WIMOAD: Weighted Integration of Multi-Omics data for Alzheimer's Disease (AD) Diagnosis

Hanyu Xiao[1], Jieqiong Wang[2]*, Shibiao Wan[1]*

[1] Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, United States, 68198

[2] Department of Neurological Sciences, University of Nebraska Medical Center, Omaha, NE, United States, 68198

*Corresponding contact:

Dr. Shibiao Wan, swan@unmc.edu

Dr. Jieqiong Wang, jiwang@unmc.edu

Abstract

As the most common subtype of dementia, Alzheimer's disease (AD) is characterized by a progressive decline in cognitive functions, especially in memory, thinking, and reasoning ability. Early diagnosis and interventions enable the implementation of measures to reduce or slow further regression of the disease, preventing individuals from severe brain function decline. The current framework of AD diagnosis depends on A/T/(N) biomarkers detection from cerebrospinal fluid or brain imaging data, which is invasive and expensive during the data acquisition process. Moreover, the pathophysiological changes of AD accumulate in amino acids, metabolism, neuroinflammation, etc., resulting in heterogeneity in newly registered patients. Recently, next generation sequencing (NGS) technologies have found to be a non-invasive, efficient and less-costly alternative on AD screening. However, most of existing studies rely on single omics only. To address these concerns, we introduce WIMOAD, a weighted integration of multi-omics data for AD diagnosis. WIMOAD synergistically leverages specialized classifiers for patients' paired gene expression and methylation data for multi-stage classification. The resulting scores were then stacked with MLP-based meta-models for performance improvement. The prediction results of two distinct meta-models were integrated with optimized weights for the final decision-making of the model, providing higher performance than using single omics only. Remarkably, WIMOAD achieves significantly higher performance than using single omics alone in the classification tasks. The model's overall performance also outperformed most existing approaches, highlighting its ability to effectively discern intricate patterns in multi-omics data and their correlations with clinical diagnosis results. In addition, WIMOAD also stands out as a biologically interpretable model by leveraging the SHapley Additive exPlanations (SHAP) to elucidate the contributions of each gene from each omics to the model output. We believe WIMOAD is a very promising tool for accurate

AD diagnosis and effective biomarker discovery across different progression stages, which eventually will have consequential impacts on early treatment intervention and personalized therapy design on AD.

Keywords: Alzheimer's Disease, Multi-omics, Weighted Score Fusion, Early Diagnosis, DNA Methylation

Introduction

Alzheimer's disease (AD) is the most common subtype of dementia, characterized by a progressive decline in cognitive functions, notably in memory, thinking, and reasoning [1]. It is closely associated with aging and exerts a persistent impact on cognitive functions [2]. With a national care cost growth of $24 billion from a year ago, reaching $345 billion overall in 2023, this neurodegenerative disease poses significant challenges for individuals and their families [3]. But according to previous study [4], AD is not an inevitable process of aging and there is the possibility to prevent or delay the development of this demensia in certain proportion of people. For primary healthcare and disease screening, the ability to achieve early and efficient diagnosis of AD is crucial for effective intervention and treatment [5].

Typically, AD is characterized by the A/T/N framework [6]. The "A" component refers to amyloidosis-beta peptide accumulation [7–9], and the "T" aspect, tauopathy, represents hyperphosphorylated tau protein aggregation [10,11]. The "N" component, focusing on specific aspects of neurodegeneration [12], gives an overall picture of neuronal and synaptic loss in the patients' brains. So far, the majority of research relies on phenotypic data, particularly brain imaging like Magnetic Resonance Imaging (MRI), Computed Tomography (CT) and Positron Emission Tomography (PET) [13,14]. With the advancements in artificial intelligence (AI)

algorithms [15], Chen et al. [16] have implanted U-Net, Multi-Layer Perceptron, and Graph Neural Network for 3-class AD diagnosis, and Al-Otaibi et al. [17] demonstrate the deep transfer learning on brain imaging with AutoEncoder structure, providing high classification performance. To aggregate different information extracted from multiple types of images, MMTFN introduced by Miao et al. [18] constructs a 3D multi-scale residual block layers and a Transformer network that jointly learns the representations from MRI and PET images of 720 subjects and gets a 94.61% accuracy between AD and Normal Control. Although the models are promising, utilizing the imaging data as model inputs results in However, idealized brain imaging of patients remains limited, and the neuropathological diagnosis is invasive and harmful to patients [19]. As pathophysiological changes gradually accumulate in amino acids, metabolism, and neuroinflammation, newly registered patients show considerable heterogeneity in the impaired cognitive domains which will lead to increasing diagnostic costs [20,21], underscoring the need for more precise and individualized diagnostic approaches [22–24].

With the progress in sequencing techniques, genetic data is increasingly being utilized as external validation in AD studies as the less-expensive and less-invasive measurement [25]. For example, researchers have identified many genetic risk factors for AD (e.g., *APOE* [26], *CR1* [27], *ABCA7* [28], etc.) identified by Single Nucleotide Polymorphism (SNP) in Genome-Wide Association Studies (GWAS) [29,30]. Transcriptomic analysis is also essential for biomarker detection in complex diseases like AD. Guo and Yang [31] applied a transcriptome-wide association study (TWAS) with reference transcriptomic data from brain and blood tissues and detected 141 risk genes while Methys et al. [32] utilized advanced single-cell transcriptome analysis and found cell-type specific disease-associated changes across various degrees of AD, which can provide a molecular and cellular foundation for further investigation. As one of the main

components of the epigenetic data and highly correlated with aging [33], DNA methylation level is found to be increased in peripheral cells of AD patients while correlating with worse cognitive performances and *APOE* polymorphism [34,35]. However, considering the intricate nature of the aging process and the progression of neurodegenerative disorders, relying on one data modality only may underestimate other related risk factors in this complicated process, since one omics can not convey all the information needed.

To enhance the effectiveness of current AD research, integrating genetic data could greatly improve the accuracy, reliability, and interpretability of the computational model [36–38]. However, how to combine data from different omics layers to provide a holistic view of biological systems remains the major challenge of this field. One general solution is to summarize all results from transcriptomic, proteomics, metabolomics, etc., on brains and other tissues and form a comprehensive understanding of the impact of one gene alterations in individual clinical trajectories [39–42]. Factor analysis, which represents high-dimensional variables to a smaller number of latent factors, is also brought up in multi-omics research (MOFA, multi-omics factor analysis) [43]. iCluster [44], JIVE [45] , and SLIDE [46] are all commonly used tools that jointly model associations and the variance-covariance structure within each data type while reducing the dimensionality for clustering. In AD studies, Bao et al. [47] proposed a structural Bayesian factor analysis framework named SBFA that incorporates imaging and biological data for functional assessment questionnaire (FAQ) score prediction. In addition, various integration or 'fusion' methodologies have been introduced through data concatenation with AI-based algorithms [48–50], but models that focus on AD studies are rare [51]. Clinical information is also incorporated in the integration process for better diagnosis performance [52].

Despite these advancements, significant gaps remain in integration studies. Firstly, most genomic studies focus on SNPs or gene expression data, with less attention on methylation data, which is highly related to aging and AD [53–55]. Secondly, widely used direct data concatenation [56] for integration may lose some key information for each data modality, as each omics will have different representations and data formats. To fill this gap, we proposed WIMOAD, which assigns distinct weights for the prediction score of each omics classifier and integrate the results from different data modalities to do the final decision-making, for different stages diagnosis of AD. Our major contribution can be summarized as follows:

(1) We proposed a stacked weighted score-based multiomics (gene expression and methylation data from ADNI) fusion model for Alzheimer's disease diagnosis, which has surpassed the performance of using single omics alone, as well as the existing integration methods.
(2) The stacking part of the ensemble model has dramatically improved the overall classification outcome on both single omics and the integration of two omics
(3) The proposed model is accurate, easy to use, time-saving, and interpretable from a biological view as we apply the Shapley Value [57] to quantify the contribution of individual genes for model decision-making, which will help for new biomarker detection.

Materials and Methods

Datasets

The data used in this paper are from the genetic section of the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI is a longitudinal multicenter study that collected clinical, imaging, genetic, and biochemical biomarkers for early detection and tracking of recruited cohorts across different time points. For our model, we collected the data of

591 people's gene expression and methylation profiles as model input following the criteria that the genetic profiles from different omics are paired for a certain sample (Originally we have 744 gene expression profiles and 649 methylation data records. The rest of the samples which only has one omics data were eliminated). Among them, there are 203 Normal Controls (CN) subjects (age: 74.45 ± 5.78, F/M: 101/102), 180 Early Mild Cognitive Impairments (EMCI) subjects (age: 71.44 ± 7.11, F/M: 81/ 99), 113 Late Mild Cognitive Impairments (LMCI) subjects (age: 72.74 ± 7.67; F/M: 45/68), which is 293 Mild Cognitive Impairments (MCI) and 95 Alzheimer's Diseases (AD) (age: 74.28 ± 7.59, F/M: 35/60). The demographic information of the data is shown in **Table.1**. For subsequent binary group classification tasks, we have reprocessed the original categories as follows: all samples, excluding the AD group, were categorized into a "patient" (PT) group to facilitate 'PT-AD' binary classification. Furthermore, the EMCI and LMCI groups were combined into a single MCI group, enabling the execution of other binary classification tasks related to MCI.

Table. 1 The demographic information of the Selected Participants. *Data are mean ± standard deviation (std). CN: Normal Controls; EMCI: Early Mild Cognitive Impairments; LMCI: Late Mild Cognitive Impairments; MCI: Mild Cognitive Impairments; AD: Alzheimer's Diseases; F: Female; M: Male*

Diagnosis	Samples	Age (mean±std)	Sex (F/M)
CN	N = 203	74.45 ± 5.78	101/102
EMCI	N = 180	71.44 ± 7.11	81/ 99
LMCI	N = 113	72.74 ± 7.67	45/68
AD	N = 95	74.28 ± 7.59	35/60

Overview of WIMOAD Framework

WIMOAD is a weighted score fusion model based on combining multiple base classifiers [58]. The pipeline is shown in **Fig. 1.** After establishing the database, gene expression and methylation data were extracted and paired according to patient ID to serve as model inputs. The model processed these omics separately, extracting the most variable genes from both omics within two categorized groups to use as features. For each data type, five commonly used machine learning classifiers, Support Vector Machine (SVM) [59] , Random Forest (RF) [60], Naïve Bayes (NB) [61], Logistic Regression (LR) [62] and K-Nearest Neighbors (KNN) [63], were applied independently to create new training sets with the prediction scores for meta-models, feedforward Multi-Layer Perceptron (MLP) [64,65]. Finally, the meta-model prediction results from both gene expression and methylation were combined using a weighted fusion mechanism [56]. The ensembled result was used to make the final decision on AD diagnosis. Subsequent optimization was performed for each classifier and the ensemble weight to enhance the integration model performance. The model was validated under 10 times 10-fold cross-validation (CV). In each CV round, the predicted score of each model was linearly combined by assigned weights for the final decision of the whole model. Once trained, the models were interpreted using SHAP to explain the results.

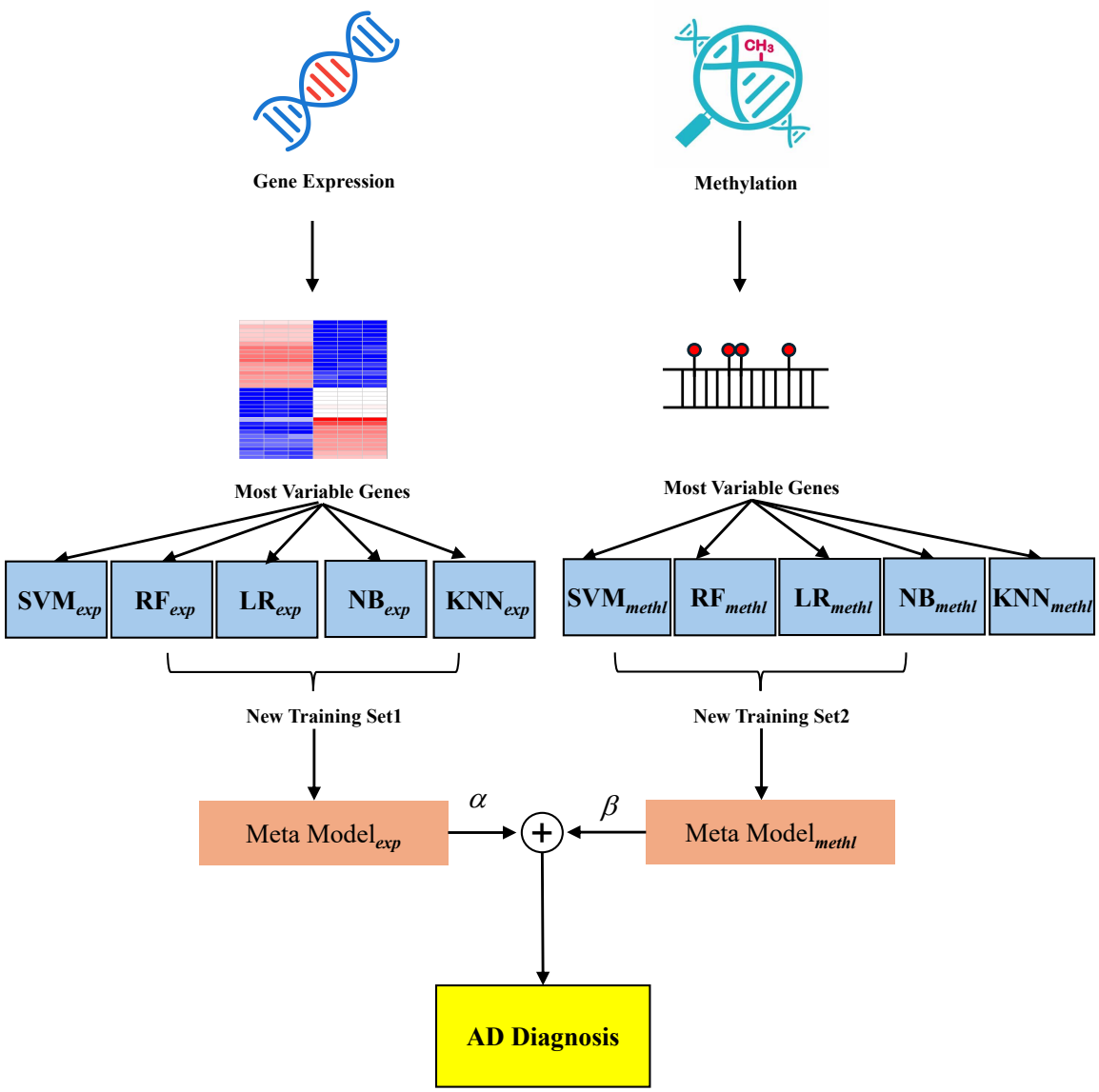


Fig. 1. The Workflow of WIMOAD. *The process begins by identifying the most variable features from paired gene expression and methylation data for classification. For each omics data, different classifiers were trained. The outputs of the basic classifiers were considered as the new training sets for two distinct meta-models, which used the predictions of base classifiers as inputs and generated the overall prediction scores. For multi-omics integration, each meta-model is assigned a weight for ensemble learning, which also controls the contributions of each meta-model to the final decision.* SVM_{exp}*: Support Vector Machine for gene expression data.* SVM_{methl}*: Support*

Vector Machine for gene methylation data. RF: Random Forest classifier. NB: Naïve Bayes classifier. KNN: K-Nearest Neighbor classifier. LR: Logistic Regression.

Preprocessing of multi-omics data

The gene expression profiling was provided with Affymetrix Human Genome U219 Array from peripheral blood samples. The raw expression values generated by this platform were first normalized using the Robust Multi-chip Average (RMA) method, resulting in 530,467 probes corresponding to 49,293 transcripts from 744 samples. These probes were subsequently mapped and annotated according to the human genome reference (hg19). Given that a single gene may be associated with multiple probes, we selected the probe data corresponding to the first occurrence of each gene in the processed matrix to represent the expression level of that gene for each individual. Genes with missing information in the annotated data were excluded from further analysis. Finally, the filtered data contains 20,270 annotated genes, and the expression matrix underwent a log transformation for scaling, which aimed to improve the accuracy of classification results.

Whole-genome DNA methylation profiling was conducted using the Illumina Infinium HumanMethylationEPIC BeadChip Array. The original data samples were normalized with the dasen method for downstream quality control (QC) including p-value criteria filtering, sex and sample ID verification, with 649 samples remained. The database provided raw data for these 649 participants who had undergone the QC process for further analysis. We obtained beta values for a total of 865,860 CpG sites by analyzing the channel signals. These CpG sites were subsequently mapped to the human genome reference (hg19), resulting in methylation data for 20,594 genes.

The workflow of the multi-omics data preparation is summarized in **Fig. 2**.

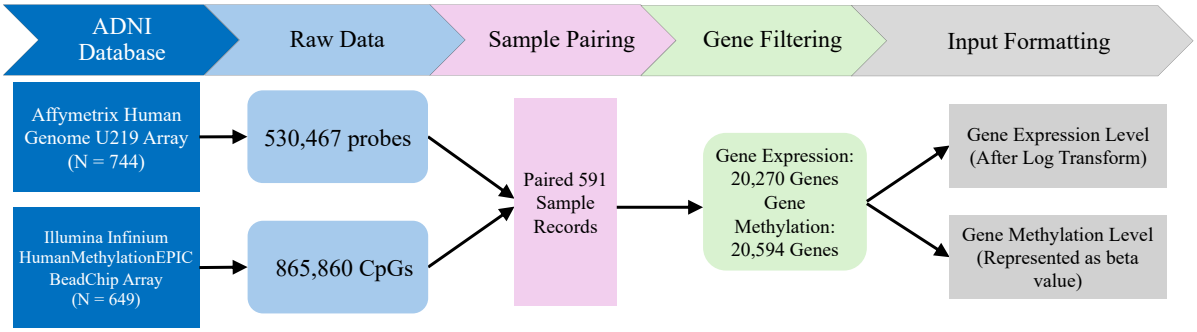


Fig. 2. The Preprocessing Steps for Multi-Omics Data.

Feature Selection

For a supervised learning model, in the case of a high-dimensionality curse and to enhance prediction efficiency while simultaneously reducing the consumption of computational resources, feature selection is a key process for model prediction. We selected 1000 genes that show statistically significant within-group variance separately for different omics inputs based on the ANOVA F-value [66] calculated by the 'SelectKBest' package in scikit-learn with 'f_classif' function. For comparison, we also employed median absolute deviation (MAD) and Fano factor for gene selection [67].

Weighted Score Fusion

In omics integration research, a common approach is to concatenate different types of data directly before classification. However, in this study, Exp and Methl data exhibit substantial differences in their representations and feature characteristics, which will result in suboptimal classification outcomes when directly concatenated or combined pairwise. Consequently, we employed a score fusion method to construct an integration model for multi-omics data. Initially, we assigned trained meta-model to each dataset separately for binary classification. Subsequently,

209 we performed a weighted linear aggregation of the obtained prediction scores to derive the final

210 prediction score of the model, which calculated as:

$$s_{\text{Int}} = \alpha * s_{Exp} + \beta * s_{Methl} \quad \text{s.t. } \alpha + \beta = 1, \alpha \geq 0, \beta \geq 0 \tag{1}$$

211

212 Where s_{Int} is the integrated prediction score of two meta-models, which represents the

213 probability of a given sample belonging to a specific class. s_{Exp} as the score generated by the

214 gene expression meta-model and s_{Methl} as the score generated by the gene methylation meta-

215 model. The α and β are the weight coefficients to balance the scores. These coefficients are

216 determined by the validation data in the 10 times 10-fold CV through screening from α = 0 to α

217 =1 in the linear combination. This approach ensures a more accurate and interpretable integration

218 of the diverse omics data types, accommodating the unique features of each dataset and enhancing

219 the overall classification performance.

220

221 **Evaluation of the Model Performance**

222 10 times 10-fold CV [68] was used to evaluate our WIMOAD. Specifically, we measured

223 accuracy (Acc), precision (Prec), Recall (Rec), F1-Score (F1), Matthews correlation coefficient

224 (MCC), Specificity (Sp), G-measure (G), Jaccard Index (Jacc) and Area Under Curve (AUC):

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Prec = \frac{TP}{TP + FP} \tag{2}$$

$$Rec = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

$$Sp = \frac{TN}{TN + FP} \quad (6)$$

$$G = \sqrt{\left(\frac{TP}{TP + FP}\right) \times \left(\frac{TP}{TP + FN}\right)} \quad (7)$$

$$Jacc = \frac{TP}{TP + FP + FN} \quad (8)$$

Model Interpretation with SHAP

To develop an explainable model, we utilized the Kernel SHAP Explainer [57,69] for multi-kernel classifiers for different omics input. Given that different omics data modalities convey distinct types of information, interpreting each modality separately allows us to identify key genes contributing to the prediction results, providing a comprehensive understanding of the biological processes involved and highlighting critical genes that may be overlooked when considering a single data source. In addition, since we introduced the stacking strategy, multiple explainers were applied to different classifiers in each omics to see whether there are overlaps among the base models in contributing gene selection. We filtered the top 10 genes in this process for each kernel explainer based on the selected features and the running time.

Results

Machine Learning and Deep Learning Classifiers Comparison for Selected Samples

WIMOAD In this paper, we initially selected SVM, Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Naïve Bayes (NB) and Multilayer Perceptron (MLP) as the classifiers. The Accuracy and AUC evaluation metrics of these classfiers's performance on gene expression (Exp) and methylation (Methl) data with group CN vs. EMCI are

shown in **Fig. 3.** With 10 times 10-fold CV, no classifier shew a performance higher than 60% in both accuracy and AUC. In addition, as a commonly used deep learning method, convolutional Multilayer Perceptron (MLP) did not exhibit higher AUC scores than conventional machine learning classifiers in the majority of groups for both omics. We finally applied some of the commonly used classifiers as the base models for further stacking study to achieve higher overall performance.

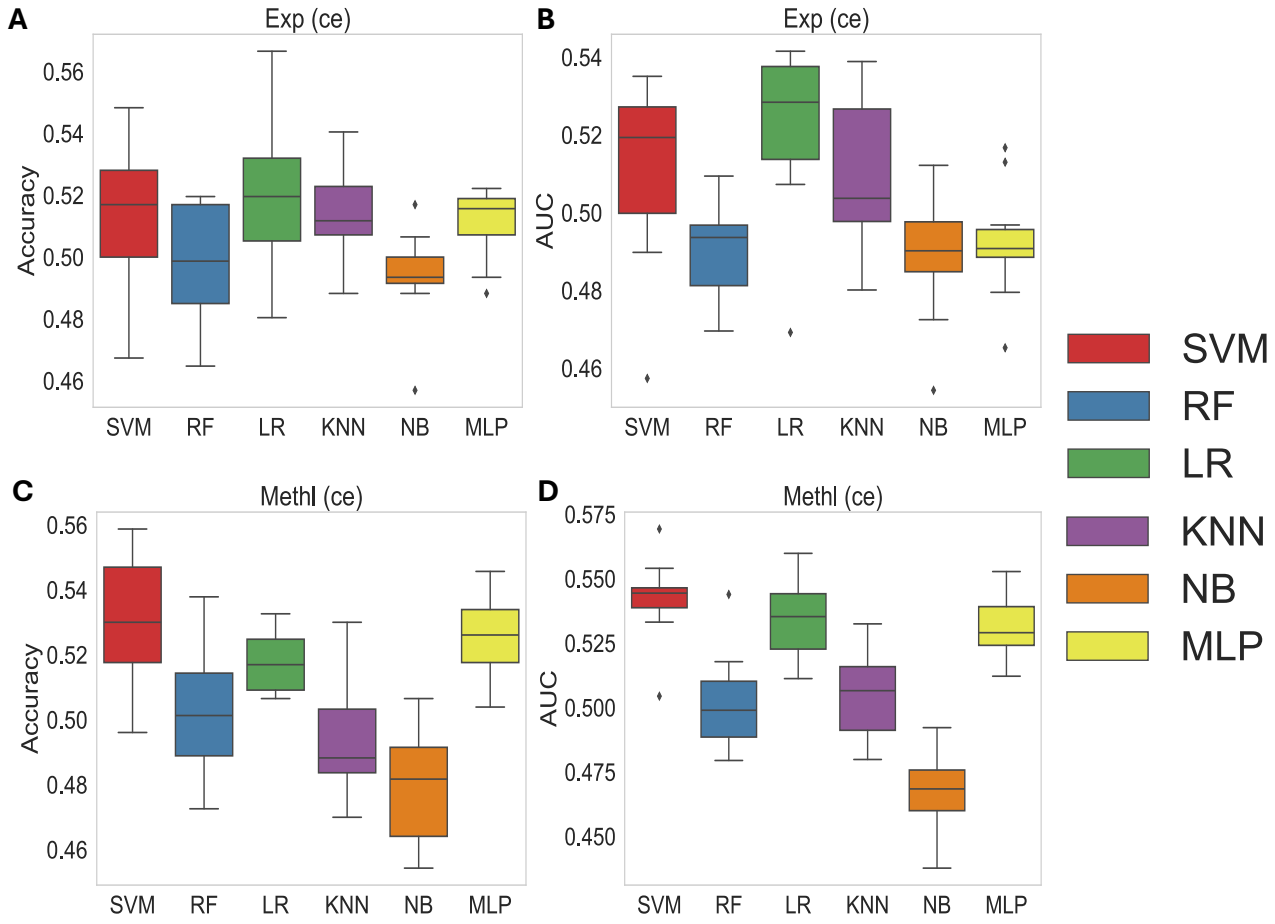


Figure. 3. Comparing the Performance using one Classifier Directly on the Collected data for Binary Classification. *All classifiers are trained with the same feature dimensions and 10 times 10-fold CV on CN vs. EMCI group. The performance was measured using the metric introduced previously.* ***(A-B)*** *Accuracy and AUC score comparison on gene expression data.* ***(C-D)*** *Accuracy*

and AUC score comparison on gene methylation data. SVM: Support Vector Machine LR: Logistic Regression; MLP: Multilayer Perceptron; RF: Random Forest; NB: Naïve Bayes; KNN: K-Nearest Neighbor.

The Stacking Ensemble Learning has Dramatically Improved the Overall Outcome

Classifiers ensemble is due to the premise that ensembles can often achieve better performance than individual classifiers. Except for general voting, stacking is also commonly used, which combines the predictions of base-level classifiers together with the class label to establish the meta-level dataset for decision-making, and is found to outperform voting [1][70]. We applied the stacking technique using a three-layer (one hidden layer) MLP as the meta-model to enhance the five base classifier outputs on single omics classification [71]. **Fig. 4** shows the CN vs. EMCI group results in comparison before (SVM as the only classifier) and after introducing stacking, including gene expression and methylation. Overall, there is about 20% improvement in the performance matrix (Accuracy, Precision, Specificity, AUC) after applying stacking. Among the three feature selection methods, ANOVA F-test selection achieved the highest performance after stacking. We then select the ANOVA F-test for the feature selection block during the integration model establishment.

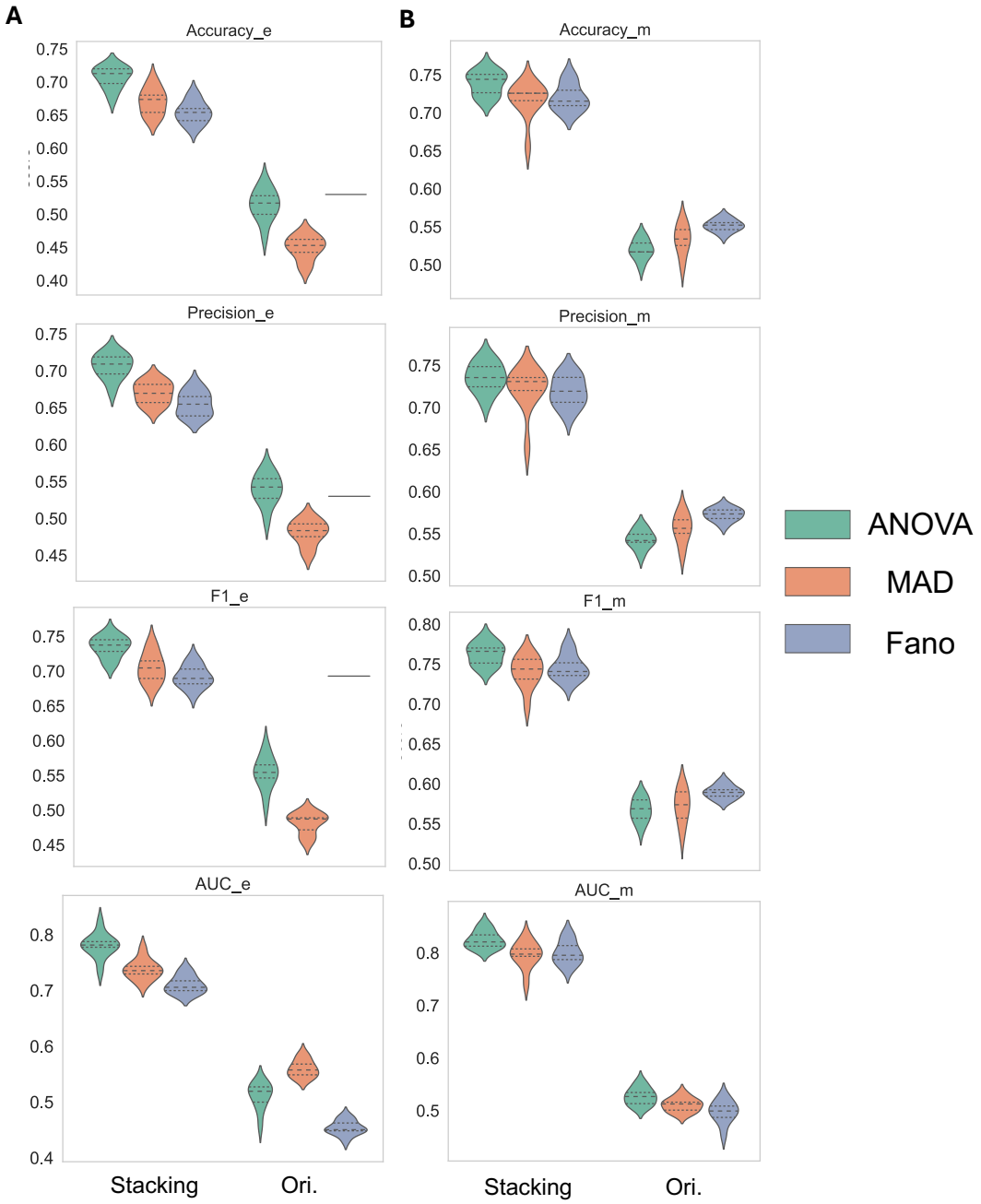


Figure. 4. Model Improvement After Stacking. *The results are based on CN vs. EMCI Group.* ***(A)*** *Classification performance improvement using gene expression data only before and after stacking.* ***(B)*** *Classification performance improvement using gene methylation data only before and after stacking. "_e": gene expression; "_m": gene methylation; ANOVA: ANOVA F-test for*

feature selection; MAD: Median Absolute Deviation; Stacking: Results for stacking models; Ori.: Results using one classifier (SVM) only.

Integration Model Achieved Higher Performance Than One Modality Only

WIMOAD is a weighted score fusion model for binary group classification, with distinctly assigned weight coefficients to balance the contribution of each data modality when reducing the negative effect that results from the data collection to the minimum. **Fig. 5** show how the coefficient of the Exp meta-model impacts the prediction accuracy of the final output. With optimized weights, the value of feature integration and the potential for original sampling exceed the performance of both Exp and Methl meta-model outputs. According to the AUC comparison, the integration model can outperform both omics when assigning weight from 0.2 to 0.8, when achieving the peaks around 0.5. Only the CN vs. EMCI group archives the peak when the weight for the Exp meta-model is 0.4. For convenience of the test, we assigned the weight coefficient as 0.5 for each meta-model for further study.

Our constructed WIMOAD integration model demonstrated an improvement in performance relative to single modality models, effectively mitigating the impact of poorly performing data on the final classification results with pre-optimized weight coefficients for both omics. As illustrated in **Fig. 6**, the integration model significantly enhanced the overall performance compared to using one omics only.

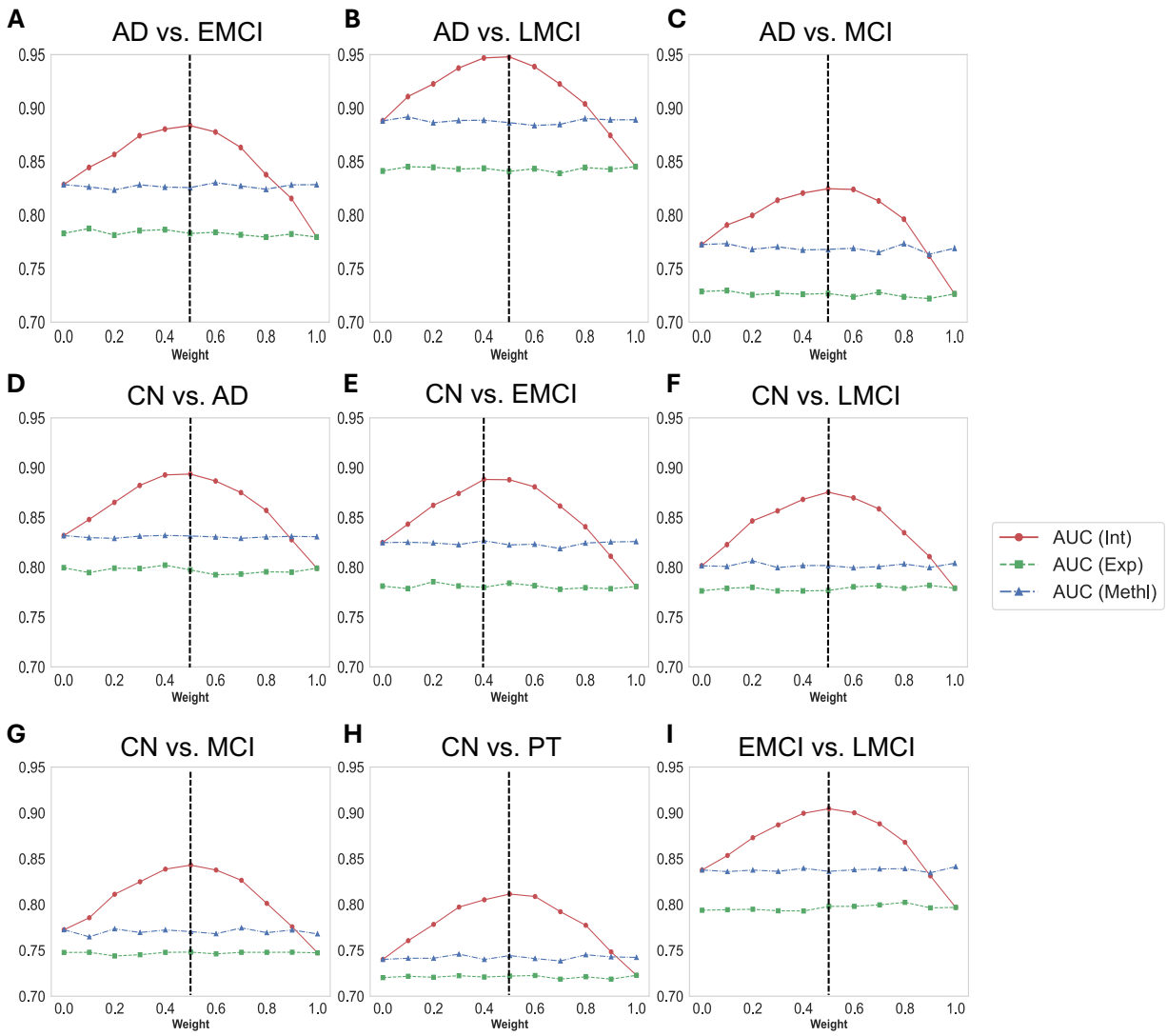
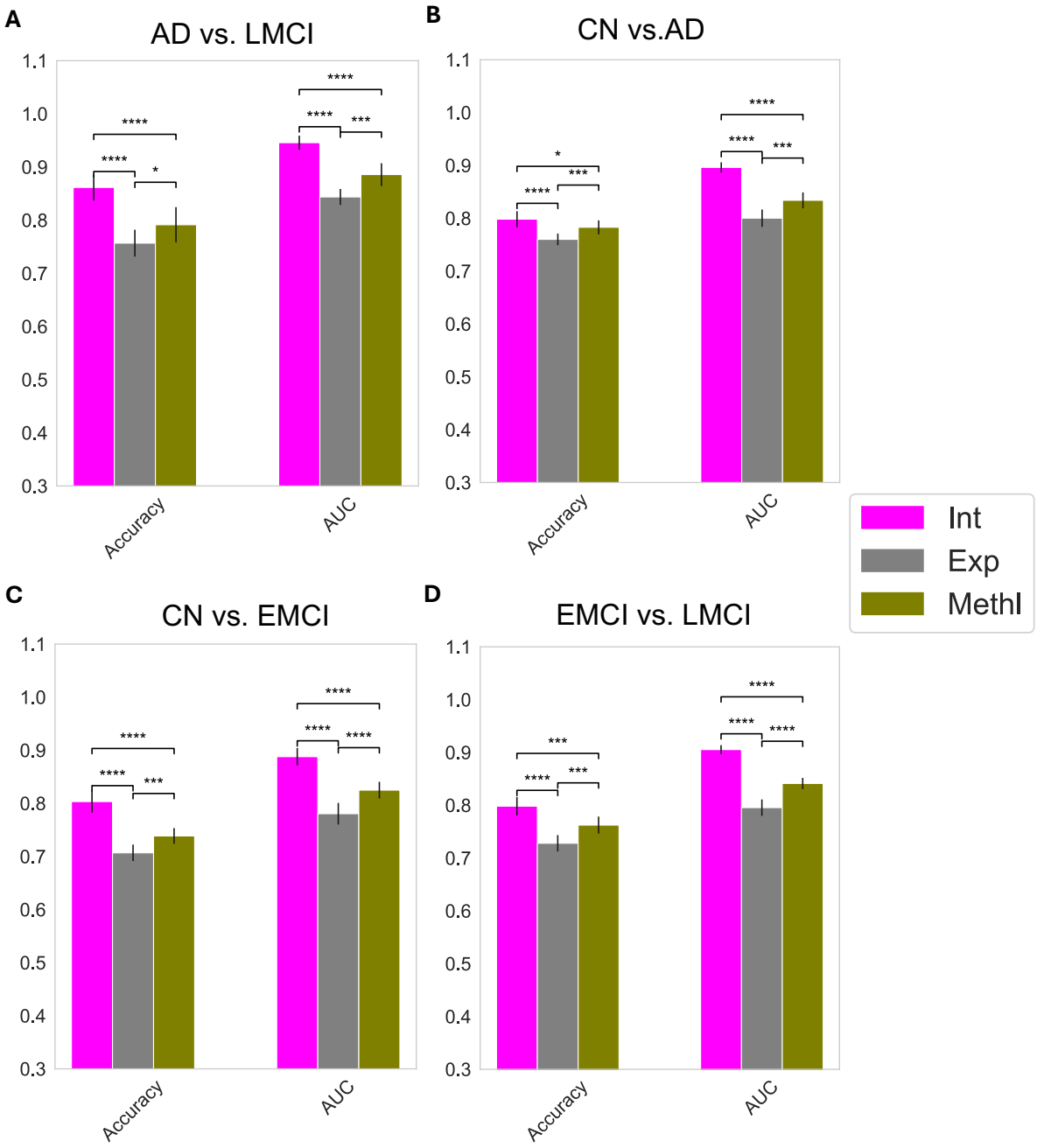


Figure. 5 Variation in AUC of the Integration Model with Changes in the Integration Coefficient. *the x-axis represents the increase of the integration coefficient α, which is the weight assigned to the prediction results of Exp classifier. The y-axis represents the accuracy of the model. The vertical dashed black line represents the highest AUC with respect to the weight coefficient α. In most tasks (8 out of 9), the integration has the best performance when α = 0.5.* ***(A)*** *AD vs. EMCI group.* ***(B)*** *AD vs. LMCI group.* ***(C)*** *AD vs. MCI group.* ***(D)*** *CN vs. AD group.* ***(E)*** *CN vs. EMCI group.* ***(F)*** *CN vs. LMCI group.* ***(G)*** *CN vs. MCI group.* ***(H)*** *CN vs. PT group.* ***(I)*** *EMCI vs. LMCI group.*

304



305

306 ***Figure. 6 Integration Performances of WIMOAD.*** *The x-axis represents the evaluation matrix,*

307 *and the y-axis represents the values. The results were generated under the best coefficient selected*

308 *(α = 0.5)and cross-validated 10 times.* ***(A)*** *AD vs. LMCI group.* ***(B)*** *CN vs. AD group.* ***(C)*** *CN vs.*

LMCI group. ***(D)*** *EMCI vs. LMCI group. '*':p<0.05; '**': p<0.01; '***': p<0.001; '****': p<0.0001.*

Comparison with State-of-the-art Predictors

Table. 2 compares the performance of WIMOAD against the state-of-the-art predictors for AD diagnosis using the paired ADNI data in our case. Across all the nine groups, the WIMOAD demonstrates consistently higher accuracies (77.6% on average compared to 70.4% using IntegrationLearner [72] and 45.6% using MOGLAM [73]) and AUCs (86.9% on average compared to 69.4% with IntegrationLearner and 53.7% with MOGLAM) compared to the existing integration methods.

Table. 2 Comparing state-of-the-art methods. *All model apply ADNI data as input source.*

Groups	WIMOAD		IntegrationLearner [72]		MOGLAM [73]	
	Acc	AUC	Acc	AUC	Acc	AUC
AD vs. EMCI	**0.776**	**0.882**	0.712	0.686	0.333	0.531
AD vs. LMCI	**0.862**	**0.946**	0.698	0.743	0.450	0.495
AD vs. MCI	**0.776**	**0.830**	0.767	0.660	0.237	0.487
CN vs. AD	**0.798**	**0.896**	0.730	0.706	0.310	0.494
CN vs. EMCI	**0.803**	**0.888**	0.662	0.706	0.474	0.536
CN vs. LMCI	**0.773**	**0.873**	0.715	0.709	0.355	0.673
CN vs. MCI	**0.743**	**0.845**	0.671	0.678	0.592	0.574
CN vs. PT	**0.709**	**0.810**	0.685	0.671	0.733	0.489
EMCI vs. LMCI	**0.740**	**0.847**	0.695	0.685	0.621	0.556
Avg	**0.776**	**0.869**	0.704	0.694	0.456	0.537

320

321 **Contributing Genes Identification According to Shapley Values**

322 We leveraged SHAP explainer to enhance the interpretability of our approach by analyzing
323 the importance of each most variable genes selected for model output. As demonstrated in **Fig. 7**
324 **(A-E),** the gene contributions represented by their respective SHAP values' magnitudes were
325 ranked for gene expression data of group CN vs. EMCI, elucidating the top 10 genes exerting the
326 most substantial influence on model predictions. Remarkably, discernible variations emerged
327 across different binary groups and omics data types. It becomes evident that the regulatory
328 dynamics, manifested through gene upregulation or downregulation, yield bidirectional effects on
329 the model's decision boundaries, influencing the classification outcome for individual samples.
330 After the ntersection of top5 contributing genes among five classifiers, *ABRA (Actin-binding Rho-*
331 *activating protein)* is the gene present in the overlap. In the SHAP summary plot, if the ABRA
332 gene expression level is high, the model is more likely to predict the sample as EMCI.

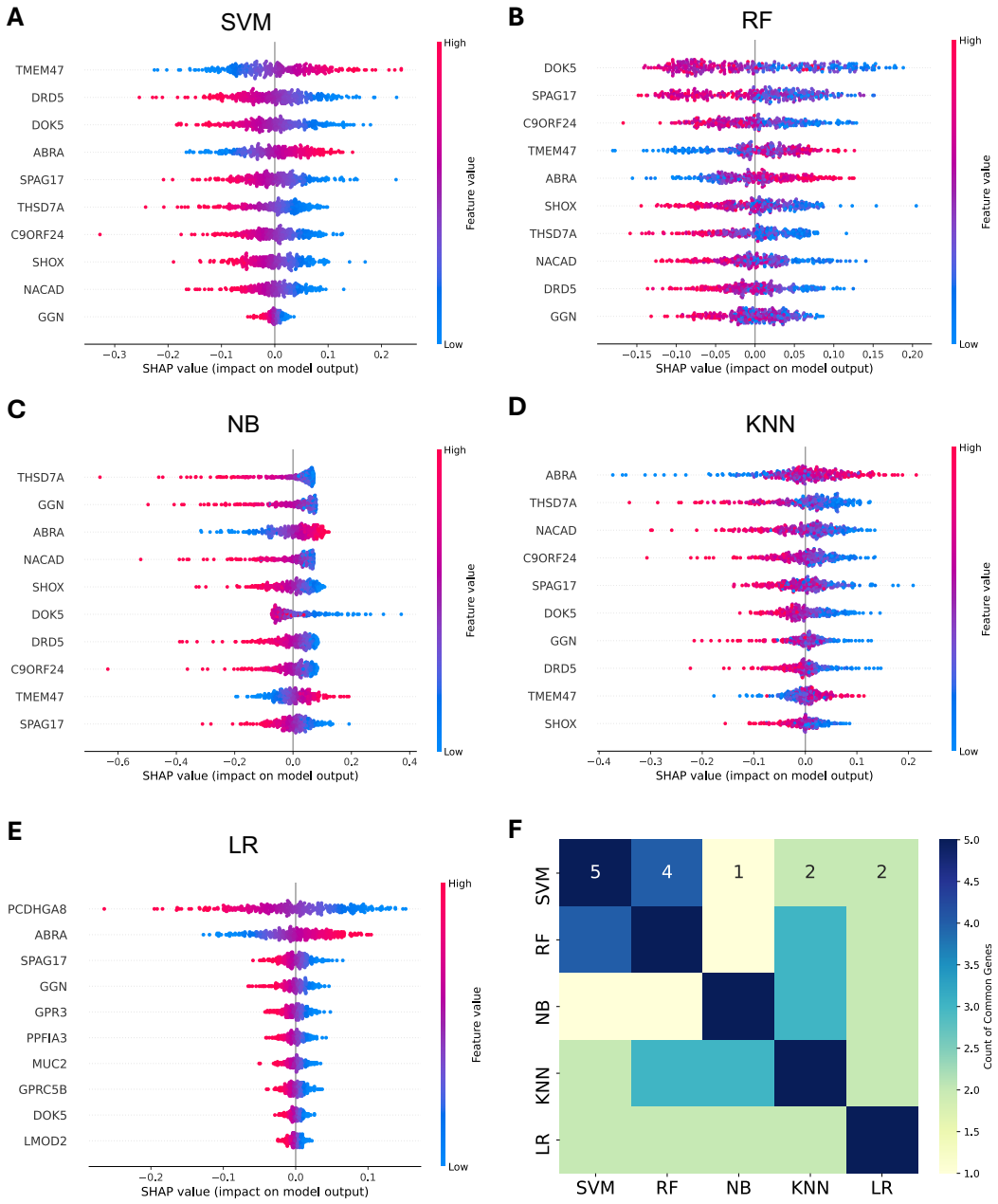


Fig. 7. SHAP Plots for Model Explanation and Contributing Genes Detection. *Top 10 most contributing genes and their influence on the model classification (sample being classified as EMCI) were exhibited.* ***(A-E)*** *SHAP summary plots for gene expression classifier of CN vs. EMCI group. the colors show the gene expression/methylation level of certain genes, and the SHAP*

values of the certain gene for each sample are denoted in the x-axis. Higher SHAP values for a certain gene represent the higher possibility that with the expression/methylation value, the model will classify the sample as AD. ***(F)*** *The heatmap showing the overlapping genes of five gene sets generated from the top5 contributing genes in each classifier.*

Discussion

In this study, we introduce WIMOAD (Weighted Integration for Superior Alzheimer's Diagnosis), a supervised binary classification model that integrates the stacked classification results from gene expression and methylation data through a weighted score fusion approach for early diagnosis of AD. Additionally, the model applied SHAP to interpret the contributions of different omics data and revealed distinct contributing genes across various data sources. According to the 10 times 10-fold CV results, WIMOAD improves the overall performance by integrating two omics in the binary classification task, especially in the classification case between health control and early mild cognitive impairment.

WIMOAD is an integration model based on meta-learning. As the convolutional and MLP-based classifiers and algorithms that applied deep learning did not provide better performances with the datasets according to the classifiers comparison, we established meta-models that take the predictions from different classifiers and the test label as a training dataset for model improvement for each omics. By assigning weights for the score generated by each classifier to different omics data profiles, there is a general increase in the model output, which results in one or more peaks that the performance matrix of the model can surpass using single omics in the classification task.

After the establishment of the model, we tested other multimodal fusion models, such as IntegrationLearner from Mallick et al. [72], a novel Bayesian ensemble method that combines

361 information across several longitudinal and cross-sectional omics data layers, and MoGLAM from
362 Ouyang et al. [73], which integrates a dynamic graph convolutional network, attention mechanism,
363 and omic integrated representation learning modules for fusing DNA methylation, miRNA, and
364 mRNA expression profiles for disease classification. Comparative analysis revealed that
365 WIMOAD consistently outperformed these methods across all classification groups. A likely
366 reason for WIMOAD's superior results is its use of weighted score fusion to aggregate predictions
367 from different classifiers, followed by decision-making, rather than directly concatenating data
368 from various sources as input for predictions.

369 For the interpretability of the model, WIMOAD applied SHAP for each data modality.
370 Instead of directly combining data, WIMOAD can extract specific representations from different
371 data modalities simultaneously and fully use all the information for the prediction. By quantifying
372 the contributions of the most variable genes separately, WIMOAD will contribute to the detection
373 of new biomarkers in multi-omics for early diagnosis, biomarker discovery, and precision therapy
374 design in AD studies. Given that the SVM model can currently only utilize KernelSHAP—an
375 algorithm within SHAP with relatively high computational complexity and longer runtime—we
376 have limited our presentation to the top ten genes (both expression and methylation) that most
377 significantly influence the model's predictions. Integrating SHAP into the decision-making process
378 allows for the visualization of how gene expression/methylation levels affect model predictions as
379 well. For instance, a higher expression level of a particular gene correlates with a higher
380 corresponding Shapley value, indicating that when the model detects high expression of this gene
381 in a sample, it is more likely to classify the sample into a specific category. This demonstrates that
382 the gene's expression level has a direct impact on the model's final prediction. Consequently,
383 incorporating the SHAP explainer makes it feasible to identify new biomarkers. Additionally, in

binary classification cases, the results obtained from different groups could potentially serve as markers for identifying the various stages in the progression from healthy (CN) to MCI (EMCI and LMCI) and AD.

The limitations of the WIMOAD model primarily center on the number of modalities it deals with. WIMOAD currently integrates only gene expression and methylation data, whereas most state-of-the-art integration models incorporate three or more data modalities. During the development of WIMOAD, we attempted to include proteomics profiles [74] into consideration. However, only 129 samples met the criteria of having gene expression, methylation, and proteomics data after filtering, and these samples were only sufficient for the CN-LMCI binary classification task. As a result, the model is limited to two types of omics data. Notably, since our data all comes from the peripheral blood, the biomarker detection in the study needs further investigation about how it links with the change in the brain, and how it will contribute to the mechanism of the AD process.

Conclusion

In this paper, we proposed a weighted score fusion model named WIMOAD for multi-omics integration in AD diagnosis. It is a meta-learning-based model that extracts information from both gene expression and paired methylation profiles of samples for model decision-making. Compared to the most recent models presented that incorporate statistical analysis and deep learning algorithms, WIMOAD has surpassed most classification tasks with genetic data. By adding the SHAP explainer in the workflow, top contributing genes or biomarkers from different omics and how they affect the model classification results can be visualized. Additionally, WIMOAD is also flexible in the number of data modalities included and straightforward to

implement. The future direction of our research will include incorporating commonly utilized imaging data to develop a more comprehensive multi-modality-based diagnostic model that enhances AD diagnostics' robustness and clinical applicability in disease pathology.

Conflict of Interest

The authors have declared that no competing interests exist.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA036727. This work was supported by the American Cancer Society under award number IRG-22-146-07-IRG, and by the Buffett Cancer Center, which is supported by the National Cancer Institute under award number CA036727. This work was supported by the Buffet Cancer Center, which is supported by the National Cancer Institute under award number CA036727, in collaboration with the UNMC/Children's Hospital & Medical Center Child Health Research Institute Pediatric Cancer Research Group. This study was supported, in part, by the National Institute on Alcohol Abuse and Alcoholism (P50AA030407-5126, Pilot Core grant). This study was also supported by the Nebraska EPSCoR FIRST Award (OIA-2044049). This work was also partially supported by the National Institute of General Medical Sciences under Award Numbers P20GM103427 and P20GM130447. This study was in part financially supported by the Child Health Research Institute at UNMC/Children's Nebraska. This work was also partially supported by the University of Nebraska Collaboration Initiative Grant from the Nebraska Research Initiative (NRI). The content is solely the responsibility of the authors and does not necessarily represent the official views from the funding organizations.

Authors' contributions

SW conceived and designed the study. HX developed the algorithm, performed the experiments, analyzed the data and implemented the WIMOAD package. All authors participated in writing the paper. The manuscript was approved by all authors.

Data availability

All the data used in this manuscript are publicly available in the corresponding references. WIMOAD is available at https://github.com/wan-mlab/WIMOAD.

Reference

[1] C.R. Jack, D.A. Bennett, K. Blennow, M.C. Carrillo, B. Dunn, S.B. Haeberlein, D.M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, E. Liu, J.L. Molinuevo, T. Montine, C. Phelps, K.P. Rankin, C.C. Rowe, P. Scheltens, E. Siemers, H.M. Snyder, R. Sperling, Contributors, C. Elliott, E. Masliah, L. Ryan, N. Silverberg, NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease, Alzheimer's & Dementia 14 (2018) 535–562. https://doi.org/10.1016/j.jalz.2018.02.018.

[2] L. Pini, M. Pievani, M. Bocchetta, D. Altomare, P. Bosco, E. Cavedo, S. Galluzzi, M. Marizzoni, G.B. Frisoni, Brain atrophy in Alzheimer's Disease and aging, Ageing Research Reviews 30 (2016) 25–48. https://doi.org/10.1016/j.arr.2016.01.002.

[3] Alzheimers_report.pdf, (n.d.).

[4] F.E. Matthews, B.C.M. Stephan, L. Robinson, C. Jagger, L.E. Barnes, A. Arthur, C. Brayne, Cognitive Function and Ageing Studies (CFAS) Collaboration, A two decade dementia incidence comparison from the Cognitive Function and Ageing Studies I and II, Nat Commun 7 (2016) 11398. https://doi.org/10.1038/ncomms11398.

[5] J. Rasmussen, H. Langerman, Alzheimer's Disease – Why We Need Early Diagnosis, DNND Volume 9 (2019) 123–130. https://doi.org/10.2147/DNND.S228939.

[6] C.R. Jack, D.A. Bennett, K. Blennow, M.C. Carrillo, H.H. Feldman, G.B. Frisoni, H. Hampel, W.J. Jagust, K.A. Johnson, D.S. Knopman, R.C. Petersen, P. Scheltens, R.A. Sperling, B. Dubois, A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers, Neurology 87 (2016) 539–547. https://doi.org/10.1212/WNL.0000000000002923.

[7] S. Sadigh-Eteghad, B. Sabermarouf, A. Majdi, M. Talebi, M. Farhoudi, J. Mahmoudi, Amyloid-Beta: A Crucial Factor in Alzheimer's Disease, Medical Principles and Practice 24 (2014) 1–10. https://doi.org/10.1159/000369101.

[8] S.S. Sisodia, D.L. Price, Role of the β-amyloid protein in Alzheimer's disease, The FASEB Journal 9 (1995) 366–370. https://doi.org/10.1096/fasebj.9.5.7896005.

[9] T. Grimmer, M. Riemenschneider, H. Förstl, G. Henriksen, W.E. Klunk, C.A. Mathis, T. Shiga, H.-J. Wester, A. Kurz, A. Drzezga, Beta Amyloid in Alzheimer's Disease: Increased Deposition in Brain Is Reflected in Reduced Concentration in Cerebrospinal Fluid, Biological Psychiatry 65 (2009) 927–934. https://doi.org/10.1016/j.biopsych.2009.01.027.

[10] Y. Wang, E. Mandelkow, Tau in physiology and pathology, Nat Rev Neurosci 17 (2016) 22–35. https://doi.org/10.1038/nrn.2015.1.

[11] S. Muralidar, S.V. Ambi, S. Sekaran, D. Thirumalai, B. Palaniappan, Role of tau protein in Alzheimer's disease: The prime pathological player, International Journal of Biological Macromolecules 163 (2020) 1599–1617. https://doi.org/10.1016/j.ijbiomac.2020.07.327.

[12] L. Gjerum, B.B. Andersen, M. Bruun, A.H. Simonsen, O.M. Henriksen, I. Law, S.G. Hasselbalch, K.S. Frederiksen, Comparison of the clinical impact of 2-[18F]FDG-PET and cerebrospinal fluid biomarkers in patients suspected of Alzheimer's disease, PLoS One 16 (2021) e0248413. https://doi.org/10.1371/journal.pone.0248413.

[13] I. Bazarbekov, A. Razaque, M. Ipalakova, J. Yoo, Z. Assipova, A. Almisreb, A review of artificial intelligence methods for Alzheimer's disease diagnosis: Insights from neuroimaging to sensor data analysis, Biomedical Signal Processing and Control 92 (2024) 106023. https://doi.org/10.1016/j.bspc.2024.106023.

[14] Y. Tong, Z. Li, H. Huang, L. Gao, M. Xu, Z. Hu, Research of spatial context convolutional neural networks for early diagnosis of Alzheimer's disease, J Supercomput 80 (2024) 5279–5297. https://doi.org/10.1007/s11227-023-05655-9.

[15] S.K. Arjaria, A.S. Rathore, D. Bisen, S. Bhattacharyya, Performances of Machine Learning Models for Diagnosis of Alzheimer's Disease, Ann. Data. Sci. 11 (2024) 307–335. https://doi.org/10.1007/s40745-022-00452-2.

[16] K. Chen, Y. Weng, A.A. Hosseini, T. Dening, G. Zuo, Y. Zhang, A comparative study of GNN and MLP based machine learning for the diagnosis of Alzheimer's Disease involving data synthesis, Neural Netw 169 (2024) 442–452. https://doi.org/10.1016/j.neunet.2023.10.040.

[17] S. Al-Otaibi, M. Mujahid, A.R. Khan, H. Nobanee, J. Alyami, T. Saba, Dual Attention Convolutional AutoEncoder for Diagnosis of Alzheimer's Disorder in Patients Using Neuroimaging and MRI Features, IEEE Access 12 (2024) 58722–58739. https://doi.org/10.1109/ACCESS.2024.3390186.

[18] S. Miao, Q. Xu, W. Li, C. Yang, B. Sheng, F. Liu, T.T. Bezabih, X. Yu, MMTFN: Multi-modal multi-scale transformer fusion network for Alzheimer's disease diagnosis, International Journal of Imaging Systems and Technology 34 (2024) e22970. https://doi.org/10.1002/ima.22970.

[19] S. Suganyadevi, A.S. Pershiya, K. Balasamy, V. Seethalakshmi, S. Bala, K. Arora, Deep Learning Based Alzheimer Disease Diagnosis: A Comprehensive Review, SN COMPUT. SCI. 5 (2024) 391. https://doi.org/10.1007/s42979-024-02743-2.

[20] C.R. Jack, D.M. Holtzman, Biomarker Modeling of Alzheimer's Disease, Neuron 80 (2013) 1347–1358. https://doi.org/10.1016/j.neuron.2013.12.003.

[21] Bateman Randall J., Xiong Chengjie, Benzinger Tammie L.S., Fagan Anne M., Goate Alison, Fox Nick C., Marcus Daniel S., Cairns Nigel J., Xie Xianyun, Blazey Tyler M., Holtzman David M., Santacruz Anna, Buckles Virginia, Oliver Angela, Moulder Krista, Aisen Paul S., Ghetti Bernardino, Klunk William E., McDade Eric, Martins Ralph N., Masters Colin L., Mayeux Richard, Ringman John M., Rossor Martin N., Schofield Peter R., Sperling Reisa A., Salloway Stephen, Morris John C., Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease, New England Journal of Medicine 367 (2012) 795–804. https://doi.org/10.1056/NEJMoa1202753.

[22] R.P. Friedland, E. Koss, J.V. Haxby, C.L. Grady, J. Luxenberg, M.B. Schapiro, J. Kaye, Alzheimer Disease: Clinical and Biological Heterogeneity, Ann Intern Med 109 (1988) 298–311. https://doi.org/10.7326/0003-4819-109-4-298.

[23] D. Ferreira, A. Nordberg, E. Westman, Biological subtypes of Alzheimer disease, Neurology 94 (2020) 436–448. https://doi.org/10.1212/WNL.0000000000009058.

[24] J.M. Ringman, A. Goate, C.L. Masters, N.J. Cairns, A. Danek, N. Graff-Radford, B. Ghetti, J.C. Morris, Dominantly Inherited Alzheimer Network, Genetic Heterogeneity in Alzheimer Disease and Implications for Treatment Strategies, Curr Neurol Neurosci Rep 14 (2014) 499. https://doi.org/10.1007/s11910-014-0499-8.

[25] H. Hampel, S. Lista, C. Neri, A. Vergallo, Time for the systems-level integration of aging: Resilience enhancing strategies to prevent Alzheimer's disease, Progress in Neurobiology 181 (2019) 101662. https://doi.org/10.1016/j.pneurobio.2019.101662.

[26] Utility of the Apolipoprotein E Genotype in the Diagnosis of Alzheimer's Disease | New England Journal of Medicine, (n.d.). https://www.nejm.org/doi/full/10.1056/nejm199802193380804 (accessed May 28, 2024).

[27] J.-C. Lambert, S. Heath, G. Even, D. Campion, K. Sleegers, M. Hiltunen, O. Combarros, D. Zelenika, M.J. Bullido, B. Tavernier, L. Letenneur, K. Bettens, C. Berr, F. Pasquier, N. Fiévet, P. Barberger-Gateau, S. Engelborghs, P. De Deyn, I. Mateo, A. Franck, S. Helisalmi, E. Porcellini, O. Hanon, European Alzheimer's Disease Initiative Investigators, M.M. de Pancorbo, C. Lendon, C. Dufouil, C. Jaillard, T. Leveillard, V. Alvarez, P. Bosco, M. Mancuso, F. Panza, B. Nacmias, P. Bossù, P. Piccardi, G. Annoni, D. Seripa, D. Galimberti, D. Hannequin, F. Licastro, H. Soininen, K. Ritchie, H. Blanché, J.-F. Dartigues, C. Tzourio, I. Gut, C. Van Broeckhoven, A. Alpérovitch, M. Lathrop, P. Amouyel, Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease, Nat Genet 41 (2009) 1094–1099. https://doi.org/10.1038/ng.439.

[28] P. Hollingworth, D. Harold, R. Sims, A. Gerrish, J.-C. Lambert, M.M. Carrasquillo, R. Abraham, M.L. Hamshere, J.S. Pahwa, V. Moskvina, K. Dowzell, N. Jones, A. Stretton, C. Thomas, A. Richards, D. Ivanov, C. Widdowson, J. Chapman, S. Lovestone, J. Powell, P. Proitsi, M.K. Lupton, C. Brayne, D.C. Rubinsztein, M. Gill, B. Lawlor, A. Lynch, K.S. Brown, P.A. Passmore, D. Craig, B. McGuinness, S. Todd, C. Holmes, D. Mann, A.D. Smith, H. Beaumont, D. Warden, G. Wilcock, S. Love, P.G. Kehoe, N.M. Hooper, E.R.L.C. Vardy, J. Hardy, S. Mead, N.C. Fox, M. Rossor, J. Collinge, W. Maier, F. Jessen, E. Rüther, B. Schürmann, R. Heun, H. Kölsch, H. van den Bussche, I. Heuser, J. Kornhuber, J. Wiltfang, M. Dichgans, L. Frölich, H. Hampel, J. Gallacher, M. Hüll, D. Rujescu, I. Giegling, A.M. Goate, J.S.K. Kauwe, C. Cruchaga, P. Nowotny, J.C. Morris, K. Mayo, K. Sleegers, K. Bettens, S. Engelborghs, P.P. De Deyn, C. Van Broeckhoven, G. Livingston, N.J. Bass, H. Gurling, A. McQuillin, R. Gwilliam, P. Deloukas, A. Al-Chalabi, C.E. Shaw, M. Tsolaki, A.B. Singleton, R. Guerreiro, T.W. Mühleisen, M.M. Nöthen, S. Moebus, K.-H. Jöckel, N. Klopp, H.-E. Wichmann, V.S. Pankratz, S.B. Sando, J.O. Aasly, M. Barcikowska, Z.K. Wszolek, D.W. Dickson, N.R. Graff-Radford, R.C. Petersen, Alzheimer's Disease Neuroimaging Initiative, C.M. van Duijn, M.M.B. Breteler, M.A. Ikram, A.L. DeStefano, A.L. Fitzpatrick, O. Lopez, L.J. Launer, S. Seshadri, CHARGE consortium, C. Berr, D. Campion, J. Epelbaum, J.-F. Dartigues, C. Tzourio, A. Alpérovitch, M. Lathrop, EADI1 consortium, T.M. Feulner, P. Friedrich, C. Riehle, M. Krawczak, S. Schreiber, M. Mayhaus, S. Nicolhaus, S. Wagenpfeil, S. Steinberg, H. Stefansson, K. Stefansson, J. Snaedal, S. Björnsson, P.V. Jonsson, V. Chouraki, B. Genier-Boley, M. Hiltunen, H. Soininen, O. Combarros, D. Zelenika, M. Delepine, M.J. Bullido, F. Pasquier, I. Mateo, A. Frank-Garcia, E. Porcellini, O. Hanon, E. Coto, V. Alvarez, P. Bosco, G. Siciliano, M. Mancuso, F. Panza, V. Solfrizzi, B. Nacmias, S. Sorbi, P. Bossù, P. Piccardi, B. Arosio, G. Annoni, D. Seripa, A. Pilotto, E. Scarpini, D. Galimberti, A. Brice, D. Hannequin, F. Licastro, L. Jones, P.A. Holmans, T. Jonsson, M. Riemenschneider, K. Morgan, S.G. Younkin, M.J. Owen, M. O'Donovan, P. Amouyel, J. Williams, Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease, Nat Genet 43 (2011) 429–435. https://doi.org/10.1038/ng.803.

[29] A. Forte, S. Lara, C. Peña-Bautista, M. Baquero, C. Cháfer-Pericás, New approach for early and specific Alzheimer disease diagnosis from different plasma biomarkers, Clinica Chimica Acta 556 (2024) 117842. https://doi.org/10.1016/j.cca.2024.117842.

[30] M.I. Kamboh, F.Y. Demirci, X. Wang, R.L. Minster, M.M. Carrasquillo, V.S. Pankratz, S.G. Younkin, A.J. Saykin, Alzheimer's Disease Neuroimaging Initiative, G. Jun, C. Baldwin, M.W. Logue, J. Buros, L. Farrer, M.A. Pericak-Vance, J.L. Haines, R.A. Sweet, M. Ganguli, E. Feingold, S.T. Dekosky, O.L. Lopez, M.M. Barmada, Genome-wide association study of Alzheimer's disease, Transl Psychiatry 2 (2012) e117. https://doi.org/10.1038/tp.2012.45.

[31] S. Guo, J. Yang, Bayesian genome-wide TWAS with reference transcriptomic data of brain and blood tissues identified 141 risk genes for Alzheimer's disease dementia, Alz Res Therapy 16 (2024) 120. https://doi.org/10.1186/s13195-024-01488-7.

[32] H. Mathys, J. Davila-Velderrain, Z. Peng, F. Gao, S. Mohammadi, J.Z. Young, M. Menon, L. He, F. Abdurrob, X. Jiang, A.J. Martorell, R.M. Ransohoff, B.P. Hafler, D.A. Bennett, M. Kellis, L.-H. Tsai, Single-cell transcriptomic analysis of Alzheimer's disease, Nature 570 (2019) 332–337. https://doi.org/10.1038/s41586-019-1195-2.

[33] S. Horvath, K. Raj, DNA methylation-based biomarkers and the epigenetic clock theory of ageing, Nat Rev Genet 19 (2018) 371–384. https://doi.org/10.1038/s41576-018-0004-3.

[34] A. Di Francesco, B. Arosio, A. Falconi, M.V. Micioni Di Bonaventura, M. Karimi, D. Mari, M. Casati, M. Maccarrone, C. D'Addario, Global changes in DNA methylation in Alzheimer's disease peripheral blood mononuclear cells, Brain, Behavior, and Immunity 45 (2015) 139–144. https://doi.org/10.1016/j.bbi.2014.11.002.

[35] X. Wei, L. Zhang, Y. Zeng, DNA methylation in Alzheimer's disease: In brain and peripheral blood, Mechanisms of Ageing and Development 191 (2020) 111319. https://doi.org/10.1016/j.mad.2020.111319.

[36] S. Qiu, M. Sun, Y. Xu, Y. Hu, Integrating multi-omics data to reveal the effect of genetic variant rs6430538 on Alzheimer's disease risk, Front. Neurosci. 18 (2024). https://doi.org/10.3389/fnins.2024.1277187.

[37] D. Das, J. Ito, T. Kadowaki, K. Tsuda, An interpretable machine learning model for diagnosis of Alzheimer's disease, PeerJ 7 (2019) e6543. https://doi.org/10.7717/peerj.6543.

[38] J. Zhang, X. Sun, X. Jia, B. Sun, S. Xu, W. Zhang, Z. Liu, Integrative multi-omics analysis reveals the critical role of the PBXIP1 gene in Alzheimer's disease, Aging Cell 23 (2024) e14044. https://doi.org/10.1111/acel.14044.

[39] T. Oka, Y. Matsuzawa, M. Tsuneyoshi, Y. Nakamura, K. Aoshima, H. Tsugawa, Multiomics analysis to explore blood metabolite biomarkers in an Alzheimer's Disease Neuroimaging Initiative cohort, Sci Rep 14 (2024) 6797. https://doi.org/10.1038/s41598-024-56837-1.

[40] A. Badhwar, G.P. McFall, S. Sapkota, S.E. Black, H. Chertkow, S. Duchesne, M. Masellis, L. Li, R.A. Dixon, P. Bellec, A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap, Brain 143 (2020) 1315–1331. https://doi.org/10.1093/brain/awz384.

[41] R. Nativio, Y. Lan, G. Donahue, S. Sidoli, A. Berson, A.R. Srinivasan, O. Shcherbakova, A. Amlie-Wolf, J. Nie, X. Cui, C. He, L.-S. Wang, B.A. Garcia, J.Q. Trojanowski, N.M. Bonini, S.L. Berger, An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease, Nat Genet 52 (2020) 1024–1035. https://doi.org/10.1038/s41588-020-0696-0.

[42] P. Kodam, R. Sai Swaroop, S.S. Pradhan, V. Sivaramakrishnan, R. Vadrevu, Integrated multi-omics analysis of Alzheimer's disease shows molecular signatures associated with disease progression and potential therapeutic targets, Sci Rep 13 (2023) 3695. https://doi.org/10.1038/s41598-023-30892-6.

[43] D. Wang, J. Gu, Integrative clustering methods of multi-omics data for molecule-based cancer classifications, Quant Biol 4 (2016) 58–67. https://doi.org/10.1007/s40484-016-0063-4.

[44] R. Shen, A.B. Olshen, M. Ladanyi, Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis, Bioinformatics 25 (2009) 2906–2912. https://doi.org/10.1093/bioinformatics/btp543.

[45] E.F. Lock, K.A. Hoadley, J.S. Marron, A.B. Nobel, Joint and individual variation explained (JIVE) for integrated analysis of multiple data types, Ann. Appl. Stat. 7 (2013). https://doi.org/10.1214/12-AOAS597.

[46] I. Gaynanova, G. Li, Structural learning and integrative decomposition of multi-view data, Biometrics 75 (2019) 1121–1132. https://doi.org/10.1111/biom.13108.

[47] J. Bao, C. Chang, Q. Zhang, A.J. Saykin, L. Shen, Q. Long, for the Alzheimer's Disease Neuroimaging Initiative, Integrative analysis of multi-omics and imaging data with incorporation of biological information via structural Bayesian factor analysis, Briefings in Bioinformatics 24 (2023) bbad073. https://doi.org/10.1093/bib/bbad073.

[48] N. Mahendran, D.R. Vincent P M, Deep belief network-based approach for detecting Alzheimer's disease using the multi-omics data, Computational and Structural Biotechnology Journal 21 (2023) 1651–1660. https://doi.org/10.1016/j.csbj.2023.02.021.

[49] S.R. Stahlschmidt, B. Ulfenborg, J. Synnergren, Multimodal deep learning for biomedical data fusion: a review, Briefings in Bioinformatics 23 (2022) bbab569. https://doi.org/10.1093/bib/bbab569.

[50] P. Lu, L. Hu, A. Mitelpunkt, S. Bhatnagar, L. Lu, H. Liang, A hierarchical attention-based multimodal fusion framework for predicting the progression of Alzheimer's disease, Biomedical Signal Processing and Control 88 (2024) 105669. https://doi.org/10.1016/j.bspc.2023.105669.
[51] M. Trinh, R. Shahbaba, C. Stark, Y. Ren, Alzheimer's disease detection using data fusion with a deep supervised encoder, Front. Dement. 3 (2024). https://doi.org/10.3389/frdem.2024.1332928.
[52] J.L. Ávila-Jiménez, V. Cantón-Habas, M. del P. Carrera-González, M. Rich-Ruiz, S. Ventura, A deep learning model for Alzheimer's disease diagnosis based on patient clinical records, Computers in Biology and Medicine 169 (2024) 107814. https://doi.org/10.1016/j.compbiomed.2023.107814.
[53] V. Bollati, D. Galimberti, L. Pergoli, E. Dalla Valle, F. Barretta, F. Cortini, E. Scarpini, P.A. Bertazzi, A. Baccarelli, DNA methylation in repetitive elements and Alzheimer disease, Brain, Behavior, and Immunity 25 (2011) 1078–1083. https://doi.org/10.1016/j.bbi.2011.01.017.
[54] M. Barrachina, I. Ferrer, DNA Methylation of Alzheimer Disease and Tauopathy-Related Genes in Postmortem Brain, Journal of Neuropathology & Experimental Neurology 68 (2009) 880–891. https://doi.org/10.1097/NEN.0b013e3181af2e46.
[55] S. Scarpa, R.A. Cavallaro, F. D'Anselmi, A. Fuso, Gene silencing through methylation: An epigenetic intervention on Alzheimer disease, Journal of Alzheimer's Disease 9 (2006) 407–414. https://doi.org/10.3233/JAD-2006-9406.
[56] M. Pammi, N. Aghaeepour, J. Neu, Multiomics, artificial intelligence, and precision medicine in perinatology, Pediatr Res 93 (2023) 308–315. https://doi.org/10.1038/s41390-022-02181-x.
[57] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, (n.d.).
[58] B. Pavlyshenko, Using Stacking Approaches for Machine Learning Models, in: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018: pp. 255–258. https://doi.org/10.1109/DSMP.2018.8478522.
[59] C. Cortes, V. Vapnik, Support-vector networks, Mach Learn 20 (1995) 273–297. https://doi.org/10.1007/BF00994018.
[60] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995: pp. 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994.
[61] G.I. Webb, Naïve Bayes, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of Machine Learning, Springer US, Boston, MA, 2010: pp. 713–714. https://doi.org/10.1007/978-0-387-30164-8_576.
[62] A. Das, Logistic Regression, in: A.C. Michalos (Ed.), Encyclopedia of Quality of Life and Well-Being Research, Springer Netherlands, Dordrecht, 2014: pp. 3680–3682. https://doi.org/10.1007/978-94-007-0753-5_1689.
[63] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN Model-Based Approach in Classification, in: R. Meersman, Z. Tari, D.C. Schmidt (Eds.), On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, Springer, Berlin, Heidelberg, 2003: pp. 986–996. https://doi.org/10.1007/978-3-540-39964-3_62.
[64] M.-C. Popescu, V.E. Balas, L. Perescu-Popescu, N. Mastorakis, Multilayer perceptron and neural networks, WSEAS Trans. Cir. and Sys. 8 (2009) 579–588.
[65] F. Murtagh, Multilayer perceptrons for classification and regression, Neurocomputing 2 (1991) 183–197. https://doi.org/10.1016/0925-2312(91)90023-5.
[66] N. Omer Fadl Elssied, O. Ibrahim, A. Hamza Osman, A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification, RJASET 7 (2014) 625–638. https://doi.org/10.19026/rjaset.7.299.

[67] D. Grün, L. Kester, A. van Oudenaarden, Validation of noise models for single-cell transcriptomics, Nat Methods 11 (2014) 637–640. https://doi.org/10.1038/nmeth.2930.
[68] P. Refaeilzadeh, L. Tang, H. Liu, Cross-Validation, in: L. LIU, M.T. ÖZSU (Eds.), Encyclopedia of Database Systems, Springer US, Boston, MA, 2009: pp. 532–538. https://doi.org/10.1007/978-0-387-39940-9_565.
[69] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D.E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, S.-I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat Biomed Eng 2 (2018) 749–760. https://doi.org/10.1038/s41551-018-0304-0.
[70] L. Todorovski, S. Džeroski, Combining Classifiers with Meta Decision Trees, Machine Learning 50 (2003) 223–249. https://doi.org/10.1023/A:1021709817809.
[71] V. Boateng, B. Yang, Ensemble Stacking with the Multi-Layer Perceptron Neural Network Meta-Learner for Passenger Train Delay Prediction, in: 2023 IEEE Conference on Artificial Intelligence (CAI), 2023: pp. 21–22. https://doi.org/10.1109/CAI54212.2023.00017.
[72] H. Mallick, A. Porwal, S. Saha, P. Basak, V. Svetnik, E. Paul, An integrated Bayesian framework for multi-omics prediction and classification, Statistics in Medicine 43 (2024) 983–1002. https://doi.org/10.1002/sim.9953.
[73] D. Ouyang, Y. Liang, L. Li, N. Ai, S. Lu, M. Yu, X. Liu, S. Xie, Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification, Computers in Biology and Medicine 164 (2023) 107303. https://doi.org/10.1016/j.compbiomed.2023.107303.
[74] D.A. Butterfield, D. Boyd-Kimball, A. Castegna, Proteomics in Alzheimer's disease: insights into potential mechanisms of neurodegeneration, J Neurochem 86 (2003) 1313–1327. https://doi.org/10.1046/j.1471-4159.2003.01948.x.

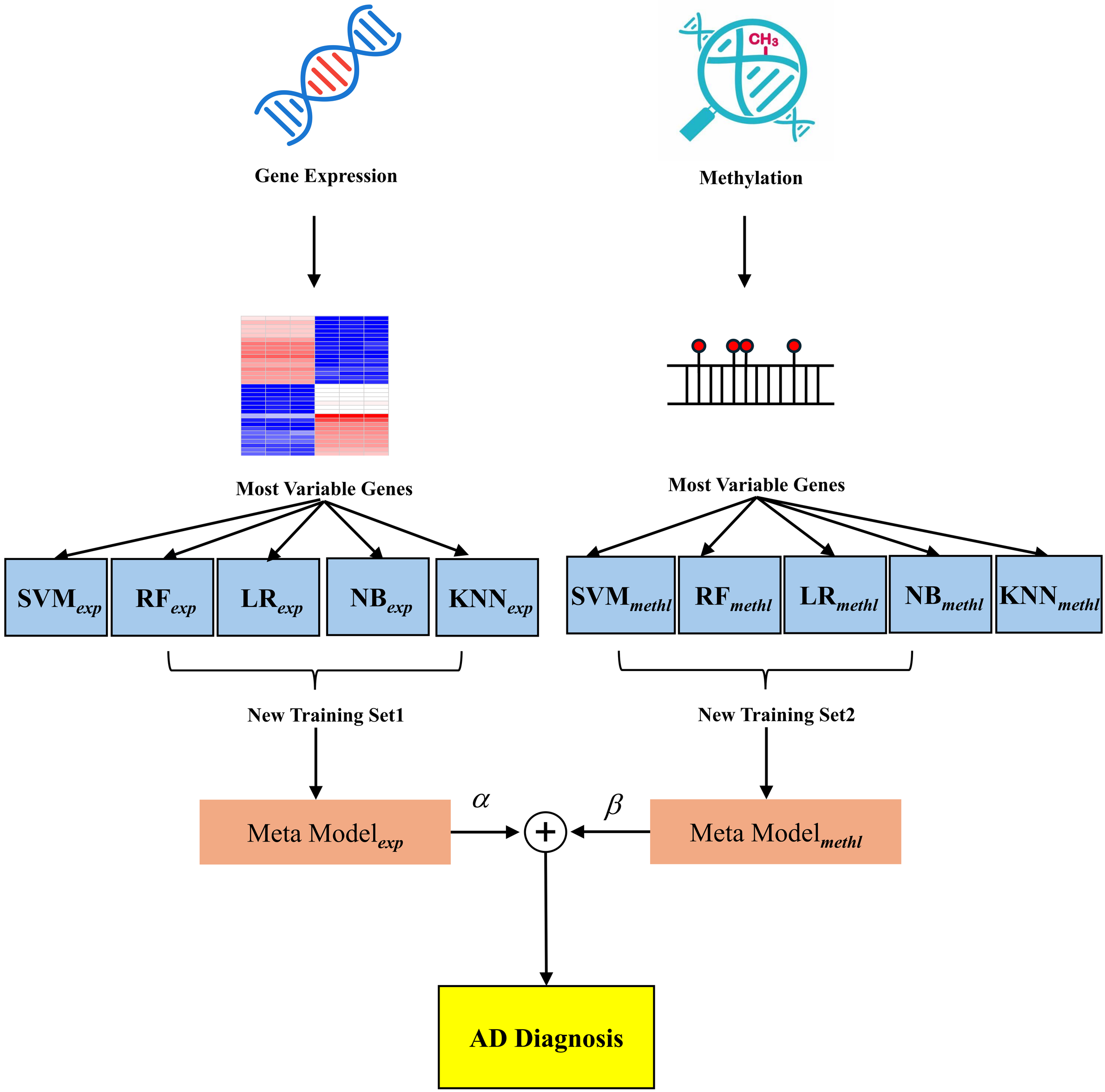
Table. 1 The demographic information of the Selected Participants. *Data are mean ± standard deviation (std). CN: Normal Controls; EMCI: Early Mild Cognitive Impairments; LMCI: Late Mild Cognitive Impairments; MCI: Mild Cognitive Impairments; AD: Alzheimer's Diseases; F: Female; M: Male*

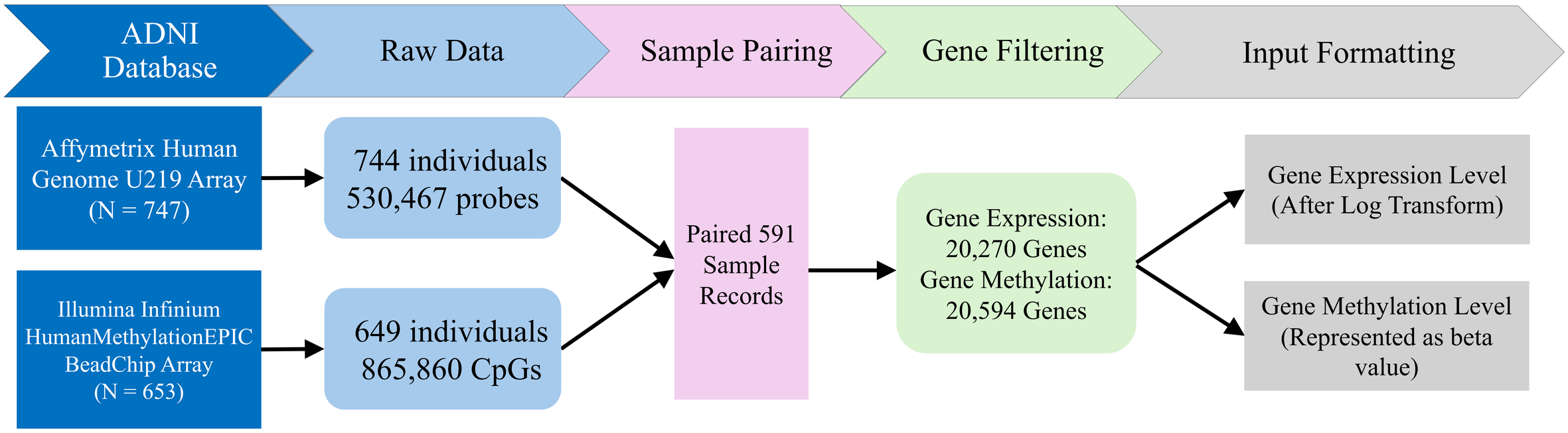
Diagnosis	Samples	Age (mean±std)	Sex (F/M)
CN	N = 203	74.45 ± 5.78	101/102
EMCI	N = 180	71.44 ± 7.11	81/ 99
LMCI	N = 113	72.74 ± 7.67	45/68
AD	N = 95	74.28 ± 7.59	35/60

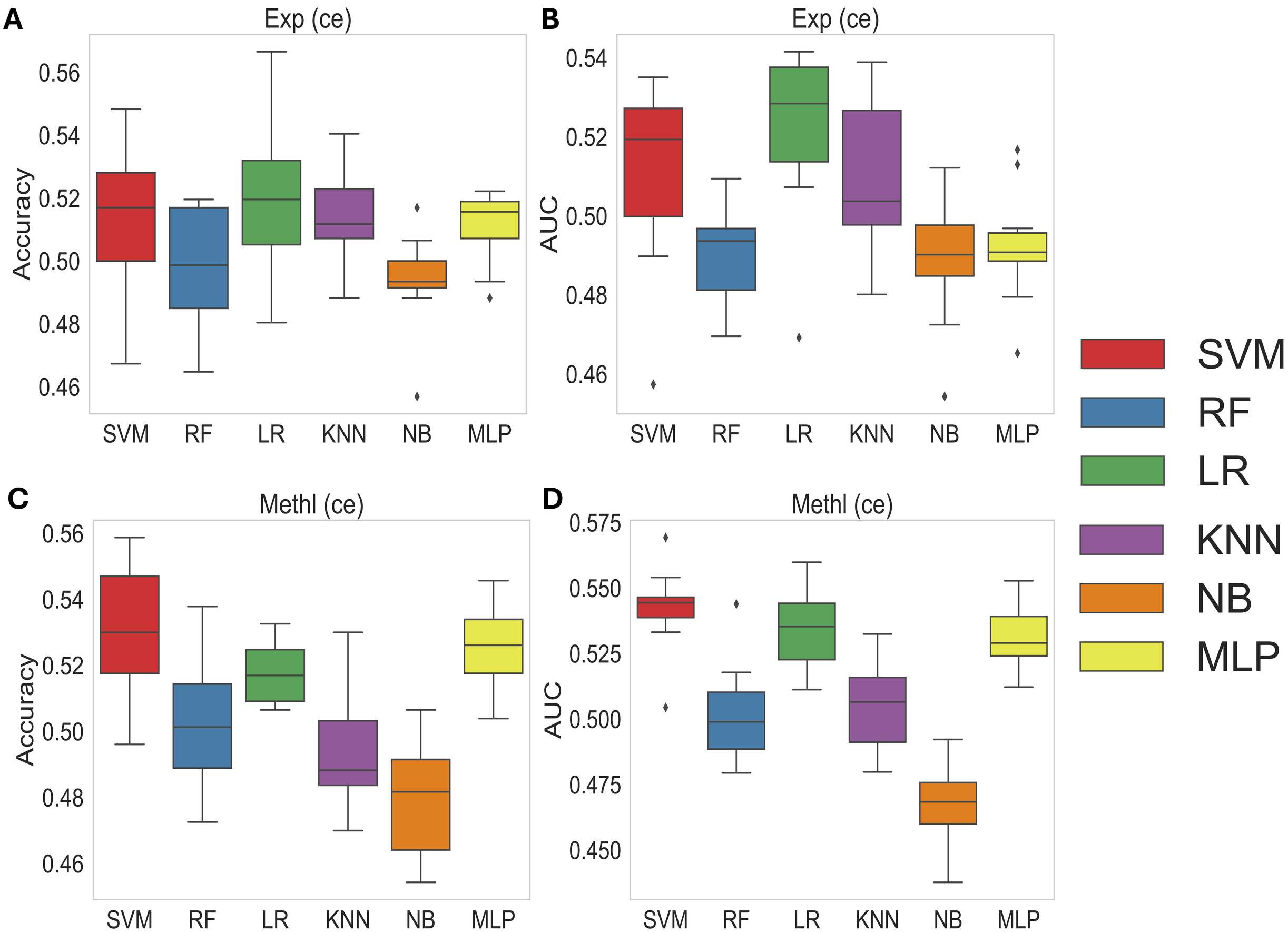
Table. 2 Comparing state-of-the-art methods. *All model apply ADNI data as input source.*

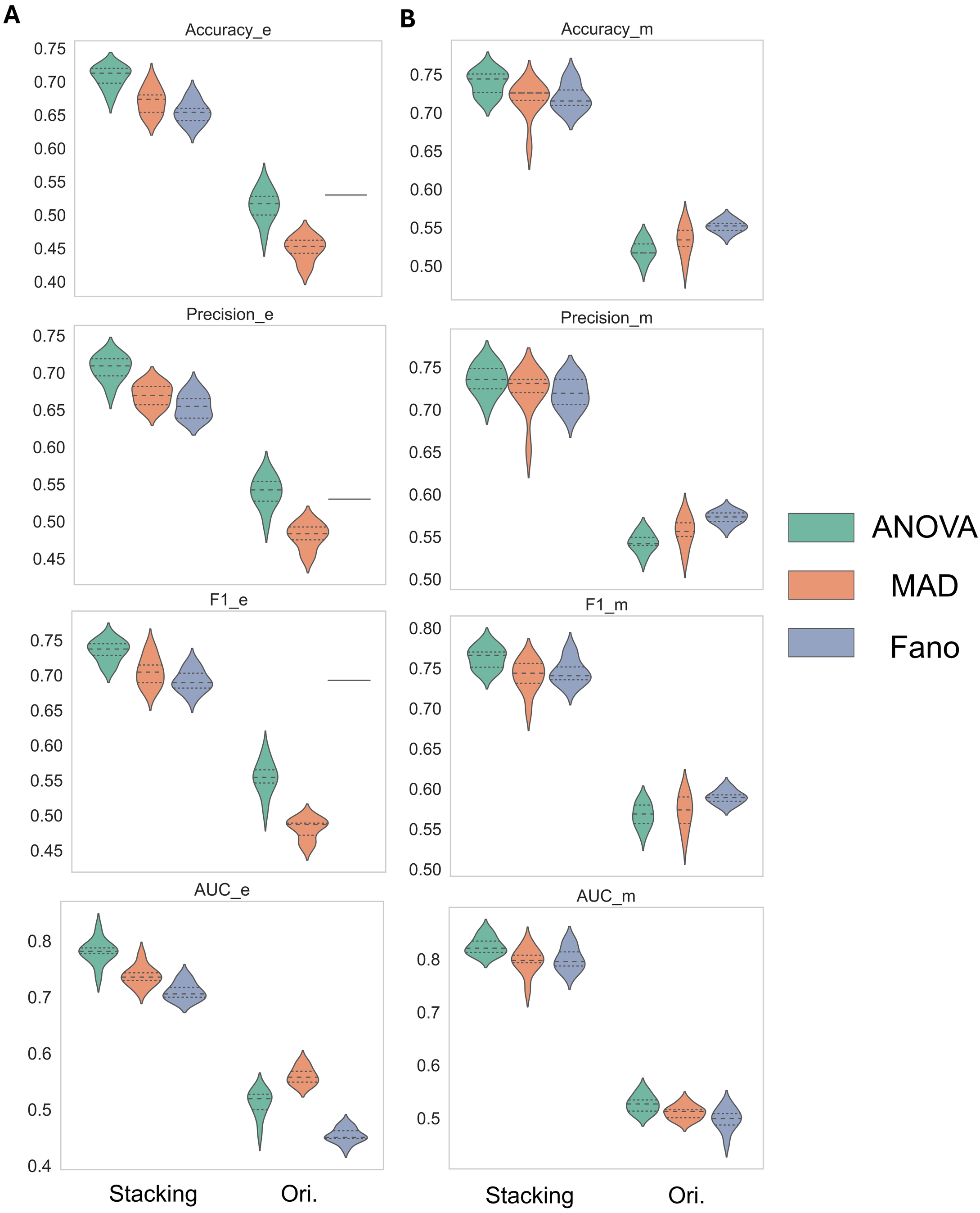
Groups	WIMOAD		IntegrationLearner [72]		MOGLAM [73]	
	Acc	AUC	Acc	AUC	Acc	AUC
AD vs. EMCI	**0.776**	**0.882**	0.712	0.686	0.333	0.531
AD vs. LMCI	**0.862**	**0.946**	0.698	0.743	0.450	0.495
AD vs. MCI	**0.776**	**0.830**	0.767	0.660	0.237	0.487
CN vs. AD	**0.798**	**0.896**	0.730	0.706	0.310	0.494
CN vs. EMCI	**0.803**	**0.888**	0.662	0.706	0.474	0.536
CN vs. LMCI	**0.773**	**0.873**	0.715	0.709	0.355	0.673
CN vs. MCI	**0.743**	**0.845**	0.671	0.678	0.592	0.574
CN vs. PT	**0.709**	**0.810**	0.685	0.671	0.733	0.489
EMCI vs. LMCI	**0.740**	**0.847**	0.695	0.685	0.621	0.556

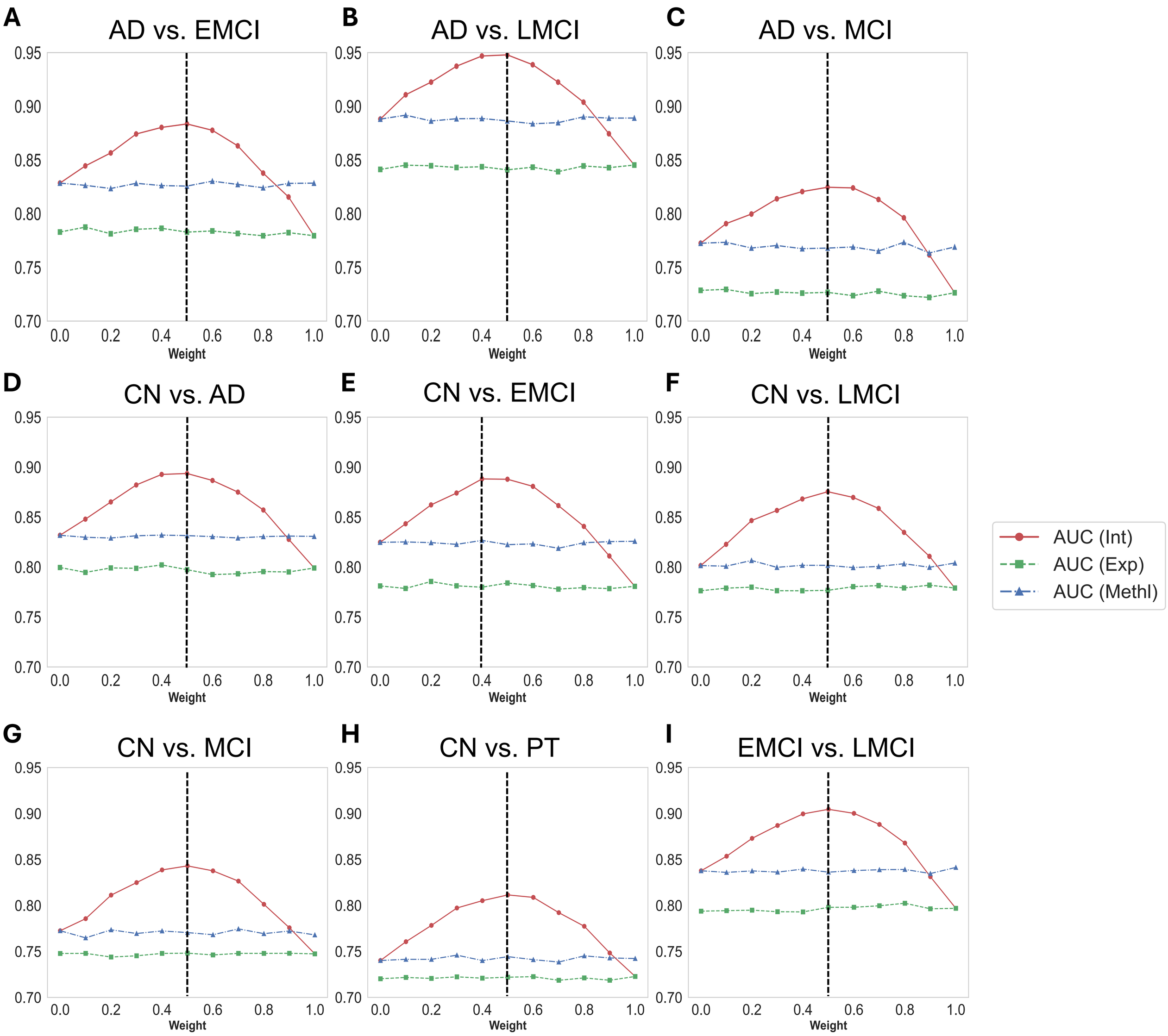
Avg	**0.776**	**0.869**	0.704	0.694	0.456	0.537

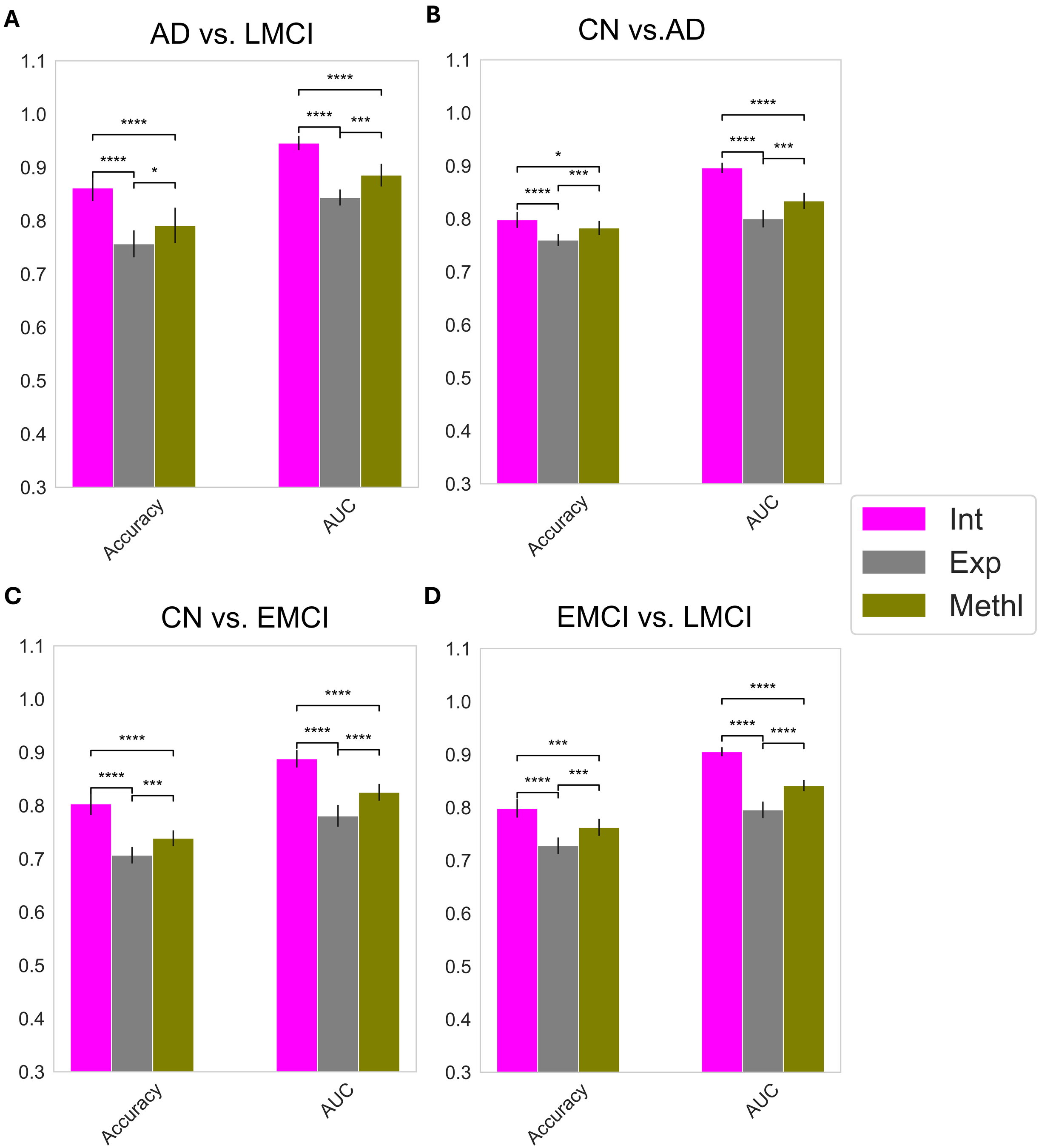

Gene Expression
Methylation
CH3
Most Variable Genes
Most Variable Genes
SVM exp
RF exp
LR exp
NB exp
KNN exp
SVM methl
RF methl
LR methl
NB methl
KNN methl
New Training Set1
New Training Set2
Meta Model exp
α
β
Meta Model methl
AD Diagnosis

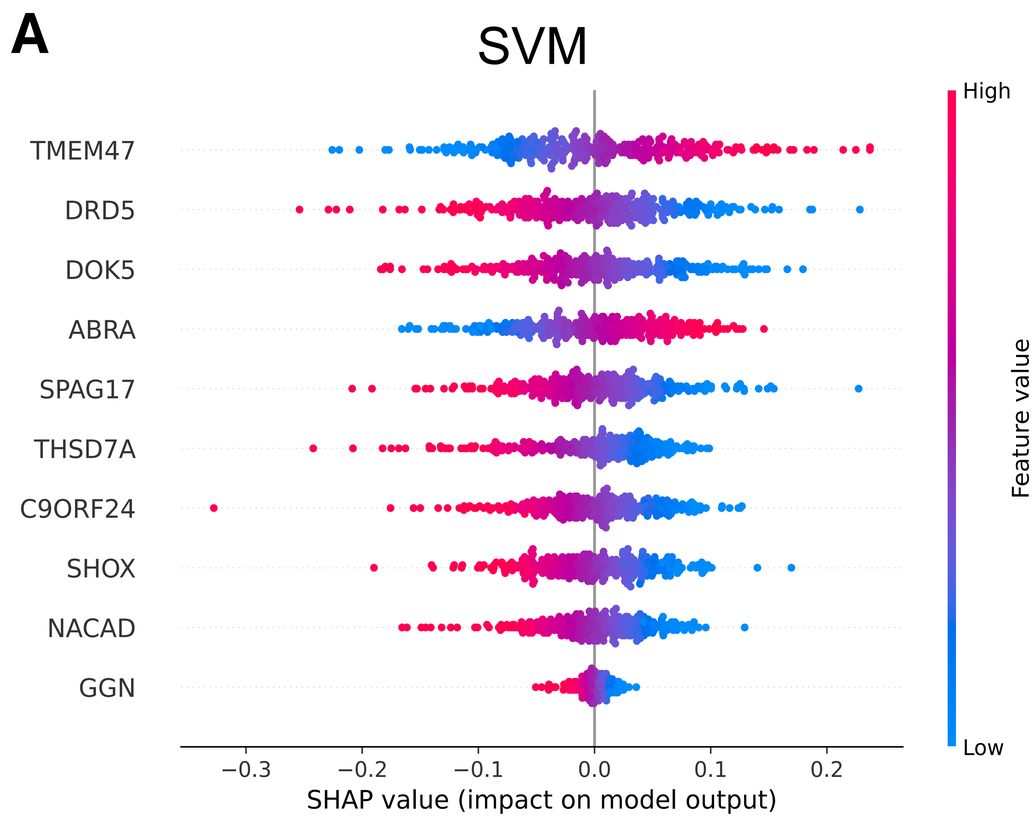

ADNI Database
Raw Data
Sample Pairing
Gene Filtering
Input Formatting
Affymetrix Human Genome U219 Array (N = 747)
Illumina Infinium HumanMethylationEPIC BeadChip Array (N = 653)
744 individuals 530,467 probes
649 individuals 865,860 CpGs
Paired 591 Sample Records
Gene Expression: 20,270 Genes Gene Methylation: 20,594 Genes
Gene Expression Level (After Log Transform)
Gene Methylation Level (Represented as beta value)

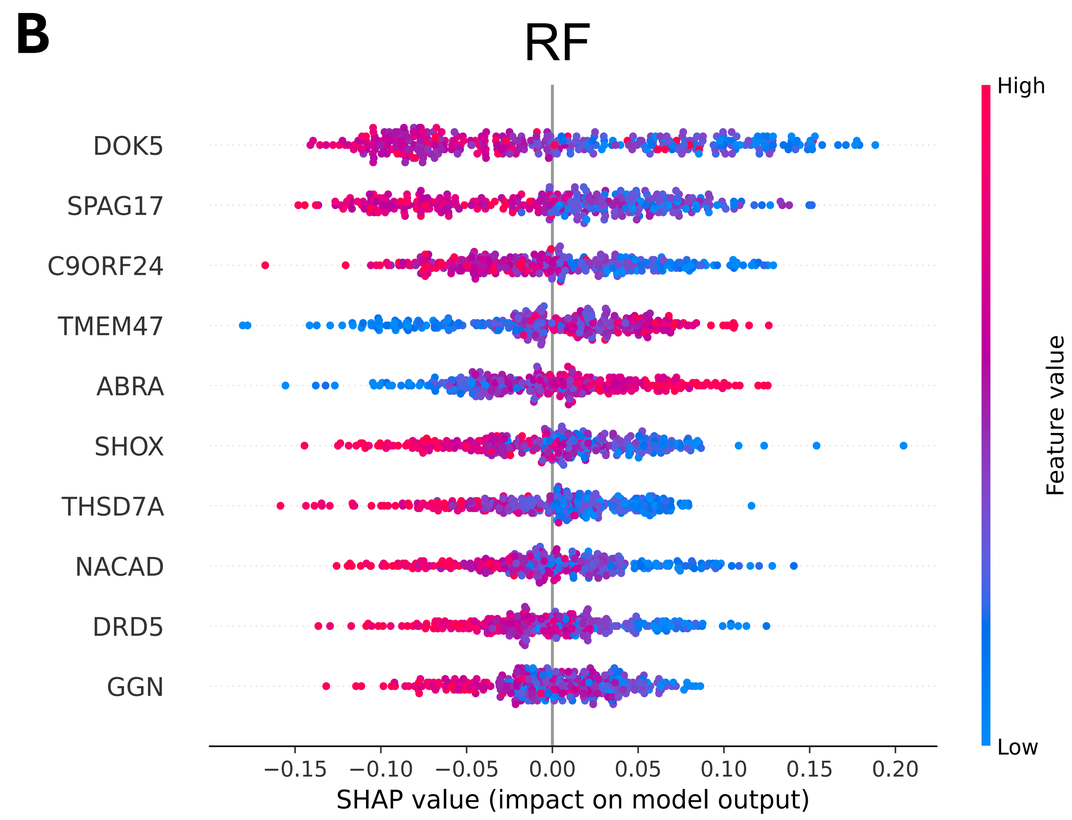

A
Exp (ce)
Accuracy
0.56
0.54
0.52
0.50
0.48
0.46
SVM
RF
LR
KNN
NB
MLP
B
Exp (ce)
AUC
0.54
0.52
0.50
0.48
0.46
SVM
RF
LR
KNN
NB
MLP
C
Methl (ce)
Accuracy
0.56
0.54
0.52
0.50
0.48
0.46
SVM
RF
LR
KNN
NB
MLP
D
Methl (ce)
AUC
0.575
0.550
0.525
0.500
0.475
0.450
SVM
RF
LR
KNN
NB
MLP
SVM
RF
LR
KNN
NB
MLP

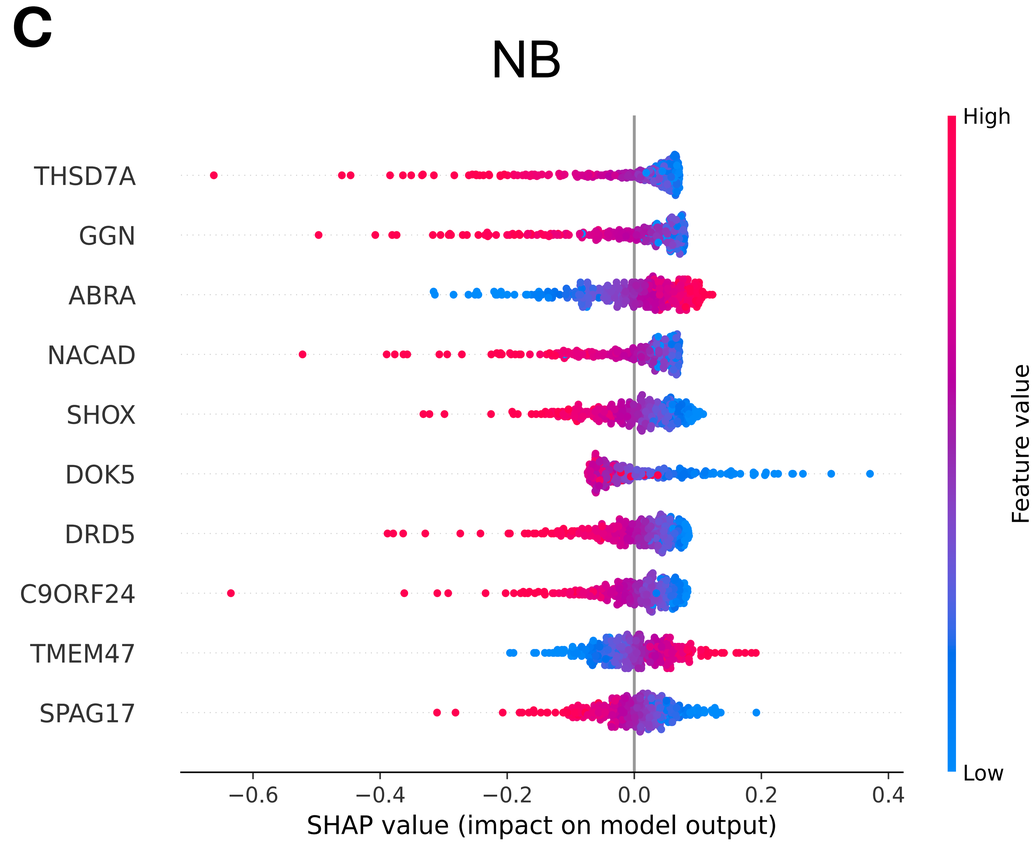

A
B
Accuracy_e
Accuracy_m
Precision_e
Precision_m
F1_e
F1_m
AUC_e
AUC_m
Stacking
Ori.
ANOVA
MAD
Fano

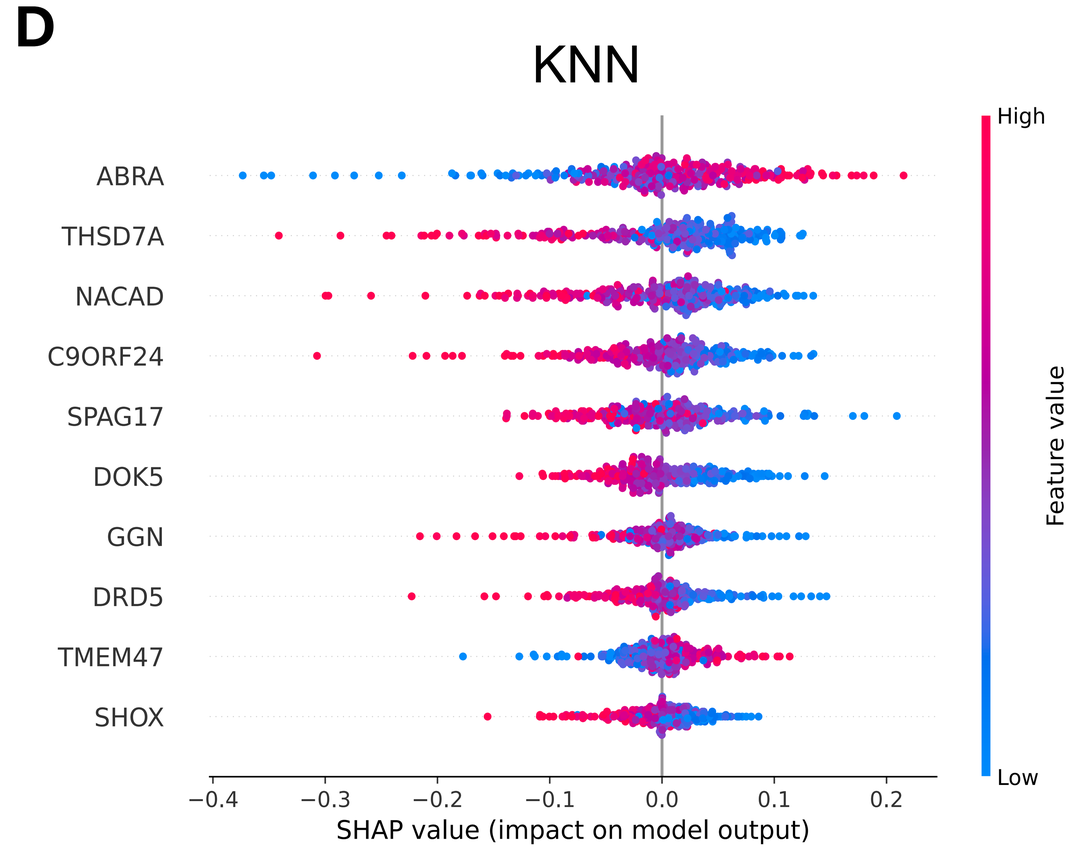

A
AD vs. EMCI
B
AD vs. LMCI
C
AD vs. MCI
D
CN vs. AD
E
CN vs. EMCI
F
CN vs. LMCI
G
CN vs. MCI
H
CN vs. PT
I
EMCI vs. LMCI
0.95
0.90
0.85
0.80
0.75
0.70
0.0
0.2
0.4
0.6
0.8
1.0
Weight
AUC (Int)
AUC (Exp)
AUC (Methl)

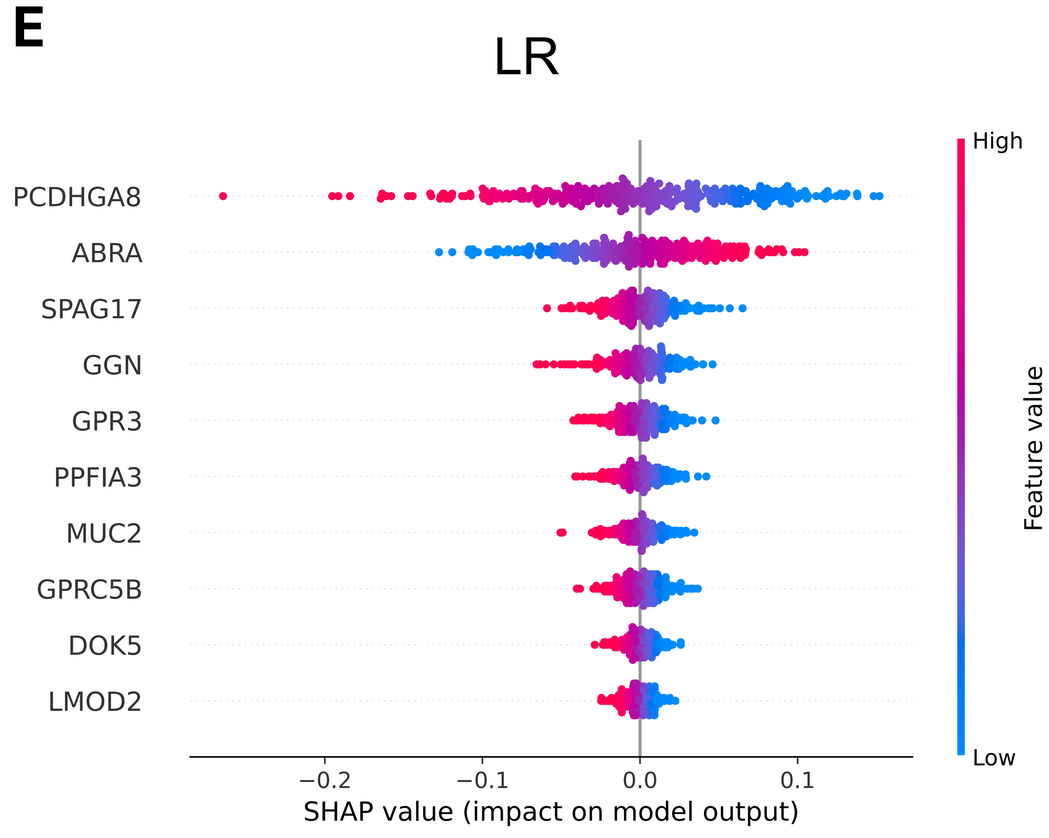

A
AD vs. LMCI
B
CN vs.AD
C
CN vs. EMCI
D
EMCI vs. LMCI
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Accuracy
AUC
Int
Exp
Methl

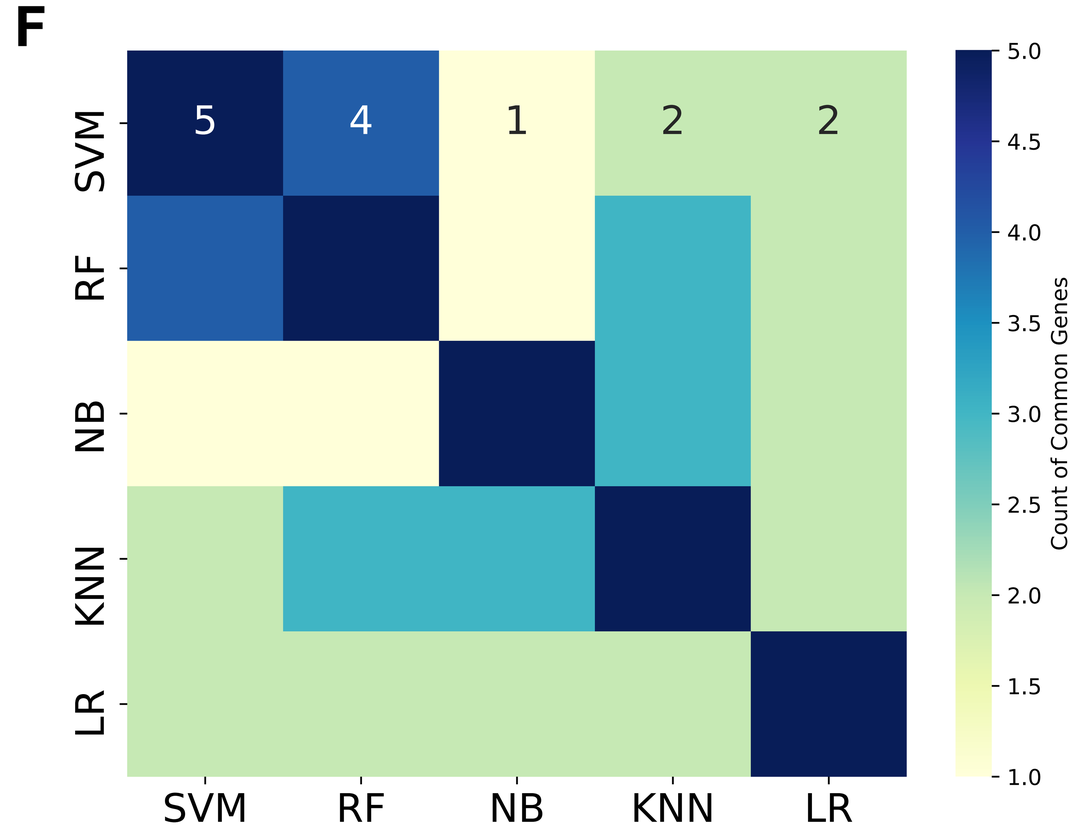

A
SVM
TMEM47
DRD5
DOK5
ABRA
SPAG17
THSD7A
C9ORF24
SHOX
NACAD
GGN
−0.3
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low


B
RF
DOK5
SPAG17
C9ORF24
TMEM47
ABRA
SHOX
THSD7A
NACAD
DRD5
GGN
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
SHAP value (impact on model output)
High
Feature value
Low


C
NB
THSD7A
GGN
ABRA
NACAD
SHOX
DOK5
DRD5
C9ORF24
TMEM47
SPAG17
−0.6
−0.4
−0.2
0.0
0.2
0.4
SHAP value (impact on model output)
High
Feature value
Low


D
KNN
ABRA
THSD7A
NACAD
C9ORF24
SPAG17
DOK5
GGN
DRD5
TMEM47
SHOX
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low


E
LR
PCDHGA8
ABRA
SPAG17
GGN
GPR3
PPFIA3
MUC2
GPRC5B
DOK5
LMOD2
−0.2
−0.1
0.0
0.1
SHAP value (impact on model output)
High
Feature value
Low


F
SVM
RF
NB
KNN
LR
5
4
1
2
2
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Count of Common Genes