Lu H (2024) Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. PLOS ONE 19(4): e0301541.

[Breiman2001] Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

[Nguyen2020] Nguyen, Xuan & Nguyen, Thi Thanh & La, Duong Duc & Kumar, Gopalakrishnan & Rene, Eldon & Nguyen, D. Duc & Chang, Soon-Woong & Chung, W. & Xuan Hoan, Nguyen & Nguyen, V.. (2021). Development of machine learning - based

models to forecast solid waste generation in residential areas: A case study from Vietnam. Resources, Conservation and Recycling. 167. 105381.

[Chen2016] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. 2016.

[Dorogush2018] Dorogush, A.V., Ershov, V. and Gulin, A., 2018. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

[Ke2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[Khan2024] Khan AA, Chaudhari O, Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Syst Appl. 2024;244: 122778.

[Bracher-Smith2025] Bracher-Smith M, Melograna F, Ulm B, Bellenguez C, Grenier-Boley B, Duroux D, et al. Machine learning in Alzheimer's disease genetics. Nature Communications. 2025;16: 1–16.

[Privé2019] Privé F, Vilhjálmsson BJ, Aschard H, Blum MGB. Making the Most of Clumping and Thresholding for Polygenic Scores. Am J Hum Genet. 2019;105: 1213–1221.

[Lundberg2017] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[Janzing2020] Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causal problem." In International Conference on artificial intelligence and statistics, pp. 2907-2916. PMLR, 2020.

[DeLong1988] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44: 837–845.

[Leonenko2021] Leonenko, G., Baker, E., Stevenson-Hoare, J. et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. Nat Commun 12, 4506 (2021).

[gtex2013] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45: 580–585.

[Kuzma2025] Kuzma A, Valladares O, Greenfest-Allen E, Nicaretta H, Kirsch M, Ren Y, et al. NIAGADS: A data repository for Alzheimer's disease and related dementia genomics. Alzheimers Dement. 2025;21: e70255.

[Abdelnour2022] Abdelnour, C., Agosta, F., Bozzali, M., Fougère, B., Iwata, A., Nilforooshan, R., Takada, L.T., Viñuela, F. and Traber, M., 2022. Perspectives and challenges in patient stratification in Alzheimer's disease. Alzheimer's research & therapy, 14(1), p.112.

[Winchester2023] Winchester, L.M., Harshfield, E.L., Shi, L., Badhwar, A., Khleifat, A.A., Clarke, N., Dehsarvi, A., Lengyel, I., Lourida, I., Madan, C.R. and Marzi, S.J., 2023. Artificial intelligence for biomarker discovery in Alzheimer's disease and dementia. Alzheimer's & Dementia, 19(12), pp.5860-5871.

[Aslam2019] Aslam M, Ullah A, Paramasivam N, Kandasamy N, Naureen S, Badshah M, et al. Segregation and potential functional impact of a rare stop-gain PABPC4L variant in familial atypical Parkinsonism. Sci Rep. 2019;9: 13576.

[Sato2023] Sato K, Takayama K-I, Inoue S. Stress granules sequester Alzheimer's disease-associated gene transcripts and regulate disease-related neuronal proteostasis. Aging (Albany NY). 2023;15: 3984–4011.

[Vastrad2024] Vastrad B, Vastrad C. Identification of candidate biomarkers and signaling pathways associated with Alzheimer's disease using bioinformatics analysis of next generation sequencing data. bioRxiv. 2024. doi:10.1101/2024.12.03.626535

[Kuo2021] Kuo C-L, Pilling LC, Liu Z, Atkins JL, Levine ME. Genetic associations for two biological age measures point to distinct aging phenotypes. Aging Cell. 2021;20: e13376.

[Henriques2021] Henriques BJ, Katrine Jentoft Olsen R, Gomes CM, Bross P. Electron transfer flavoprotein and its role in mitochondrial energy metabolism in health and disease. Gene. 2021;776: 145407.

[Song2021] Song T, Song X, Zhu C, Patrick R, Skurla M, Santangelo I, et al. Mitochondrial dysfunction, oxidative stress, neuroinflammation, and metabolic alterations in the progression of Alzheimer's disease: A meta-analysis of in vivo magnetic resonance spectroscopy studies. Ageing Res Rev. 2021;72: 101503.

[Sun2020] Sun D, Liu S, Li S, Zhang M, Yang F, Wen M, et al. Structural insights into human acid-sensing ion channel 1a inhibition by snake toxin mambalgin1. Elife. 2020;9. doi:10.7554/eLife.57096.

[Vessey2005] Vessey JP, Stratis AK, Daniels BA, Da Silva N, Jonz MG, Lalonde MR, et al. Proton-mediated feedback inhibition of presynaptic calcium channels at the cone photoreceptor synapse. J Neurosci. 2005;25: 4108–4117.

[Zha2013] Zha X-M. Acid-sensing ion channels: trafficking and synaptic function. Mol Brain. 2013;6: 1.

[Werren2024] Werren EA, Peirent ER, Jantti H, Guxholli A, Srivastava KR, Orenstein N, et al. Biallelic variants in CSMD1 are implicated in a neurodevelopmental disorder with intellectual disability and variable cortical malformations. Cell Death Dis. 2024;15: 379.

[Hong2016] Hong S, Beja-Glasser VF, Nfonoyim BM, Frouin A, Li S, Ramakrishnan S, et al. Complement and microglia mediate early synapse loss in Alzheimer mouse models. Science. 2016;352: 712–716.

Author contributions statements

Conceptualization: M.I.C., A.O.

Data curation: M.I.C.

Formal analysis: M.I.C.

Funding acquisition: A.O.

Investigation: A.O., M.I.C.

Methodology: M.I.C., A.O.

Project administration: A.O.

Software: M.I.C.

Supervision: A.O.

Validation: M.I.C., A.O.

Visualization: M.I.C.

Writing – original draft: M.I.C.

Writing – review & editing: M.I.C., A.O.

Supplementary

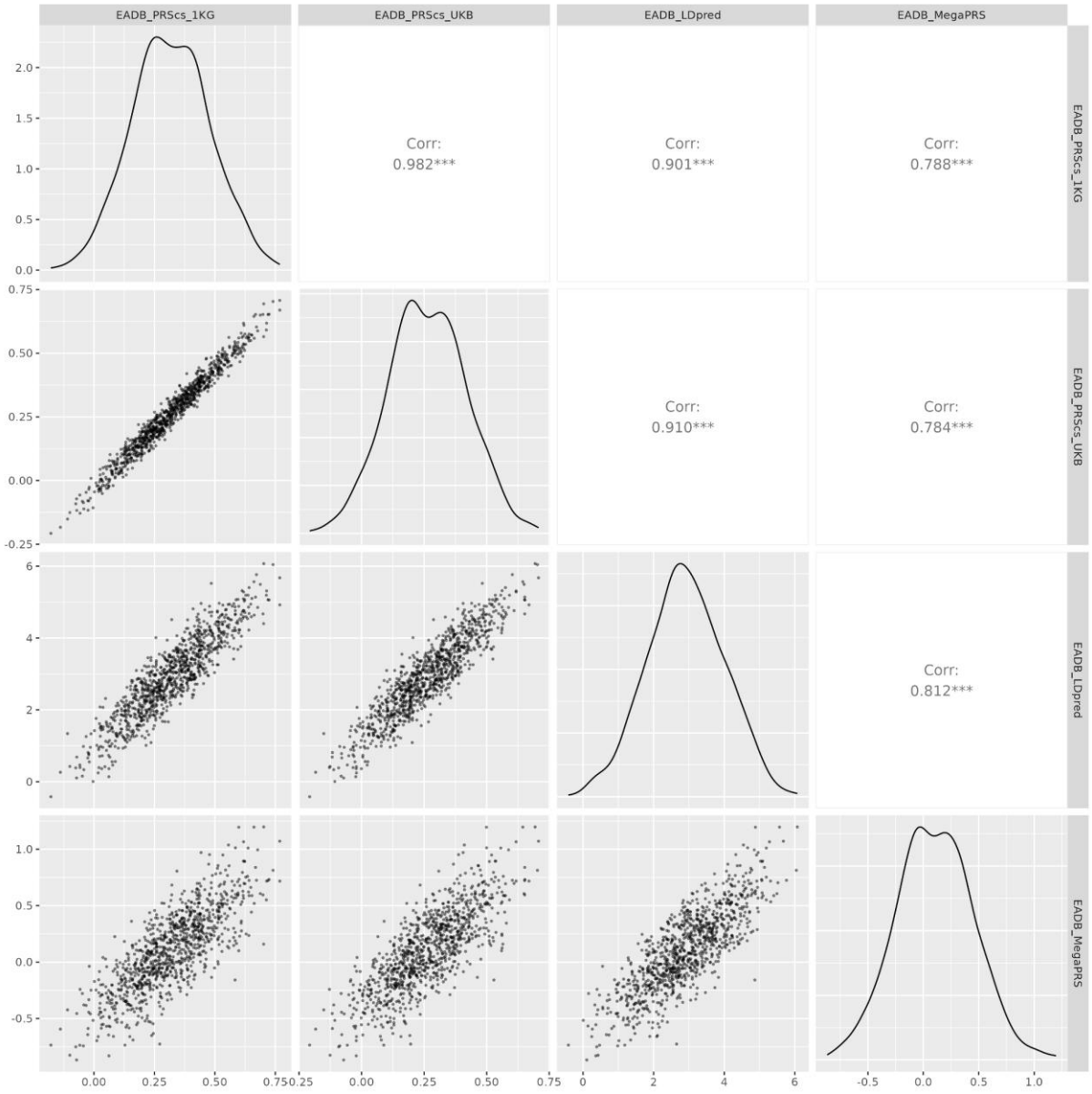


Figure S1. Pairwise comparison of individual's score calculated by PRS as M-whole between each method. Overall, the PRS were highly concordant across methods (Pearson's r = 0.78–0.98), with the strongest correlation observed between PRS-CS using the 1000 Genomes and UK Biobank LD reference panels (r = 0.98).

MODEL	Extreme	N cases	N control	p-value	AUC
PRS-CS 1KG	Low	6	14	\	0.674
	High	15	5	0.00781096	0.674
PRS-CS UKB	Low	6	14	\	0.697
	High	15	5	0.007566506	0.697
LDpred-inf	Low	6	14	\	0.699
	High	15	4	0.007063603	0.699
QuickPRS	Low	5	15	\	0.571
	High	10	10	0.1406845	0.571
RF	Low	8	12	\	0.618
	High	14	8	0.129	0.618
XGB	Low	5	15	\	0.702
	High	13	7	0.0138	0.702
CatBoost	Low	8	12	\	0.708
	High	16	4	0.0130	0.708
LGBM	Low	4	16	\	0.854
	High	18	2	0.000120	0.854
LR	Low	8	12	\	0.679
	High	15	5	0.0291	0.679
NB	Low	8	12	\	0.552
	High	84	51	0.0648	0.552

PRS-CS (ML setup)	Low	14	6	\	0.628
	High	9	11	0.114	0.628

Table S1: Model's performance in identifying individuals at high and low risk.

Extremes groups were defined by ranking individuals according to their prediction scores and selecting the top and bottoms 2% of the distribution.

In ML, each individual's prediction score corresponded to the mean prediction score across all replicates.

MODEL	APOE included	AUC mean	F1 mean	Precision mean	Recall mean	Sd AUC
CatBoost	Y	0.69	0.652	0.679	0.627	0.00551
LightGBM	Y	0.68	0.610	0.640	0.583	0.00482
LR	Y	0.56	0.528	0.514	0.542	0.0132
Naive Bayes	Y	0.57	0.538	0.536	0.540	0.00914
Random Forest	Y	0.59	0.497	0.566	0.442	0.0149
XGBoost	Y	0.67	0.619	0.638	0.600	0.00694
PRS-CS (ML setup)	Y	0.54	/	/	/	0.00577
CatBoost	N	0.544	0.457	0.517	0.409	0.00950
LightGBM	N	0.544	0.482	0.517	0.451	0.00769
LR	N	0.544	0.520	0.509	0.531	0.00931
Naive Bayes	N	0.559	0.530	0.527	0.533	0.0112

Random Forest	N	0.530	0.444	0.509	0.395	0.0136
XGBoost	N	0.543	0.619	0.500	0.518	0.0186
PRS-CS (ML setup)	N	0.53	/	/	/	0.0001

Table S2: Predictive performance of the ML models and PRS-CS in the M-APOE and M-noAPOE settings. Mean predictive performance of the machine learning models and PRS-CS with the ML setup when including (M-APOE) and excluding (M-noAPOE) APOE variants. Performance metrics were averaged across repeated cross-validation runs, while the reported SD represents the between-run standard deviation of the AUC calculated across different random seeds.

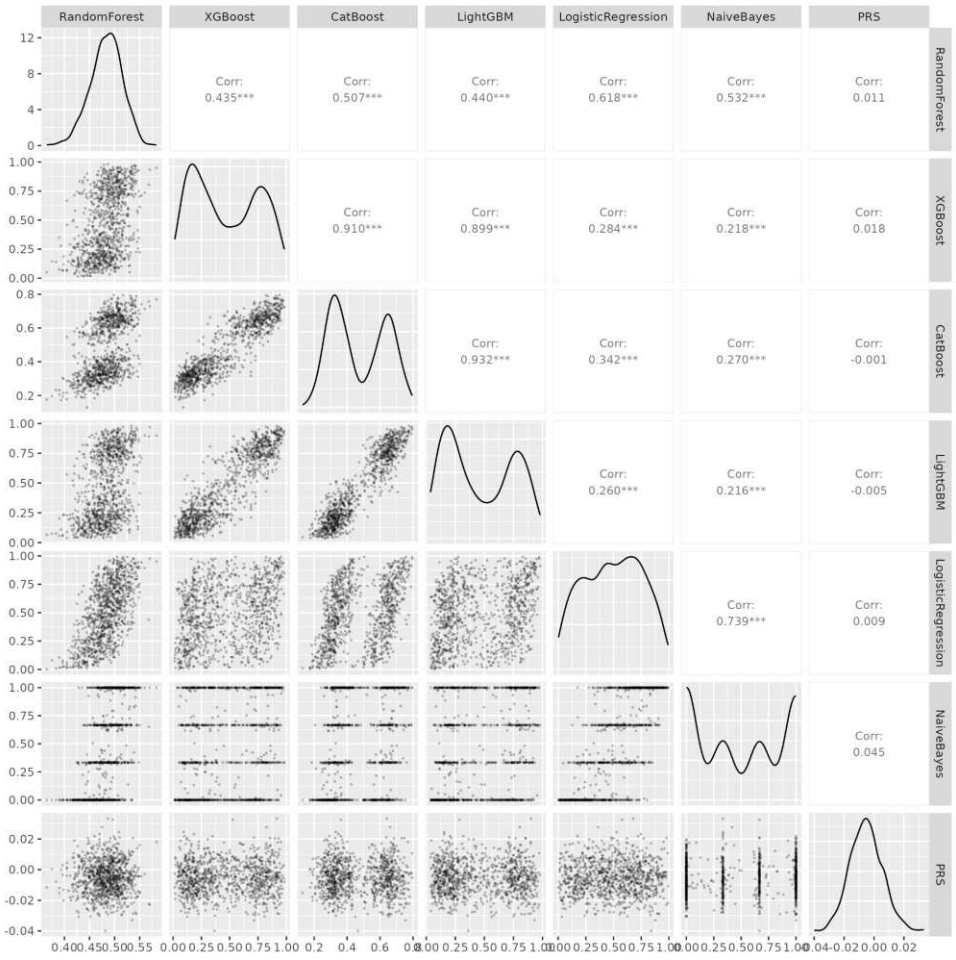


Figure S2: Pairwise comparison of an individual's score calculated by ML models as M-whole between each method. XGBoost, CatBoost and LightGBM showed high concordance (Pearson's r = 0.90–0.93), whereas Random Forest displayed only moderate correlations with the other tree-based models (r = 0.44–0.51). Logistic Regression and Naïve Bayes were moderately correlated with each other (r = 0.74), while the PRS showed little or no correlation with any of the machine learning models (r ≈ 0).

MODEL 1	MODEL 2	p-value
CatBoost	Logistic Regression	1.558331e-11 ****
XGBoost	Logistic Regression	1.387481e-10 ****
LightGBM	Logistic Regression	6.073516e-09 ****
CatBoost	Naive Bayes	3.146779e-08 ****
XGBoost	Naive Bayes	9.884754e-08 ****
LightGBM	Naive Bayes	1.433174e-06 ****
XGBoost	PRS-CS (1000G)	3.143783e-05 ****
Random Forest	CatBoost	4.151501e-05 ****
Random Forest	XGBoost	8.490373e-05 ****
Random Forest	Logistic Regression	1.358100e-04 ****
LightGBM	PRS-CS (1000G)	2.271296e-04 ****
LightGBM	Random Forest	9.088509e-04 **
Random Forest	Naive Bayes	1.080134e-02 *
Logistic Regression	Naive Bayes	1.034358e-01 ns
Logistic Regression	PRS-CS (1000G)	1.326195e-01 ns
XGBoost	LightGBM	2.053082e-01 ns
CatBoost	LightGBM	2.304282e-01 ns
Random Forest	PRS-CS (1000G)	3.504951e-01 ns

Naive Bayes	PRS-CS (1000G)	4.723021e-01 ns
XGBoost	CatBoost	8.302035e-01 ns

Table S3: DeLong test. Pairwise comparisons of AUC-ROC between the machine learning models and the state-of-the-art PRS-CS were performed using DeLong's two-sided test for correlated ROC curves. The comparisons shown here use the original state-of-the-art PRS-CS, rather than the PRS-CS evaluated under the machine learning experimental setup. Unadjusted *P*-values are reported. ns, $P \geq 0.05$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0005$.

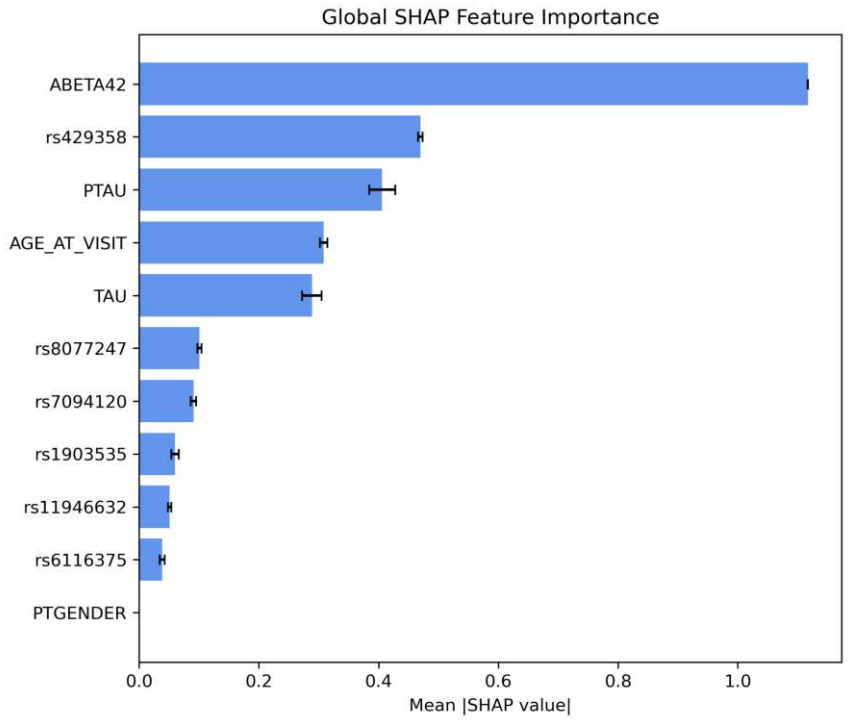


Figure S3: Global SHAP feature importance in the model including APOE variants. Mean absolute SHAP (SHAP) values for the six variants consistently selected from SHAP across repeated runs of the best-performing CatBoost model. Bars represent the mean absolute SHAP value across replicates, while error bars indicate the standard deviation. Features are ranked according to their mean contribution to the model predictions. The APOE variant rs429358 shows the highest contribution to the prediction.

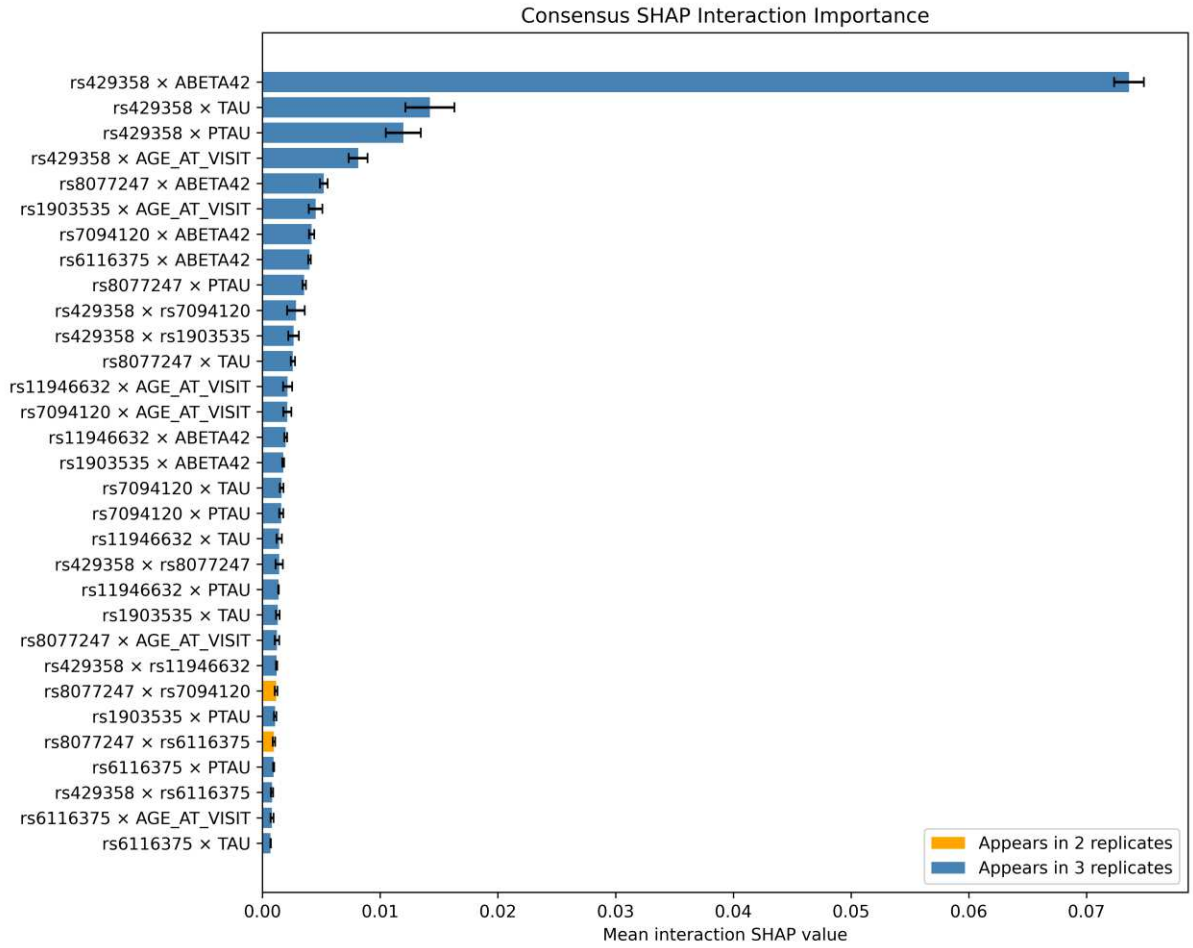


Figure S4: SHAP interaction values including the APOE variants in the model.

Mean SHAP interaction values for pairwise interactions consistently identified across repeated runs of the best-performing CatBoost model. Bars represent the mean interaction SHAP value across replicates, with error bars indicating the standard deviation. Interactions are ranked according to their mean interaction value. Colors indicate the number of replicates in which each interaction was identified (orange: 2 replicates; blue: 3 replicates).

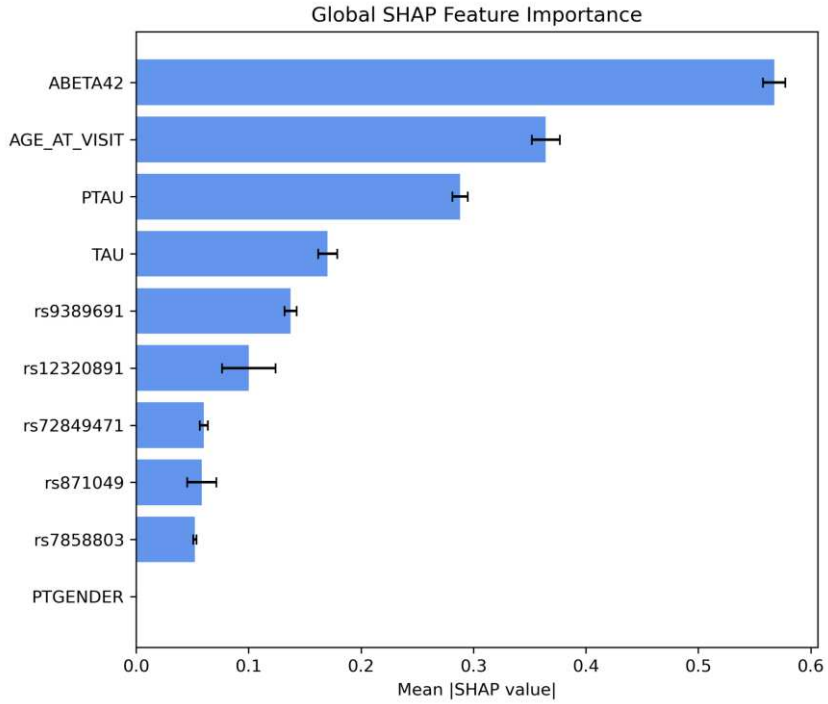


Figure S5: S Global SHAP feature importance in the model excluding APOE variants.

Mean absolute SHAP (SHAP) values for variants consistently identified by SHAP across repeated runs of the best-performing CatBoost model. Bars represent the mean absolute SHAP value across replicates, while error bars indicate the standard deviation. Features are ranked according to their mean contribution to the model predictions.

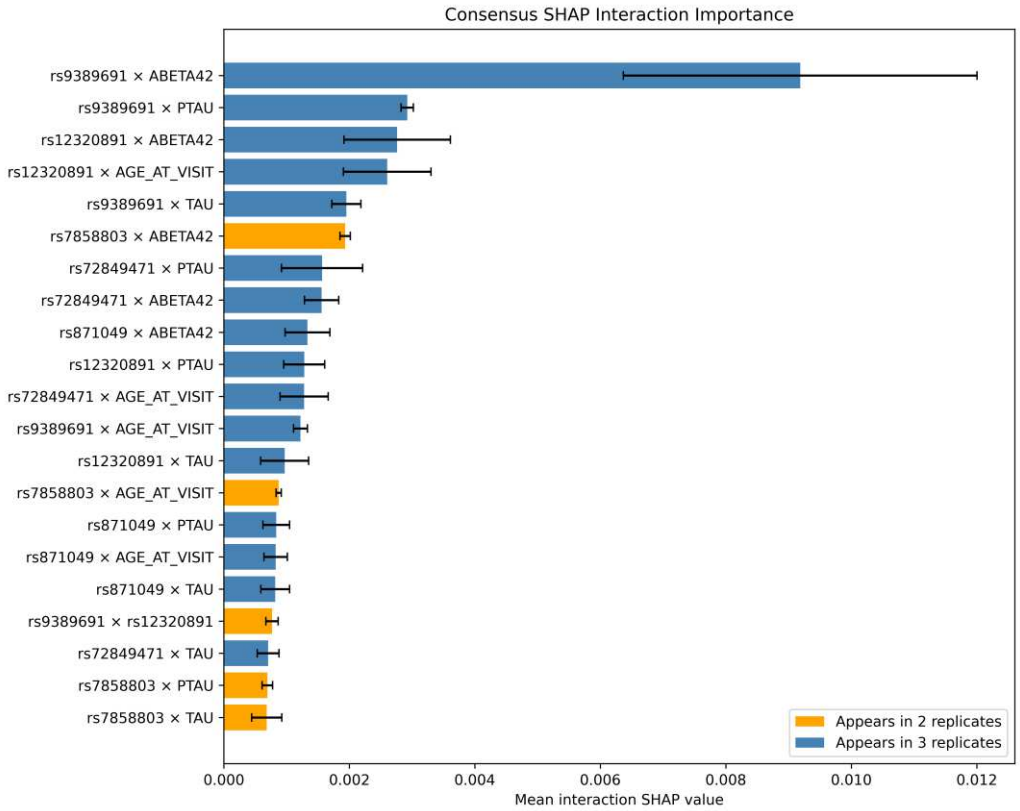


Figure 6: SHAP interaction values excluding APOE variants from the model

Mean SHAP interaction values for pairwise interactions consistently identified across repeated runs of the best-performing CatBoost model. Bars represent the mean interaction SHAP value across replicates, with error bars indicating the standard deviation. Interactions are ranked according to their mean interaction value. Colors indicate the number of replicates in which each interaction was identified (orange: 2 replicates; blue: 3 replicates).