Journal of High School Science

Predictive overlap, feature reduction, and robustness in Machine Learning-based cross-sectional classification of Alzheimer's disease

Xia M

Submitted: October 12, 2025, Revised: version 1, April 6, 2026, version 2, May 9, 2026
Accepted: May 9, 2026

Abstract
Alzheimer's disease (AD) is a progressive neurodegenerative disorder with substantial clinical and economic burden, making accurate and interpretable diagnosis classification an important goal. While beta amyloid and tau are central biomarkers, broader clinical datasets often contain partially overlapping predictors, raising questions about predictive overlap, interpretability, and robustness in machine learning (ML)-based classification. This study compared five ML models —logistic regression, support vector machine (SVM), extreme gradient boosting (XGBoost), ridge classifier, and k-nearest neighbors (KNN)—for cross-sectional AD diagnosis classification using 35,635 samples from the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS). Performance was evaluated across 50 repeated stratified splits using standard classification metrics. Predictive overlap was assessed using out-of-fold (OOF) single-feature Spearman correlation heatmaps, with supplementary descriptive metrics including mean pairwise absolute difference (MPAD), mean within-sample variance, and root mean square pairwise distance (RMSPD). Three experiments were conducted: baseline modeling, feature reduction to 13 variables, and Gaussian noise perturbation. XGBoost achieved the strongest overall performance, with the highest mean accuracy, strongest discrimination, and best calibration. After feature reduction, performance declined only modestly while preserving the overall ranking of algorithms. Under synthetic perturbation, most models remained stable, with KNN showing the greatest sensitivity to noise. Overall, the results suggest that predictive signal is distributed across partially substitutable features and that smaller, more interpretable feature sets can retain substantial classification performance in cross-sectional AD diagnosis. These findings further motivate future research examining whether redundancy patterns may provide additional insight into how AD-related classification signals are distributed across broader layers of clinical features.

Keywords
Alzheimer's disease, Machine Learning, Explainable artificial intelligence, Predictive modeling, Feature selection, Biomarker redundancy, Distributed disease representation, XGBoost, Clinical decision support systems, Neurodegenerative disease classification

Matthew Xia, Millburn High School, 462 Millburn Ave, Millburn, NJ 07041, USA. matthewxia27@gmail.com

https://doi.org/10.64336/001c.162155

1. Introduction

AD is a progressive neurodegenerative disorder characterized by memory loss, cognitive decline, and eventual loss of independent function (1). It places a substantial burden on patients, families, and health care systems. In the United States alone, AD and related dementias were associated with an estimated cost of $305 billion in 2020, and this burden is expected to increase substantially as the population ages (2). Because AD imposes such profound individual and societal costs, improving methods for identifying and classifying affected individuals remains an important clinical and public health goal.

Two of the most widely studied pathological hallmarks of AD are beta amyloid plaque accumulation and tau neurofibrillary tangles (1,3). These biomarkers have played a central role in research on disease detection and progression, and numerous studies have shown that abnormal amyloid and tau burden is associated with cognitive decline (1,3). However, although these biological pathways are important for understanding disease physiology, they do not fully resolve the challenge of classification in broader clinical settings. Their relevance can vary across individuals, and the timing and interaction of these pathological processes can differ substantially across stages of disease progression (1,3). As a result, there is increasing interest in whether broader clinical, behavioral, demographic, and health-history variables can complement biomarker-driven approaches (3).

In addition to molecular and imaging biomarkers, prior studies have explored the use of cognitive testing, clinical assessments, and machine learning models to distinguish individuals with AD from those without it in research and clinical datasets (4-6). Machine learning methods are especially attractive because they can incorporate many variables simultaneously and model relationships that may not be captured by traditional statistical approaches. Prior work has applied techniques such as logistic regression, SVMs, and tree-based methods to neuroimaging, biomarker, and clinical datasets in order to classify disease status or predict progression (4-6). However, many studies focus primarily on maximizing predictive accuracy and reporting feature importance, while paying less attention to whether the predictor set contains substantial predictive overlap or whether classification performance is robust to simplification of the feature space (4-6).

This issue matters because clinical AD datasets often contain partially overlapping variables. Multiple predictors may reflect related underlying domains, such as age, genetic susceptibility, vascular health, behavioral exposures, or broader clinical history. In such settings, using larger numbers of variables does not necessarily guarantee substantially better classification. Some predictors may provide overlapping predictive information, meaning that a reduced and more interpretable feature set could potentially retain much of the useful signal of the full dataset. At the same time, different algorithms may respond differently to partially overlapping predictors, and this may help explain why some models outperform others even when trained on the same data.

Accordingly, the goal of this study was not only to compare the classification performance of five machine learning algorithms, but also to examine whether useful signals in a clinical AD dataset remain stable when the feature space is simplified or perturbed. Using the NACC UDS (7,8), logistic regression (10), SVM (11), XGBoost (9), ridge classifier (12), and KNN (13) were evaluated across three experiments. The first established baseline model performance and identified the top-ranked predictors within each model. The second tested how performance changed after removal of overlapping predictors and reduction of the feature space. The third evaluated robustness to incremental Gaussian noise in order to determine how stable each algorithm remained under controlled synthetic perturbation. This study asks whether cross sectional AD classification depends on a large predictor set or whether much of the usable signal is distributed across overlapping, partially substitutable predictors.

It was hypothesized that XGBoost would provide the strongest overall discrimination because of its ability to capture nonlinear relationships and heterogeneous predictor effects in large clinical datasets (9). It was also hypothesized that removing overlapping predictors would produce only modest declines in performance, indicating that useful signal is distributed across partially substitutable variables rather than concentrated in a very small set of uniquely indispensable predictors. The novelty of this study lies in its leakage-aware comparison of model performance together with predictive overlap, feature reduction, and robustness analyses in a large clinical AD dataset.

2. Methods

2.1 Literature review

A targeted literature review was conducted using Web of Science and Google Scholar, focusing primarily on studies from the past 10 years related to AD classification, biomarkers, machine learning, cognitive decline, and predictive modeling. Search terms included "Alzheimer's disease," "biomarkers," "artificial intelligence," "cognitive decline," "classification," and "predictive modeling." The review was intended to contextualize existing work in AD machine learning and identify gaps related to predictive overlap, feature reduction, and robustness.

2.2 Dataset and variables

For the modeling analyses, the UDS downloaded from NACC was used. The UDS contains longitudinal data collected since 2005 during standardized annual evaluations conducted at National Institute on Aging-funded Alzheimer's Disease Research Centers (ADRCs) across the United States. Participants represent the Clinical Core enrollment of the ADRCs, with cognitive status ranging from dementia to mild cognitive impairment to cognitively normal (7,8). The final analytic dataset contained 35,635 samples after preprocessing, leakage control, and complete case filtering. Predictor variables included demographic, behavioral, vascular, psychiatric, and genetic variables retained after removal of direct diagnostic leakage features and administrative identifiers. Descriptive characteristics of the final analytic cohort are

shown in Table 1.

Table 1. Characteristics of variables included in the final modeling dataset after preprocessing, leakage control, and complete-case filtering. Values are reported as mean ± standard deviation for continuous variables and n (%) for categorical variables. Summaries are based on the baseline analytic dataset. For continuous variables, obvious sentinel or nonresponse codes were excluded before calculation. For selected multi-level history variables, affirmative history was collapsed into a single category for presentation.

Characteristic	Overall (N = 35,635)	AD = 0	AD = 1
Demographics			
Sample size	35,635	9,914	25,721
Age at visit, years	73.77 ± 10.10	70.98 ± 10.51	74.84 ± 9.73
Female sex, n (%)	18,878 (53.0%)	4,789 (48.3%)	14,089 (54.8%)
Years of education	14.91 ± 3.57	14.95 ± 3.43	14.90 ± 3.62
Body mass index, kg/m²	26.78 ± 5.22	27.78 ± 5.77	26.45 ± 4.99
Systolic blood pressure, mmHg	134.91 ± 19.18	133.47 ± 19.32	135.38 ± 19.12
Diastolic blood pressure, mmHg	75.29 ± 10.73	75.88 ± 11.04	75.09 ± 10.62
Lifestyle and sleep history			
Current tobacco use, n (%)	1,386 (3.9%)	450 (4.5%)	936 (3.6%)
Lifetime tobacco exposure (100+ cigarettes), n (%)	12,678 (35.6%)	2,973 (30.0%)	9,705 (37.7%)
Occasional alcohol use, n (%)	5,066 (14.2%)	1,021 (10.3%)	4,045 (15.7%)
Sleep apnea history/presence, n (%)	1,971 (5.5%)	585 (5.9%)	1,386 (5.4%)
REM sleep behavior disorder history/presence, n (%)	485 (1.4%)	186 (1.9%)	299 (1.2%)
Insomnia history/presence, n (%)	1,474 (4.1%)	492 (5.0%)	982 (3.8%)
Family history and comorbidities			
Family history of dementia, n (%)	17,959 (50.4%)	4,280 (43.2%)	13,679 (53.2%)
Maternal history of dementia, n (%)	10,935 (30.7%)	2,666 (26.9%)	8,269 (32.1%)
Paternal history of dementia, n (%)	5,477 (15.4%)	1,367 (13.8%)	4,110 (16.0%)
Diabetes history/presence, n (%)	4,462 (12.5%)	1,181 (11.9%)	3,281 (12.8%)
Hypertension history/presence, n (%)	16,189 (45.4%)	3,270 (33.0%)	12,919 (50.2%)
Hypercholesterolemia history/presence, n (%)	16,133 (45.3%)	3,440 (34.7%)	12,693 (49.3%)
Atrial fibrillation, n (%)	1,114 (3.1%)	396 (4.0%)	718 (2.8%)
History of stroke, n (%)	1,900 (5.3%)	586 (5.9%)	1,314 (5.1%)
History of transient ischemic attack, n (%)	1,717 (4.8%)	407 (4.1%)	1,310 (5.1%)
Traumatic brain injury history, n (%)	1,557 (4.4%)	398 (4.0%)	1,159 (4.5%)
Depression diagnosis, n (%)	7,742 (21.7%)	2,489 (25.1%)	5,253 (20.4%)
Depression in last 2 years, n (%)	10,949 (30.7%)	2,842 (28.7%)	8,107 (31.5%)
Geriatric Depression Scale score	2.72 ± 2.82	3.52 ± 3.34	2.45 ± 2.57
Genetics and biomarkers			
APOE ε4 carrier (≥1 allele), n (%)	12,353 (34.7%)	2,189 (22.1%)	10,164 (39.5%)

2.3 Model descriptions

Logistic regression was used as a baseline comparison model. Logistic regression is a probabilistic statistical model that estimates the likelihood of membership in one of two classes using maximum likelihood estimation (10). In this study, the model estimates the probability that an individual belongs to the AD class based on the retained predictors.

Next, an SVM model was tested. This algorithm constructs a hyperplane that maximizes the margin between two classes (11). The objective is to maximize the distance between the hyperplane and the closest data points from each class, known as support vectors (11). Based on this boundary, the SVM classifies observations into outcome groups. In this study, the SVM model was used to determine which predictor patterns most strongly differentiate individuals with and without AD.

The third model implemented was XGBoost. XGBoost operates by combining many decision trees to form a stronger predictive model (9). It builds trees sequentially, where each new tree attempts to correct the prediction errors made by the previous ones (9). XGBoost is particularly effective at capturing nonlinear relationships and complex interactions among predictors in large datasets.

Fourth, ridge classifier was implemented. Ridge-based classification applies L2 regularization to shrink coefficients and reduce sensitivity to multicollinearity while retaining all predictors in the model (12). This can improve stability in settings with partially overlapping variables.

Fifth, the KNN classification algorithm was implemented. This algorithm assumes that observations with similar characteristics are located near each other in feature space (13). The model assigns a class label to a data point based on the most common label among its nearest neighbors (13). Because KNN is sensitive to feature scale, variables were standardized prior to model fitting.

To characterize the predictors emphasized by each algorithm, top-ranked features were summarized separately for each model using model-appropriate importance measures. XGBoost importance was derived from native tree-based importance values. Logistic regression, ridge classifier, and linear SVM used coefficient magnitude, whereas KNN used permutation importance (14). Because feature-importance definitions differ across model families, these results were interpreted qualitatively in terms of within-model ranking and recurrence across models rather than as directly comparable common-scale effect sizes.

2.4 Predictive overlap diagnostics

In this study, predictive overlap refers to the extent to which different predictors produce similar model outputs when used independently, suggesting that they may convey partially overlapping predictive information. Rather than treating overlap solely as numerical similarity of raw feature values, the primary focus was overlap in predictor-specific model behavior. Operationally, this means overlap is defined by similarity in participant-level prediction patterns from

single-feature models, not merely by correlation between raw feature values. Two variables can therefore differ in scale or clinical meaning yet still show predictive overlap if they push the model toward similar outputs across participants.

2.4.1 *Out-of-Fold predictive overlap heatmaps*

For each model, predictive overlap was assessed using OOF predictions generated from single-feature models. Using 5-fold stratified cross-validation, each feature was used independently to generate OOF prediction scores. Spearman correlations were then computed between the OOF prediction vectors for all pairs of features, producing a predictive-overlap heatmap for each model (15). Higher positive correlations indicate that two features, when used independently, lead the model toward similar outputs across observations and therefore may carry overlapping predictive signals. These heatmaps, shown later in Figures 4 and 8, were interpreted qualitatively as evidence of feature substitutability and were used to contextualize the feature-reduction experiment. Because these predictions are generated on data not used to train the single-feature models, they provide a more reliable picture of how each predictor behaves on unseen data.

2.4.2 *Supplementary descriptive redundancy summaries*

To complement the prediction-based heatmaps, three descriptive similarity summaries were also calculated for the top-ranked predictors: mean within-sample variance, MPAD, and RMSPD. These metrics were used only as secondary descriptive summaries of similarity structure and were not treated as formal standalone measures of shared predictive information. In simple terms, MPAD describes the average difference in how two predictors separate participants, whereas RMSPD is similar but gives more weight to larger mismatches.

2.4.3 *Scalar summary of predictive overlap*

To summarize overall overlap within each heatmap, the mean absolute off-diagonal correlation among the single-feature OOF prediction vectors was additionally computed. This scalar was used only as a descriptive summary of the heatmap structure rather than as a standalone formal proof of redundancy.

2.4.4 *Collinearity diagnostics*

As a supplementary structural diagnostic, collinearity among retained predictors was evaluated using standard training-data-based diagnostics where appropriate (16). These diagnostics were interpreted cautiously as complementary information about feature overlap rather than as the primary basis for the paper's conclusions.

2.5 Performance metrics and statistical analysis

To reduce dependence on any single train-test split while keeping computation feasible, the 80/20 stratified split procedure was repeated across 50 random seeds and performance was summarized across runs. For each run, accuracy, sensitivity, specificity, precision, F1 score, balanced accuracy, receiver operating characteristic area under the curve (ROC-AUC), precision-recall area under the curve (PR-AUC), and Brier score were computed (17,18).

Because the dataset is class-imbalanced and the models consistently showed high sensitivity with lower specificity, overall interpretation emphasized the sensitivity/specificity tradeoff, balanced accuracy, discrimination, and calibration rather than accuracy alone. The dataset contained more AD cases than non-AD controls, which likely contributed to the observed sensitivity-specificity tradeoff and reinforces the importance of balanced accuracy and related class-aware metrics. Calibration refers to how closely predicted probabilities match observed outcome frequencies; for example, among participants assigned a predicted probability near 0.80, a well-calibrated model would classify roughly 80% as AD-positive (18). The Brier score summarizes the mean squared difference between predicted probabilities and observed outcomes, so lower values indicate better probabilistic accuracy (18).

A fixed probability threshold of 0.5 was used for all classifiers to maintain a consistent and conventional decision rule across models. Because the study's primary goal was comparative evaluation of discrimination, feature reduction, and robustness under a common pipeline, threshold tuning was not performed. Alternative thresholds could shift the sensitivity-specificity balance, especially in an imbalanced dataset, but threshold optimization was outside the scope of this study.

For paired statistical comparisons under identical test conditions, a fixed canonical 80/20 train-test split was used. Differences in classification error patterns between models were evaluated using McNemar's test on paired predictions, and differences in ROC-AUC between correlated classifiers were assessed using DeLong's test (19,20). These pairwise tests were used to support interpretation of the canonical-split comparisons but were not the primary focus of the paper.

Table 2. Hyperparameter grids used in GridSearchCV for logistic regression, ridge classifier, XGBoost, SVM, and KNN.

Model	Hyperparameter Grid Used
Logistic Regression	C: [0.1, 1.0, 10.0]; penalty: ['l2']; solver: ['lbfgs']
Ridge Classifier	alpha: [0.1, 1.0, 10.0]
XGBoost	n_estimators: [150, 300]; max_depth: [3, 5]; learning_rate: [0.05, 0.1]; subsample: [0.8, 1.0]; colsample_bytree: [0.8, 1.0]
SVM	kernel: ['linear', 'rbf', 'poly']; C: [0.5, 1.0, 3.0]; gamma: ['scale', 'auto']; degree: [2, 3]
KNN	n_neighbors: [3, 5, 9, 15]; weights: ['uniform', 'distance']; p: [1,2]

Model calibration was assessed on the canonical test split using both calibration curves and Brier scores, allowing comparison between predicted probabilities and observed outcome frequencies (18). Hyperparameters for each model were optimized using GridSearchCV with five-fold cross-validation on the training data, with model-specific parameter grids and ROC-AUC-based selection (21). This allowed comparisons to be based on

tuned implementations within a practical search space rather than untuned defaults. The hyperparameter grids are summarized in Table 2.

2.6 Preprocessing and experimental design

The initial NACC dataset contained 207,453 clinical visits, with multiple visits pertaining to the same patient (n = 55,268 unique individuals) (7,8). To construct a modeling-ready cohort, the dataset was aggregated to one row per individual, and each participant was assigned a single AD status label based on the first recorded AD diagnosis or, if no AD diagnosis occurred, the last available visit. The outcome variable (NACCALZD) was then restricted to a binary diagnosis classification setting by removing entries coded as unknown or unavailable, resulting in 35,874 individuals and 64 variables.

To reduce label leakage, variables judged to be direct diagnostic proxies, formal cognitive staging measures, explicit diagnostic summary measures, and subject identifiers were removed through manual column-by-column review. Complete-case filtering was then applied, excluding individuals with missing values in any modeling variable. This reduced the dataset from 35,874 to 35,635 participants, producing the final analytic cohort. A structured feature-selection and cleanup procedure was then applied to remove administrative fields, leakage-prone variables, and structurally overlapping encodings, while preserving clinically meaningful domains such as demographic factors, vascular comorbidities, psychiatric indicators, behavioral variables, and genetic markers. This process produced the feature sets used for the baseline and reduced-feature experiments.

This study was designed as a cross-sectional diagnosis-classification analysis rather than a longitudinal prediction of future AD onset or conversion. Model evaluation used repeated stratified train-test splits across 50 random seeds (0-49). For each seed, the dataset was divided into 80% training and 20% testing subsets while preserving class distributions. Performance metrics were calculated for each run and summarized as mean and standard deviation across 50 seeds. For models sensitive to feature scaling, preprocessing was performed within the training data using StandardScaler. Hyperparameters for each model were optimized using GridSearchCV with five-fold cross-validation on the training data (21).

3. Results

3.1 Experiment 1: baseline model performance

The first experiment evaluated classification performance across five machine learning algorithms trained on the baseline NACC dataset. Models were assessed using 5-fold cross validation and 80/20 stratified train-test splits across 50 random seeds. Performance was evaluated using F1 score, sensitivity, specificity, and measures of discrimination and calibration; on a canonical split, additional metrics, including precision, balanced accuracy, ROC-AUC, PR-AUC, and Brier score, were computed.

Across all runs, XGBoost achieved the highest mean classification accuracy (0.7746 ±

0.0045), outperforming logistic regression (0.7614 ± 0.0036), ridge classifier (0.7609 ± 0.0033), SVM (0.7397 ± 0.0035), and KNN (0.7397 ± 0.0038). Logistic regression and ridge classifier showed nearly identical performance, whereas SVM and KNN formed the lower-performing group. These repeated-split accuracy results are shown in Figure 1.

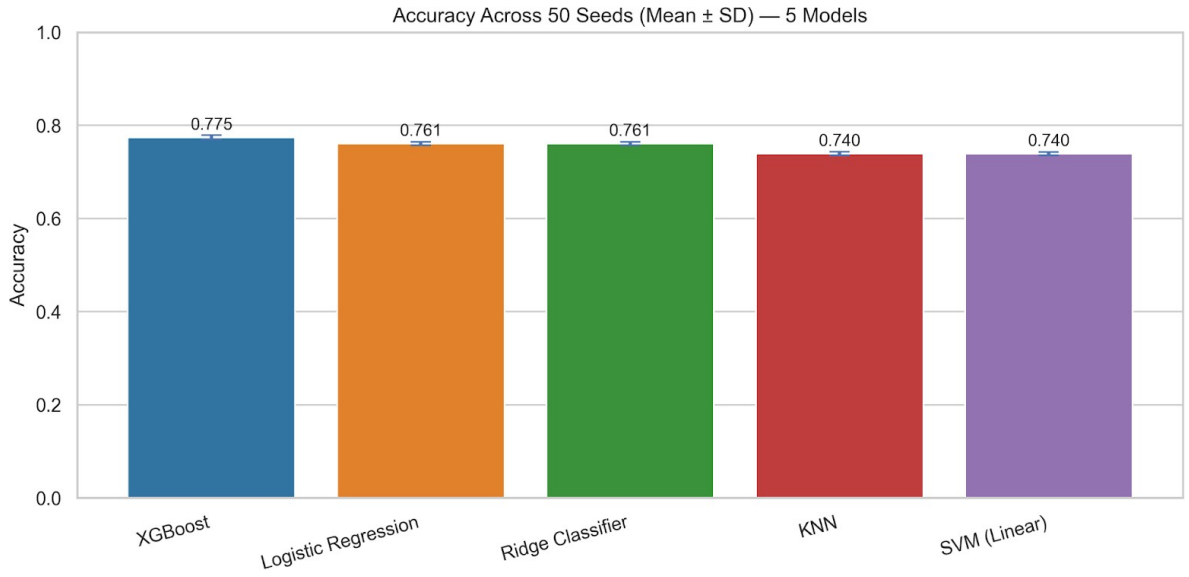

Figure 1. Mean classification accuracy across 50 random seeds for XGBoost, logistic regression, ridge classifier, KNN, and linear SVM on the baseline NACC dataset. Accuracy is shown as a repeated-split stability summary; balanced accuracy and canonical discrimination and calibration metrics are reported in Table 3. Error bars indicate standard deviation across seeds.

All models exhibited consistently high sensitivity but comparatively low specificity, indicating a systematic tendency to prioritize identification of AD-positive cases over correct classification of AD-negative cases. As a result, performance is more appropriately interpreted using sensitivity-specificity tradeoffs, balanced accuracy, discrimination, and calibration rather than accuracy alone. XGBoost demonstrated the strongest overall performance, achieving the highest ROC-AUC (0.775) and PR-AUC (0.887), the lowest Brier score (0.158), and the highest balanced accuracy on the canonical split (0.652), reflecting superior discrimination and calibration. For example, its sensitivity reached 0.929 while specificity remained 0.374, illustrating the broader pattern observed across models. This imbalance suggests the classifiers are better suited for screening applications, where high sensitivity reduces missed cases, but the elevated false positive rate necessitates confirmatory clinical evaluation. Canonical performance metrics for the baseline experiment are shown in Table 3.

Table 3. Canonical performance metrics for the five machine learning models in Experiment 1. Metrics include balanced accuracy, ROC-AUC, PR-AUC, Brier score, F1 score, sensitivity, specificity, and precision. Higher values indicate better performance for all metrics except Brier score, where lower values indicate better calibration.

Model	Balanced Accuracy	ROC–AUC	PR–AUC	Brier Score	F1 Score	Sensitivity	Specificity	Precision
XGBoost	0.652	0.775	0.887	0.158	0.856	0.929	0.374	0.794
Logistic Regression	0.612	0.725	0.849	0.171	0.848	0.941	0.283	0.773
Ridge Classifier	0.606	0.725	0.848	0.176	0.850	0.948	0.263	0.770
KNN	0.609	0.672	0.807	0.196	0.829	0.892	0.326	0.774
SVM (Linear)	0.605	0.692	0.812	0.187	0.833	0.905	0.305	0.772

Calibration curves in Experiment 1 showed that XGBoost and logistic regression were the most well-calibrated overall, indicating the strongest agreement between predicted probabilities and observed outcome frequencies. Ridge classifier showed reasonable alignment across much of the probability range but deviated in mid-to-high probability bins, suggesting less stable calibration. KNN tended to underestimate risk across portions of the curve, while linear SVM showed the most irregular calibration pattern. These findings were consistent with the Brier scores in Table 3 and supported XGBoost as the best-calibrated model in the baseline experiment. The canonical calibration curves are shown in Figure 2.

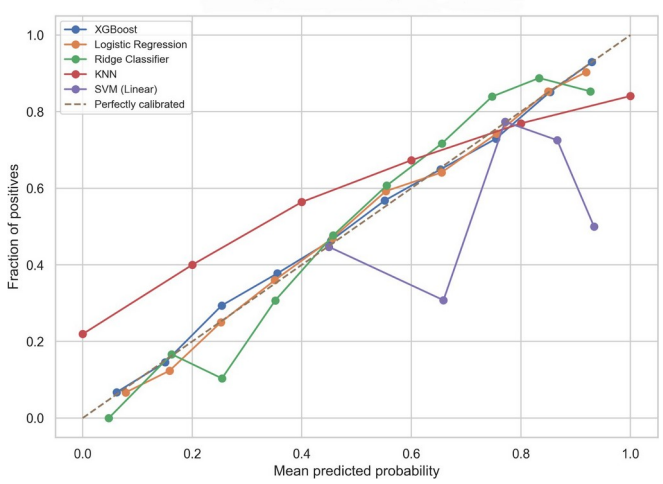

Figure 2. Calibration curves for the five machine learning models on the canonical test split in Experiment 1 using the baseline dataset.The dashed diagonal represents perfect calibration, where predicted probabilities exactly match observed outcome frequencies.

Top-ranked predictors were summarized separately for each model. Several variables appeared repeatedly among the top-ranked predictors, including age (NACCAGE), APOE-

related genotype measures (NACCNE4S and NACCAPOE), and several health- or behavior-related variables. This recurrence suggests that both established genetic risk factors and broader clinical or behavioral variables contributed meaningfully across models. These within-model top-ranked predictors are shown in Figure 3.

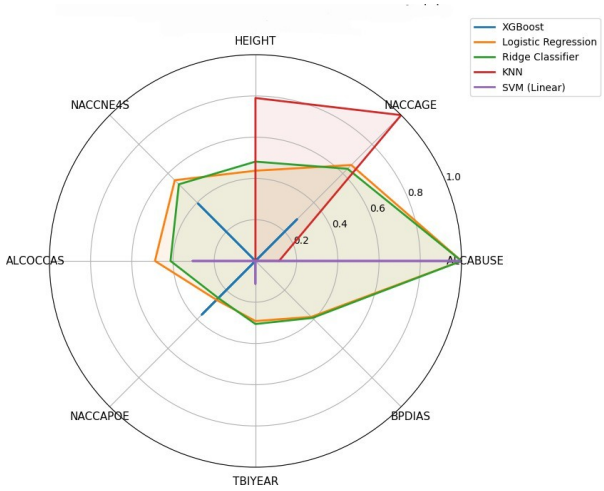

Figure 3. Radar chart of key recurring features across 5 models (normalized within model) HEIGHT. Top-ranked predictors identified using the baseline NACC dataset. Rankings were interpreted qualitatively within models rather than as directly comparable common-scale effect sizes across algorithms. Abbreviations key in Appendix.

To provide a compact descriptive summary of predictor similarity, three supplementary redundancy metrics were calculated on the top-ranked predictors for each model: mean within-sample variance, MPAD, and RMSPD. Across models, logistic regression and ridge classifier produced nearly identical redundancy values, whereas XGBoost and KNN showed slightly larger divergence values on some measures. Linear SVM produced the lowest overall divergence values, indicating a more compact predictor-output structure under that model. These descriptive summaries are reported in Table 4.

Table 4. Descriptive redundancy summary metrics for the top-ranked predictors in Experiment 1. Metrics include mean within-sample variance, MPAD, and RMSPD. These values are reported as supplementary descriptive summaries of predictor similarity structure rather than formal standalone measures of shared predictive information.

Model	Mean Within-Sample Variance	MPAD Distance	RMSPD Distance
XGBoost	0.819	**1.026**	1.349
Logistic Regression	0.780	0.962	1.317
Ridge Classifier	0.780	0.962	1.317
KNN	**0.873**	0.994	**1.393**
SVM (Linear)	0.448	0.597	0.998

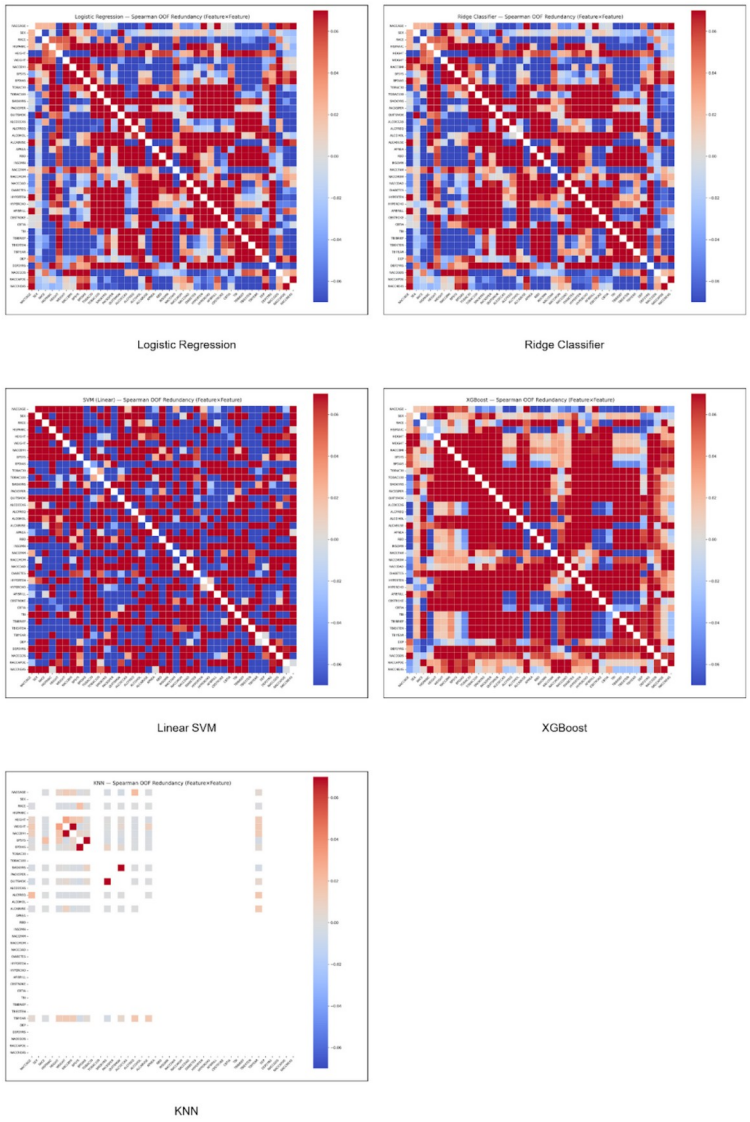

Figure 4. OOF Spearman predictive-overlap heatmaps for the top-ranked predictors in Experiment 1. Each cell represents the Spearman correlation between single-feature OOF prediction vectors, such that stronger positive correlations indicate greater overlap in predictor-specific model outputs. Abbreviations key in Appendix.

Predictive overlap among the most influential predictors was evaluated using OOF Spearman correlation heatmaps based on the output probabilities of single-feature models. These heatmaps represent the extent to which individual predictors, when used independently, generate similar model outputs across participants. Across models, the overall correlational structure suggested that different algorithms were often drawing from partially overlapping inputs even when their final performance differed. The baseline predictive-overlap heatmaps are shown in Figure 4. Exploratory model-level regressions were

examined but are not central to the present study.

3.2 Experiment 2: model performance after removing overlapping predictors

The second experiment evaluated model performance after removal of overlapping predictors, yielding a reduced dataset containing 13 predictors. One representative variable from each of 12 clinical categories was selected using a two-step process. First, category-level ablation results were used to identify the most predictive features within each category, based on the magnitude of performance degradation (ΔAUC) upon removal. Second, among top-performing candidates, clinical relevance and interpretability were used to guide the final selection, ensuring that retained variables were both predictive and clinically meaningful rather than relying on ablation alone. In cases where multiple features within a category exhibited similar predictive contributions, preference was given to variables with clearer clinical interpretability or more widespread use in practice. For the APOE category, both APOE genotype (NACCAPOE) and number of ε4 alleles (NACCNE4S) were retained due to their distinct clinical interpretations and complementary representation of genetic risk, yielding a total of 13 predictors.

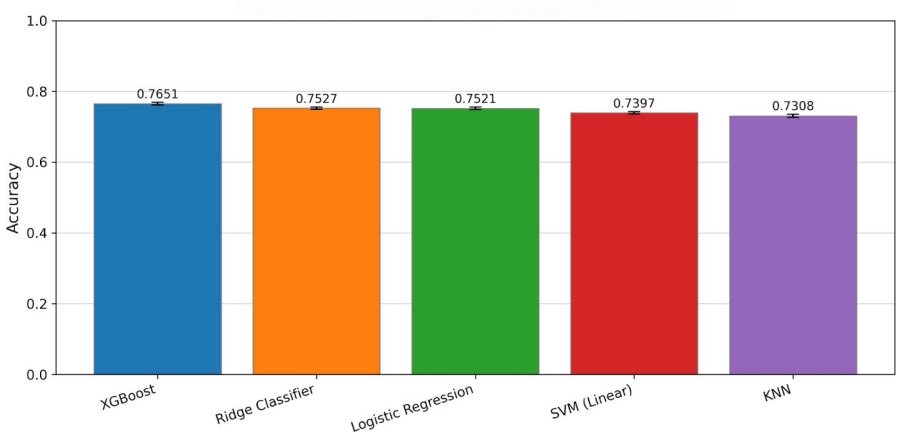

Figure 5. Mean classification accuracy across 50 random seeds for XGBoost, ridge classifier, logistic regression, KNN, and linear SVM in the reduced 13-variable dataset. Accuracy is shown as a repeated-split stability summary; balanced accuracy and canonical discrimination and calibration metrics are reported in Table 5. Error bars indicate standard deviation across seeds.

Using the same repeated 80/20 stratified train-test split framework across 50 random seeds with five-fold cross-validation, model performance was again interpreted using the prior metrics. Across all runs, XGBoost again achieved the highest mean classification accuracy (0.7651 ± 0.0037), outperforming ridge classifier (0.7527 ± 0.0032), logistic regression (0.7521 ± 0.0034), SVM (0.7397 ± 0.0035), and KNN (0.7308 ± 0.0044). Relative to the baseline experiment, performance declined modestly across all models following

feature reduction; however, the overall ranking of algorithms remained largely unchanged. These findings indicate that substantial simplification of the feature space reduced performance only modestly while preserving the broader comparative pattern across models. Repeated-split mean accuracy results for the reduced-feature experiment are shown in Figure 5.

As in Experiment 1, sensitivity remained substantially higher than specificity after feature reduction. For example, XGBoost achieved a sensitivity of 0.936 and a specificity of 0.323 on the canonical split. This again indicates screening-oriented classification behavior, with stronger case detection than exclusion of negative cases. Although this pattern supports high case sensitivity, the accompanying false-positive burden means that such predictions should be interpreted as supportive classification outputs rather than definitive standalone clinical decisions. When sensitivity and specificity were considered jointly, XGBoost also retained the highest balanced accuracy in the reduced-feature setting (0.630), supporting its strongest overall performance after feature reduction. Canonical performance metrics for the reduced-feature experiment are shown in Table 5.

Table 5. Canonical performance metrics for the five machine learning models in Experiment 2 after reduction to 13 variables. Metrics include balanced accuracy, ROC-AUC, PR-AUC, Brier score, F1 score, sensitivity, specificity, and precision. Higher values indicate better performance for all metrics except Brier score, where lower values indicate better calibration.

Model	Balanced Accuracy	ROC–AUC	PR–AUC	Brier Score	F1 Score	Sensitivity	Specificity	Precision
XGBoost	0.630	0.754	0.875	0.165	0.852	0.936	0.323	0.782
Logistic Regression	0.600	0.700	0.827	0.177	0.847	0.948	0.251	0.766
Ridge Classifier	0.596	0.698	0.826	0.232	0.848	0.953	0.238	0.764
KNN	0.615	0.673	0.807	0.197	0.827	0.883	0.347	0.778
SVM (Linear)	0.605	0.624	0.767	0.188	0.833	0.905	0.305	0.772

Calibration curves in the reduced-feature setting showed that XGBoost remained the model whose predicted probabilities most closely tracked the ideal calibration line overall, indicating the strongest agreement between predicted and observed outcome frequencies after reduction to 13 variables. Logistic regression also showed relatively good alignment across much of the probability range, whereas ridge classifier displayed greater deviation from the diagonal in several bins, suggesting less stable calibration. Linear SVM showed irregular departures from the ideal line, while KNN again tended to underestimate risk across portions of the curve. Overall, these visual findings were consistent with the Brier scores reported in Table 5, which again identified XGBoost as the best-calibrated model in Experiment 2. Notably, calibration patterns differed across models and also diverged from those observed when using the full set of predictors, indicating that feature

reduction alters probability calibration behavior in a model-dependent manner. The calibration curves for the reduced-feature experiment are shown in Figure 6.

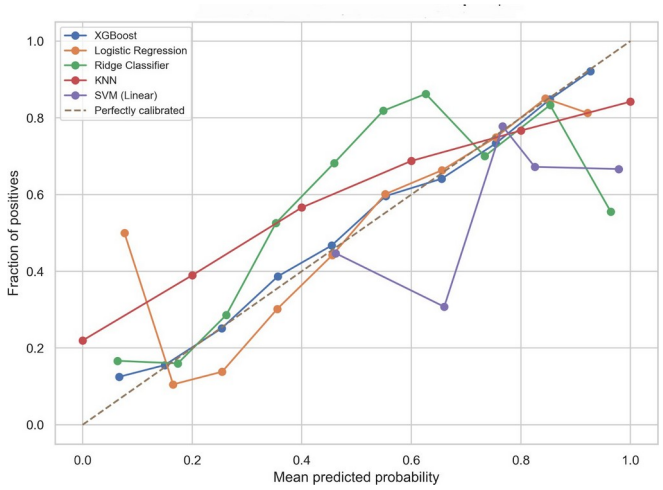

Figure 6. Calibration curves for the five machine learning models on the canonical test split in Experiment 2 after reduction to 13 variables. The dashed diagonal represents perfect calibration.

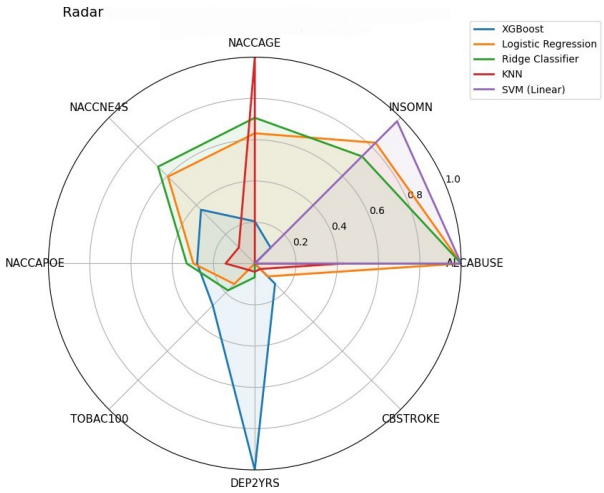

Figure 7. Radar chart of key recurring features across 5 models (normalized within model). Top-ranked predictors identified using the reduced dataset. Rankings were interpreted qualitatively within models rather than as directly comparable common-scale effect sizes across algorithms. Abbreviations key in Appendix.

To examine how signal was redistributed after feature reduction, top-ranked predictors were again summarized separately within each model. Several variables recurred among the highest-ranked predictors, including alcohol abuse, age, APOE-related measures, insomnia, smoking exposure, body mass index, maternal dementia history, stroke-related history, and traumatic brain injury. Relative to the baseline full-feature analysis, the reduced-feature setting

showed greater overlap in the identities of the top-ranked predictors across models, suggesting a concentration of predictive reliance onto a smaller core set of informative variables. In contrast, the full-feature models exhibited greater dispersion in feature importance, with different models prioritizing partially redundant predictors within the same clinical domains. This broader distribution of importance in the full setting is consistent with the presence of substitutable features, where overlapping predictors share predictive signal. After feature reduction, this redundancy was largely compressed, forcing models to converge on a more consistent set of key variables while maintaining comparable overall performance. The reduced-feature top predictors are shown in Figure 7.

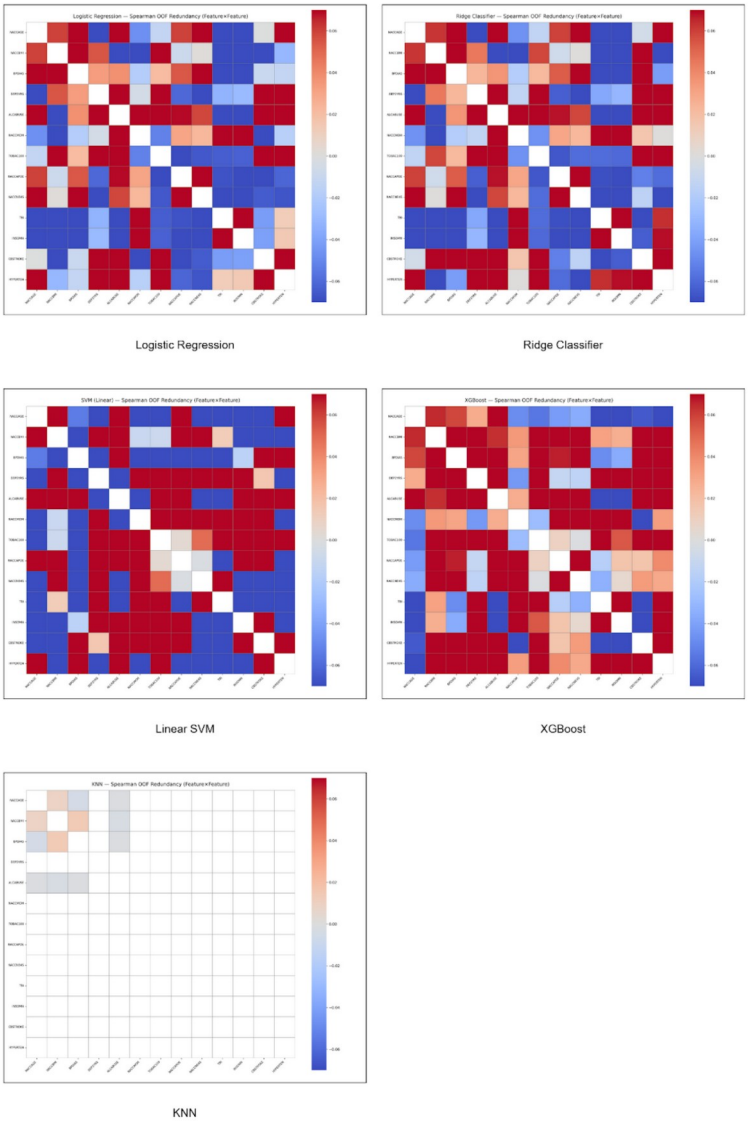

Figure 8. OOF Spearman predictive-overlap heatmaps for retained predictors in Experiment 2. Stronger positive correlations indicate greater overlap in predictor-specific model outputs. Abbreviations key in Appendix.

Predictive-overlap heatmaps in the reduced-feature setting showed that feature reduction narrowed but did not eliminate clusters of overlapping signals. In particular, APOE-related variables and the pairing of traumatic brain injury with insomnia remained visually prominent sources of overlap in several models. These findings suggest that removal of overlapping variables compressed but did not fully abolish shared predictive structure in the retained feature set. The reduced-feature predictive-overlap heatmaps are shown in Figure 8.

As a supplementary descriptive summary, redundancy among retained predictors was again summarized using mean within-sample variance, MPAD, and RMSPD. The overall profiles were fairly close across algorithms, although some differences remained. KNN showed the largest divergence values across the three measures, whereas linear SVM showed the smallest overall divergence pattern. These values are presented as descriptive summaries of similarity structure rather than as definitive measures of shared predictive information. The descriptive redundancy summaries for the reduced-feature experiment are shown in Table 6.

Table 6. Descriptive redundancy summary metrics for the retained predictors in Experiment 2. Metrics include mean within-sample variance, MPAD, and RMSPD. These values are interpreted as supplementary descriptive summaries of predictor similarity structure.

Model	Mean Within-Sample Variance	MPAD Distance	RMSPD Distance
XGBoost	0.790	0.945	1.325
Logistic Regression	0.809	0.979	1.341
Ridge Classifier	0.811	0.978	1.342
KNN	**0.835**	**0.990**	**1.362**
SVM (Linear)	0.724	0.965	0.998

Exploratory model-level regressions relating these descriptive redundancy summaries to classification metrics again showed weak associations overall and were therefore not treated as a central result. Taken together, Experiment 2 showed that removing overlapping predictors modestly reduced absolute performance but preserved the overall ranking of algorithms. XGBoost remained the strongest-performing model, logistic regression and ridge classifier remained closely aligned, and SVM and KNN continued to trail behind.

3.3 Experiment 3: robustness to incremental Gaussian noise

To assess performance robustness under controlled perturbation, Gaussian noise was added to the full predictor set at standard deviations of 0.05, 0.1, 0.2, 0.5, and 1.0. Performance was summarized using mean accuracy and standard deviation across seeds. Across the tested perturbation range, most models demonstrated minimal variation in performance, indicating that classification accuracy was generally robust to increasing levels of random numerical noise.

XGBoost consistently achieved the highest accuracy across all noise conditions, with mean accuracy remaining between approximately 0.771 and 0.773. Ridge classifier and logistic regression also showed strong stability, with both models maintaining mean accuracies near 0.761 across all standard deviations. Linear SVM exhibited little observable performance change across the tested noise levels and remained near 0.740. Together, these findings indicate that the ensemble-based and linear models were largely unaffected by progressive Gaussian perturbation. In contrast, KNN showed the greatest sensitivity to noise. Its mean accuracy declined modestly at the highest perturbation levels, making KNN the least robust classifier in this experiment. This pattern is consistent with the dependence of KNN on local distance relationships, which can be disrupted more readily by added perturbation than the decision boundaries used by linear or tree-based models.

Overall, the Gaussian-noise experiment did not reveal substantial degradation in performance for most classifiers. Instead, the results suggest that the principal models in this study, especially XGBoost, logistic regression, ridge classifier, and linear SVM, retained stable performance under progressively stronger synthetic perturbation. These results should be interpreted as a controlled numerical robustness test rather than as a full simulation of real-world clinical data corruption. Within that narrower interpretation, the findings indicate that moderate synthetic perturbation did not materially compromise classification performance for most algorithms, whereas KNN remained the most sensitive to noise. These robustness results are summarized in Figure 9.

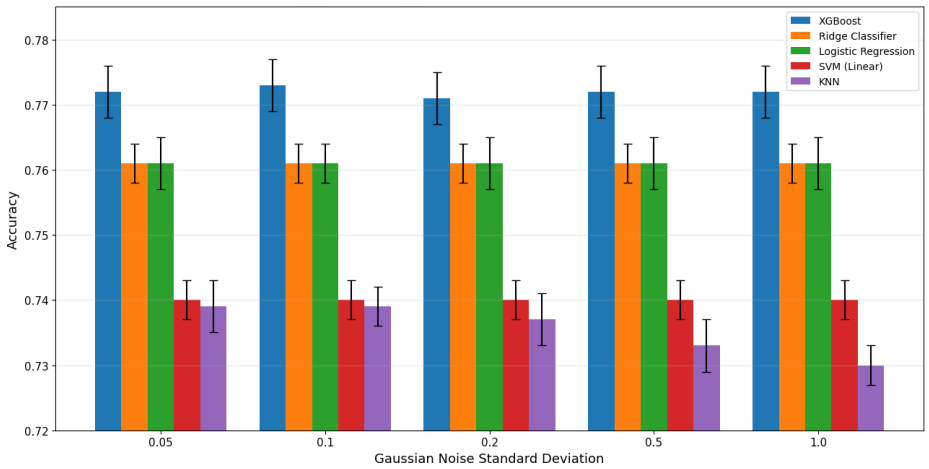

Figure 9. Mean classification accuracy of five machine learning models under increasing Gaussian noise perturbation (standard deviation = 0.05, 0.1, 0.2, 0.5, and 1.0). This figure summarizes repeated-split stability under controlled synthetic perturbation rather than balanced diagnostic performance. Error bars indicate standard deviation across seeds.

4. Discussion

4.1 Interpretation of main findings

This study evaluated five machine learning algorithms for cross-sectional AD diagnosis classification using the NACC dataset and examined predictive overlap, feature reduction, calibration, and robustness under controlled perturbation. Across experiments, XGBoost achieved the strongest overall performance, demonstrating the highest discrimination, calibration, and balanced accuracy, while logistic regression and ridge classifier remained closely competitive. In contrast, linear SVM and KNN consistently produced lower performance. This overall pattern is consistent with the broader AD machine learning literature, in which flexible ensemble or nonlinear methods often achieve the strongest discrimination, but simpler linear models frequently remain surprisingly competitive when the underlying dataset already contains strong clinical signals. Large recent reviews of AD machine learning note that performance differences across studies depend heavily on dataset composition, modality choice, and validation design rather than on algorithm name alone, which makes the close performance of XGBoost, logistic regression, and ridge classifier in the present study plausible rather than unexpected (4-6,25,26). The overall sensitivity-specificity pattern is also important for interpretation. Across both the baseline and reduced-feature experiments, all models showed consistently high sensitivity but comparatively low specificity, indicating a tendency to prioritize identification of AD-positive cases over correct exclusion of AD-negative cases. In a class-imbalanced dataset, this pattern is not unusual, but it is also clinically meaningful in the context of AD. Contemporary AD diagnostic frameworks emphasize that clinical diagnosis is not based on a single test output; instead, it integrates symptoms, cognitive performance, functional decline, biomarker evidence, and clinical judgment. Recent diagnostic updates have moved the field further toward biomarker-informed and biologically grounded definitions of AD, while still recognizing that clinical staging and patient-level decision-making require multimodal interpretation rather than reliance on one isolated signal (1,3,27,28). In that context, the current models are more appropriately interpreted as supportive classification tools with potential screening value rather than as definitive standalone diagnostic systems. High sensitivity may reduce missed cases, but the accompanying false-positive burden means that positive classifications would still require confirmatory clinical evaluation.

The finding that model performance was preserved after reduction to 13 variables further strengthens the interpretation that a useful classification signal is distributed across partially overlapping predictors rather than concentrated in a very small number of uniquely indispensable features. Although absolute performance declined modestly after feature reduction, the overall ranking of algorithms remained largely unchanged, with XGBoost retaining the strongest performance and logistic regression and ridge classifier remaining closely aligned. This fits well with prior AD machine learning work showing that feature selection can often reduce

dimensionality substantially while preserving most classification performance, especially when the retained variables represent clinically central domains such as age, APOE status, cognitive decline, vascular burden, and functional impairment. Recent AD studies and reviews frequently frame feature selection in terms of relevance and parsimony, including minimum-redundancy strategies, multimodal selection pipelines, or compact interpretable subsets, which supports the broader idea that not all available predictors contribute uniquely independent signal (5,6,25,29,30).

That interpretation is strengthened by the predictive-overlap analyses. Many AD machine learning papers report top predictors or apply feature selection, but fewer directly ask whether distinct predictors produce similar participant-level prediction patterns. In the present study, overlap was defined primarily in terms of similarity in single-feature OOF prediction behavior rather than similarity in raw feature values alone. Several top-ranked predictors produced similar participant-level outputs when used independently, suggesting that different variables may guide the models toward comparable classifications even when they differ in scale or clinical interpretation. This is an important conceptual distinction. Traditional redundancy measures often focus on correlation structure or minimum-redundancy feature filtering, whereas the present analysis focuses on overlap in predictive behavior. In a clinical AD dataset, that interpretation is biologically and clinically plausible. AD does not usually manifest through one isolated variable; instead, it appears as a syndrome spanning cognition, daily function, genetic susceptibility, neuropsychiatric change, sleep-related symptoms, and broader medical history.

Multiple predictors may therefore act as partially substitutable indicators of related underlying disease burden, even when they are not redundant in a purely statistical sense (1,3,22,27). The supplementary variance-, MPAD-, and RMSPD-based summaries were directionally consistent with the presence of similarity structure among predictors, but they were treated only as secondary descriptive diagnostics. That distinction remains important in light of the prior literature. Feature-selection studies in AD commonly seek low-redundancy subsets because high-dimensional clinical, imaging, or molecular datasets often contain overlapping features, but low pairwise redundancy does not necessarily imply distinct predictive roles, and conversely, predictors with different raw-value distributions may still drive similar model outputs. The present study therefore interprets distance- and dispersion-based summaries cautiously and places greater weight on the OOF predictive-overlap heatmaps and the preservation of performance after feature reduction. This framing helps distinguish descriptive similarity from functional substitutability and better matches the underlying clinical question: not whether variables are numerically interchangeable, but whether they carry overlapping classification signals in practice (5,25,29,30).

The calibration and discrimination results also deserve separate emphasis. Much of the AD machine learning literature still prioritizes accuracy, sensitivity, or ROC-based

discrimination, whereas calibration is discussed less often despite its importance for clinical interpretability. In the present study, XGBoost showed both the strongest discrimination and the lowest Brier score in the baseline experiment, and it remained the best-calibrated model overall after feature reduction. This is meaningful because a model can rank individuals well while still producing poorly calibrated probabilities. The shift in calibration behavior after feature reduction is therefore important: it indicates that a reduced model can preserve broad classification performance while changing how predicted probabilities align with observed outcome frequencies. Recent reviews of AI in AD diagnosis have similarly emphasized that translational usefulness depends not only on high discrimination but also on reliable, clinically interpretable probability behavior, particularly when models may be used to support triage, referral, or further biomarker workup (25,26).

The perturbation analysis adds a complementary robustness perspective. Most published robustness discussions in AD machine learning emphasize multimodal imaging pipelines, adversarial vulnerability, or instability introduced by data heterogeneity rather than controlled perturbation of tabular clinical features. In the present study, most models, including XGBoost, logistic regression, ridge classifier, and linear SVM, retained stable performance across increasing Gaussian noise levels, whereas KNN showed greater sensitivity. This pattern is consistent with more general robustness principles in machine learning and medical AI. Instance-based methods such as KNN are often more vulnerable to disruption of local geometry, while regularized linear models and ensemble tree-based models can be more stable under moderate noise. AD-specific robustness papers have also begun to emphasize outlier resistance, graph-regularized feature selection, and the need to maintain performance under imperfect data conditions, so the present perturbation findings fit a broader methodological concern even though they do not constitute a formal adversarial benchmark (29,30).

Another important result is that increased algorithmic complexity yielded only modest gains over simpler models in this setting. Although XGBoost performed best overall, logistic regression and ridge classifier remained relatively competitive across experiments. This is again consistent with the broader AD literature. Reviews repeatedly note that complex models, especially deep learning systems, may offer performance advantages in high-dimensional imaging settings, but these gains are often reduced in structured clinical datasets or when strong baseline predictors are already present. In those settings, simpler models may capture much of the usable signal while offering advantages in interpretability, implementation burden, and reproducibility (4-6, 25, 26). From a clinical perspective, this matters because AD-related decisions are typically embedded in workflows that require explanation, triangulation with other evidence, and tolerance for imperfect or incomplete data. A small improvement in peak performance may therefore not automatically justify a large increase in model complexity if simpler methods retain most of the practical signal.

The relationship between redundancy and the broader clinical picture of AD is especially important. AD is not only a biomarker-defined disease process but also a clinical syndrome with correlated manifestations across memory, orientation, executive function, function in daily life, and neuropsychiatric change. That makes partial overlap among predictors clinically believable. Age, APOE-related measures, behavioral history, sleep variables, psychiatric indicators, vascular variables, and functional proxies may differ in meaning, but they do not operate in isolation from one another. Instead, they can reflect different surfaces of the same underlying disease process or of related vulnerability pathways. The present results therefore fit a clinically coherent interpretation: predictive redundancy in AD classification is not merely a nuisance of collinearity, but may reflect the fact that AD expresses itself through multiple interrelated domains that can substitute for one another to some extent in cross-sectional classification models (1,3,22,27,28).

4.2 Limitations

Several limitations should still be acknowledged. First, this study examined cross-sectional diagnosis classification rather than longitudinal prediction of future AD onset, conversion, or trajectory. Second, evaluation was limited to internal validation within the NACC dataset, and external validation in independent cohorts remains necessary before drawing stronger conclusions about generalizability (23). Third, complete-case filtering excluded a small number of observations with missing data and may introduce mild selection bias if missingness was not random (24). Fourth, despite careful leakage control and removal of direct diagnostic proxies, some retained predictors may still function as indirect correlates of disease status in a cross-sectional diagnosis-classification setting. Fifth, the synthetic noise experiment represents a controlled perturbation analysis rather than a full simulation of real-world clinical measurement error, workflow inconsistency, or distribution shift. Finally, because feature-importance methods differ across model families, top-ranked predictors were interpreted qualitatively within models rather than as directly comparable common-scale effect sizes across algorithms.

Despite these limitations, the study provides a useful methodological perspective on predictive overlap, feature reduction, calibration, and robustness in clinical AD classification. The results suggest that predictive signal in this setting is distributed across partially overlapping variables, that substantial simplification of the feature space can preserve much of the usable classification information, and that most of the evaluated models remain stable under moderate synthetic perturbation. Taken together, these findings support the view that smaller and more interpretable predictor sets may retain substantial practical value in leakage-controlled cross-sectional AD diagnosis classification, while also highlighting the importance of evaluating model behavior in terms of discrimination, calibration, redundancy, and robustness rather than accuracy alone.

4.3 Perspectives

The present findings also motivate several

broader conceptual questions that extend beyond the specific experiments performed here.

4.3.1 *Distribution of classification signal across progressively broader layers of features*

A useful future direction emerging from the present findings is the possibility that Alzheimer's disease classification signal may be distributed across progressively broader layers of features rather than concentrated solely within the highest-ranked predictors. In the current study, performance declined only modestly after substantial feature reduction, suggesting that at least part of the predictive information may be shared across partially overlapping variables. Future work could investigate this idea more directly by systematically partitioning predictors into ranked feature bands (for example, features 1–10, 11–20, 21–30, and so forth) and independently evaluating classification performance within each band. Such analyses could help determine whether lower-ranked variables still retain meaningful disease-related signals despite contributing less strongly to overall feature-importance measures.

This framework may also help clarify how different machine learning architectures extract information from complex clinical datasets. For example, if ensemble methods such as XGBoost maintain relatively stable performance across progressively lower-ranked feature groups whereas linear models degrade rapidly, this could suggest that nonlinear ensemble models are better able to recover weak, distributed, or interaction-based disease structure embedded across many variables. Conversely, rapid degradation across all models would imply that classification depends primarily on a smaller set of dominant predictors. In this interpretation, machine learning models function not only as classifiers but also as indirect probes of how disease-related information is distributed throughout the dataset.

More broadly, these analyses may help distinguish between concentrated and distributed representations of disease. A concentrated representation would imply that a relatively small number of highly informative biomarkers carry most of the clinically relevant signal. In contrast, a distributed representation would suggest that Alzheimer's disease manifests across multiple partially overlapping domains, including demographic, vascular, behavioral, psychiatric, sleep-related, and genetic variables, each contributing modest but non-negligible information about underlying disease state. Such a framework may help contextualize why reduced-feature models can sometimes retain substantial predictive performance and why different machine learning architectures may vary in their ability to recover latent disease structure from weaker feature subsets.

4.3.2 *Predictive redundancy as an informative systems-level indicator*

An additional possibility is that predictive redundancy itself may eventually serve as an informative systems-level indicator rather than merely a statistical nuisance. Traditionally, redundancy in machine learning is often treated as something to be removed in order to improve model simplicity or reduce

multicollinearity. However, in complex biological disorders such as Alzheimer's disease, redundancy may also reflect the extent to which disease-related perturbations propagate across multiple interconnected physiological and clinical domains. In this interpretation, overlap among predictors could represent distributed manifestations of a shared latent pathological state rather than simple duplication of information.

Future studies could therefore explore whether redundancy-related measures vary systematically across disease stages, patient subgroups, or rates of progression. For example, early disease states might exhibit weaker or more localized overlap structures, whereas advanced disease could produce broader convergence among cognitive, behavioral, vascular, psychiatric, and metabolic variables as pathology increasingly affects multiple systems simultaneously. Under such a framework, increasing predictive overlap could potentially reflect growing systemic integration of disease burden across observable features.

Relatedly, redundancy-pattern analysis may help identify situations in which distinct clinical domains are converging toward similar participant-level prediction behavior despite remaining superficially different in raw measurements. For instance, sleep-related variables, vascular variables, psychiatric indicators, and genetic risk factors may each independently contribute to classification while still partially encoding the same latent disease trajectory. If validated in future work, such convergence patterns could provide insight into how heterogeneous manifestations of Alzheimer's disease become organized into shared syndrome-level representations within machine learning models.

This perspective also raises the possibility that different machine learning architectures may reveal different aspects of latent disease organization. Linear models may rely more heavily on globally dominant predictors, whereas nonlinear ensemble methods may be better able to aggregate weak distributed signals spread across many partially overlapping variables. Under this interpretation, model-dependent degradation across progressively lower-ranked feature subsets could itself become informative about the geometry and distribution of disease-related information in clinical datasets.

Importantly, these possibilities remain speculative within the context of the present study and were not directly tested here. The current analyses therefore should not be interpreted as definitive evidence of functional interchangeability among feature groups, biologically distributed disease encoding, or causal relationships among predictors. Rather, the findings motivate future investigations into whether predictive signal persistence, overlap structure, and redundancy dynamics may provide useful insight into the organization, robustness, and systems-level representation of Alzheimer's disease within machine learning frameworks.

5. Conclusion

This study compared five machine learning models for cross-sectional Alzheimer's disease diagnosis classification using the NACC

dataset and found that XGBoost performed best overall, while logistic regression and ridge classifier remained closely competitive. Performance declined only modestly after feature reduction, suggesting that substantial classification signal was retained in a smaller and more interpretable variable set. Predictive-overlap analyses further indicated that useful signal was distributed across partially substitutable features rather than concentrated in a few uniquely essential predictors. Most models also remained stable under moderate Gaussian perturbation, although KNN showed greater sensitivity to noise. Overall, these findings suggest that reduced-feature, leakage-controlled models can retain strong practical value for cross-sectional AD diagnosis classification and that evaluation should consider robustness, calibration, and predictive overlap alongside discrimination.

More broadly, the persistence of performance across reduced and partially overlapping feature sets raises the possibility that Alzheimer's disease-related information may be distributed across multiple interconnected clinical domains rather than localized within only a few dominant biomarkers. In this interpretation, machine learning models may also serve as indirect probes of broader disease related structure, potentially helping future studies investigate how predictive overlap and redundancy evolve across disease stages and heterogeneous clinical presentations.However, these interpretations remain speculative and were not directly tested in the present study.

Acknowledgements

The ADSP Phenotype Harmonization Consortium (ADSP-PHC) is funded by NIA (U24 AG074855, U01 AG068057, and R01 AG059716). Additional acknowledgements include the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA.

The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI David Holtzman, MD), P30 AG066518 (PI Lisa Silbert, MD, MCR), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI Julie A. Schneider, MD, MS), P30 AG072978 (PI Ann McKee, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Jessica Langbaum, PhD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Glenn Smith, PhD, ABPP), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson,

MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P30 AG086401 (PI Erik Roberson, MD, PhD), P30 AG086404 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), and P30 AG072959 (PI James Leverenz, MD).

6. References

1. Liu, E., Zhang, Y., & Wang, J.-Z. (2024). Updates in Alzheimer's disease: From basic research to diagnosis and therapies. *Translational Neurodegeneration, 13*(1), 45. https://doi.org/10.1186/s40035-024-00432-x

2. Alzheimer's Association. (2020). 2020 Alzheimer's disease facts and figures. *Alzheimer's & Dementia, 16*(3), 391–460. https://doi.org/10.1002/alz.12068

3. Cody, K. A., Langhough, R. E., Zammit, M. D., Clark, L., Chin, N., Christian, B. T., et al. (2024). Characterizing brain tau and cognitive decline along the amyloid timeline in Alzheimer's disease. Brain, 147(6), 2144–2157. https://doi.org/10.1093/brain/awae116

4. Bazarbekov, I., Razaque, A., Ipalakova, M., Yoo, J., Assipova, Z., & Almisreb, A. (2024). A review of artificial intelligence methods for Alzheimer's disease diagnosis: Insights from neuroimaging to sensor data analysis. *Biomedical Signal Processing and Control, 89*, 106023. https://doi.org/10.1016/j.bspc.2024.106023

5. Rezaie, Z., Banad, Y. Machine learning applications in Alzheimer's disease research: a comprehensive analysis of data sources, methodologies, and insights. *Int J Data Sci Anal,* 20, 3169–3203 (2025). https://doi.org/10.1007/s41060-024-00651-5

6. Kumar, S., Oh, I., Schindler, S., Lai, A. M., Payne, P. R. O., & Gupta, A. (2021). Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. *JAMIA open*, *4*(3), ooab052. https://doi.org/10.1093/jamiaopen/ooab052

7. Morris, J. C., Weintraub, S., Chui, H. C., Cummings, J., DeCarli, C., Ferris, S., et al. (2006). The Uniform Data Set (UDS): Clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Disease and Associated Disorders, 20*(4), 210–216. https://doi.org/10.1097/01.wad.0000213865.09806.92

8. National Alzheimer's Coordinating Center. (n.d.). *Publications and descriptions of NACC*

data. Retrieved March 20, 2026, from NACC website

9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

10. Tolles, J., & Meurer, W. J. (2016). Logistic regression: Relating patient characteristics to outcomes. *JAMA, 316*(5), 533–534. https://doi.org/10.1001/jama.2016.7653

11. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297. https://doi.org/10.1007/BF00994018

12. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

13. Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

14. Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

15. Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

16. O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*, 673–690. https://doi.org/10.1007/s11135-006-9018-6

17. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE, 10*(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

18. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

19. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

20. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845. https://doi.org/10.2307/2531595

21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

22. Belloy, M. E., Napolioni, V., & Greicius, M. D. (2019). A quarter century of APOE and Alzheimer's disease: Progress to date and the path forward. *Neuron, 101*(5), 820–838. https://doi.org/10.1016/j.neuron.2019.01.056

23. Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Jr., Moons, K. G. M., & Collins, G. S. (2024). Evaluation of clinical prediction models (part 2): How to undertake an external validation study. *BMJ, 384*, Article e074820. https://doi.org/10.1136/bmj-2023-074820

24. Ross, R. K., Breskin, A., & Westreich, D. (2020). When is a complete-case approach to missing data valid? The importance of effect-measure modification. *American Journal of Epidemiology, 189*(12), 1583–1589. https://doi.org/10.1093/aje/kwaa103

25. Christodoulou, R. C., Woodward, A., Pitsillos, R., Ibrahim, R., & Georgiou, M. F. (2025). Artificial intelligence in Alzheimer's disease diagnosis and prognosis using PET-MRI: A narrative review of high-impact literature post-Tauvid approval. *Journal of Clinical Medicine, 14*(16), 5913. https://doi.org/10.3390/jcm14165913

26. Jack, C. R., Jr., Andrews, S. J., Beach, T. G., Buracchio, T., Dunn, B., Graf, A., et al. (2024). Revised criteria for the diagnosis and staging of Alzheimer's disease. *Nature Medicine, 30*, 2121–2124. https://doi.org/10.1038/s41591-024-02988-7

27. Alzheimer's Association. (n.d.). *Criteria for diagnosis and staging of Alzheimer's disease.* Retrieved April 1, 2026, from Alzheimer's Association website

28. Alshamlan, H., Alwassel, A., Banafa, A., & Alsaleem, L. (2024). Improving Alzheimer's disease prediction with different machine learning approaches and feature selection techniques. *Diagnostics, 14*(19), 2237. https://doi.org/10.3390/diagnostics14192237

29. Zhang, C., Fan, W., Li, H., & Chen, C. (2024). Multi-level graph regularized robust multi-modal feature selection for Alzheimer's disease classification. *Knowledge-Based Systems, 293*,

111676. https://doi.org/10.1016/j.knosys.2024.111676

30. Li, Y., Chen, G., Wang, G., Zhou, Z., An, S., Dai, S., Jin, Y., Zhang, C., Zhang, M., & Yu, F. (2024). Dominating Alzheimer's disease diagnosis with deep learning on sMRI and DTI-MD. *Frontiers in Neurology, 15*, 1444795. https://doi.org/10.3389/fneur.2024.1444795

Appendix

The following tables list the top 10 features for each model. Descriptions are repeated in each model table so the table can be read independently.

Experiment 1: Baseline Dataset

XGBoost

Rank	Feature	Meaning	Explanation
1	QUITSMOK	Quit smoking	Variable related to smoking cessation or prior smoking status.
2	TOBAC30	Tobacco use in past 30 days	Indicator of recent tobacco use within approximately the last 30 days.
3	NACCNE4S	Number of APOE ε4 alleles	Number of APOE epsilon4 alleles carried by the participant.
4	NACCAPOE	APOE genotype variable	NACC-coded APOE genotype variable.
5	NACCAGE	Participant age	Age of the participant at the study visit.
6	NACCGDS	Geriatric Depression Scale score	NACC Geriatric Depression Scale score or related depression-screening summary.
7	NACCFAM	Family history variable	NACC family history variable indicating whether dementia or Alzheimer-related family history was reported.
8	SEX	Biological sex	Participant biological sex as recorded in the dataset.
9	CBSTROKE	Stroke history	Clinical history indicator for stroke or cerebrovascular accident.
10	WEIGHT	Weight	Participant body weight.

Logistic Regression

Rank	Feature	Meaning	Explanation
1	ALCABUSE	Alcohol abuse history	Indicator that the participant had a history of clinically relevant alcohol abuse or dependence.
2	NACCAGE	Participant age	Age of the participant at the study visit.
3	NACCNE4S	Number of APOE ε4 alleles	Number of APOE epsilon4 alleles carried by the participant.
4	ALCOCCAS	Occasional alcohol use	Indicator of occasional alcohol consumption.
5	HEIGHT	Height	Participant height.
6	BPDIAS	Diastolic blood pressure	The lower number in a blood pressure reading, reflecting arterial pressure between heartbeats.
7	BPSYS	Systolic blood pressure	The upper number in a blood pressure reading, reflecting arterial pressure when the heart contracts.
8	TBIYEAR	Year/timing of traumatic brain injury	Variable describing the year or timing of traumatic brain injury
9	NACCAPOE	APOE genotype variable	NACC-coded APOE genotype variable.
10	HYPERTEN	Hypertension	Indicator of high blood pressure diagnosis or history.

Ridge Classifier

Rank	Feature	Meaning	Explanation
1	ALCABUSE	Alcohol abuse history	Indicator that the participant had a history of clinically relevant alcohol abuse or dependence.
2	NACCAGE	Participant age	Age of the participant at the study visit.
3	NACCNE4S	Number of APOE ε4 alleles	Number of APOE epsilon4 alleles carried by the participant.
4	HEIGHT	Height	Participant height.
5	ALCOCCAS	Occasional alcohol use	Indicator of occasional alcohol consumption.
6	BPDIAS	Diastolic blood pressure	The lower number in a blood pressure reading, reflecting arterial pressure between heartbeats.
7	BPSYS	Systolic blood pressure	The upper number in a blood pressure reading, reflecting arterial pressure when the heart contracts.
8	TBIYEAR	Year/timing of traumatic brain injury	Variable describing the year or timing of traumatic brain injury.
9	NACCAPOE	APOE genotype variable	NACC-coded APOE genotype variable.
10	HYPERTEN	Hypertension	Indicator of high blood pressure diagnosis or history.

KNN

Rank	Feature	Meaning	Explanation
1	NACCAGE	Participant age	Age of the participant at the study visit.
2	HEIGHT	Height	Participant height.
3	NACCGDS	Geriatric Depression Scale score	NACC Geriatric Depression Scale score or related depression-screening summary.
4	SEX	Biological sex	Participant biological sex as recorded in the dataset.
5	WEIGHT	Weight	Participant body weight.
6	ALCABUSE	Alcohol abuse history	Indicator that the participant had a history of clinically relevant alcohol abuse or dependence.
7	DEP	Depression indicator	Indicator of depression history or depressive symptoms, depending on dataset coding.
8	ALCFREQ	Alcohol consumption frequency	Variable describing how often the participant consumed alcohol.
9	NACCMOM	Maternal dementia/family history variable	NACC variable related to maternal history of dementia or Alzheimer disease.
10	NACCFAM	Family history variable	NACC family history variable indicating whether dementia or Alzheimer-related family history was reported.

SVM (Linear)

Rank	Feature	Meaning	Explanation
1	ALCABUSE	Alcohol abuse history	Indicator that the participant had a history of clinically relevant alcohol abuse or dependence.
2	ALCOCCAS	Occasional alcohol use	Indicator of occasional alcohol consumption.
3	TBIEXTEN	Extended loss of consciousness from TBI	Indicator of traumatic brain injury with extended loss of consciousness.
4	TBI	Traumatic brain injury history	Indicator of traumatic brain injury history.
5	TBIYEAR	Year/timing of traumatic brain injury	Variable describing the year or timing of traumatic brain injury.
6	TBIBRIEF	Brief loss of consciousness from TBI	Indicator of traumatic brain injury with brief loss of consciousness.
7	ALCFREQ	Alcohol consumption frequency	Variable describing how often the participant consumed alcohol.
8	HEIGHT	Height	Participant height.
9	NACCAGE	Participant age	Age of the participant at the study visit.
10	INSOMN	Insomnia	Indicator of insomnia or sleep disturbance.

Experiment 2: Reduced Dataset

XGBoost

Rank	Feature	Meaning	Explanation
1	DEP2YRS	Depression within past 2 years	Indicator that depression was present or reported within the previous two years.
2	NACCNE4S	Number of APOE ε4 alleles	Number of APOE epsilon4 alleles carried by the participant.
3	TOBAC100	Lifetime tobacco use history	Indicator of lifetime tobacco exposure, often based on the threshold of having smoked at least 100 cigarettes.
4	NACCAPOE	APOE genotype variable	NACC-coded APOE genotype variable.
5	NACCAGE	Participant age	Age of the participant at the study visit.
6	CBSTROKE	Stroke history	Clinical history indicator for stroke or cerebrovascular accident.
7	NACCMOM	Maternal dementia/family history variable	NACC variable related to maternal history of dementia or Alzheimer disease.
8	BPDIAS	Diastolic blood pressure	The lower number in a blood pressure reading, reflecting arterial pressure between heartbeats.
9	INSOMN	Insomnia	Indicator of insomnia or sleep disturbance.
10	NACCBMI	Body mass index	NACC-derived body mass index. BMI summarizes body size using height and weight and is often treated as a general metabolic/health variable.

Logistic Regression

Rank	Feature	Meaning	Explanation
1	ALCABUSE	Alcohol abuse history	Indicator that the participant had a history of clinically relevant alcohol abuse or dependence.
2	INSOMN	Insomnia	Indicator of insomnia or sleep disturbance.
3	NACCAGE	Participant age	Age of the participant at the study visit.
4	NACCNE4S	Number of APOE ε4 alleles	Number of APOE epsilon4 alleles carried by the participant.
5	NACCAPOE	APOE genotype variable	NACC-coded APOE genotype variable.
6	TBI	Traumatic brain injury history	Indicator of traumatic brain injury history.
7	NACCBMI	Body mass index	NACC-derived body mass index. BMI summarizes body size using height and weight and is often treated as a general metabolic/health variable.
8	TOBAC100	Lifetime tobacco use history	Indicator of lifetime tobacco exposure, often based on the threshold of having smoked at least 100 cigarettes.
9	CBSTROKE	Stroke history	Clinical history indicator for stroke or cerebrovascular accident.
10	NACCMOM	Maternal dementia/family history variable	NACC variable related to maternal history of dementia or Alzheimer disease. It captures family history through the participant's mother.

Ridge Classifier

Rank	Feature	Meaning	Explanation
1	ALCABUSE	Alcohol abuse history	Indicator that the participant had a history of clinically relevant alcohol abuse or dependence.
2	INSOMN	Insomnia	Indicator of insomnia or sleep disturbance.
3	NACCAGE	Participant age	Age of the participant at the study visit.
4	NACCNE4S	Number of APOE ε4 alleles	Number of APOE epsilon4 alleles carried by the participant.
5	TBI	Traumatic brain injury history	Indicator of traumatic brain injury history.
6	NACCAPOE	APOE genotype variable	NACC-coded APOE genotype variable.
7	NACCBMI	Body mass index	NACC-derived body mass index. BMI summarizes body size using height and weight and is often treated as a general metabolic/health variable.
8	TOBAC100	Lifetime tobacco use history	Indicator of lifetime tobacco exposure, often based on the threshold of having smoked at least 100 cigarettes.
9	NACCMOM	Maternal dementia/family history variable	NACC variable related to maternal history of dementia or Alzheimer disease.
10	DEP2YRS	Depression within past 2 years	Indicator that depression was present or reported within the previous two years. This is more time-specific than a general depression history variable.

KNN

Rank	Feature	Meaning	Explanation
1	NACCAGE	Participant age	Age of the participant at the study visit.
2	ALCABUSE	Alcohol abuse history	Indicator that the participant had a history of clinically relevant alcohol abuse or dependence.
3	BPDIAS	Diastolic blood pressure	The lower number in a blood pressure reading, reflecting arterial pressure between heartbeats.
4	NACCAPOE	APOE genotype variable	NACC-coded APOE genotype variable.
5	NACCMOM	Maternal dementia/family history variable	NACC variable related to maternal history of dementia or Alzheimer disease. It captures family history through the participant's mother.
6	NACCNE4S	Number of APOE ε4 alleles	Number of APOE epsilon4 alleles carried by the participant.
7	NACCBMI	Body mass index	NACC-derived body mass index. BMI summarizes body size using height and weight and is often treated as a general metabolic/health variable.
8	TOBAC100	Lifetime tobacco use history	Indicator of lifetime tobacco exposure, often based on the threshold of having smoked at least 100 cigarettes.
9	DEP2YRS	Depression within past 2 years	Indicator that depression was present or reported within the previous two years. This is more time-specific than a general depression history variable.
10	CBSTROKE	Stroke history	Clinical history indicator for stroke or cerebrovascular accident.

SVM (Linear)

Rank	Feature	Meaning	Explanation
1	ALCABUSE	Alcohol abuse history	Indicator that the participant had a history of clinically relevant alcohol abuse or dependence.
2	INSOMN	Insomnia	Indicator of insomnia or sleep disturbance.
3	TBI	Traumatic brain injury history	Indicator of traumatic brain injury history.
4	HYPERTEN	Hypertension	Indicator of high blood pressure diagnosis or history.
5	CBSTROKE	Stroke history	Clinical history indicator for stroke or cerebrovascular accident.
6	TOBAC100	Lifetime tobacco use history	Indicator of lifetime tobacco exposure, often based on the threshold of having smoked at least 100 cigarettes.
7	DEP2YRS	Depression within past 2 years	Indicator that depression was present or reported within the previous two years. This is more time-specific than a general depression history variable.
8	NACCAGE	Participant age	Age of the participant at the study visit.
9	NACCNE4S	Number of APOE epsilon4 alleles	Number of APOE epsilon4 alleles carried by the participant.
10	NACCAPOE	APOE genotype variable	NACC-coded APOE genotype variable.

Interpretation note. Because feature importance was estimated using different algorithms, the magnitude of importance values should not be compared directly across models. These tables are intended to define abbreviations and summarize which variables repeatedly appeared among each model's highest-ranked predictors.