Alzheimer's Disease Classification Using Genetic Data

Subash Khanal
Dept. of Computer Science
University of Kentucky
Lexington, KY
subash.khanal33@uky.edu

Jin Chen
Inst. for Biomedical Informatics
University of Kentucky
Lexington, KY
chen.jin@uky.edu

Nathan Jacobs
Dept. of Computer Science
University of Kentucky
Lexington, KY
jacobs@cs.uky.edu

Ai-Ling Lin
Dept. of Radiology
University of Missouri
Columbia, MO
ai-ling.lin@health.missouri.edu

***Abstract*—There has been a recent surge of interest in using genetic data to build ML-based accurate and interpretable disease classification models. In this line of research, we separately assess the potential of the peripheral blood gene expression data as well as the Single Nucleotide Polymorphism (SNP) data in building ML models for AD classification. We present a systematic approach on feature selection and ML model design using both types of genetic data provided by the Alzheimer's Disease Neuroimaging Initiatives (ADNI). Our two-step feature selection produced a curated list of important genes. In addition to these selected genetic features, to examine the role of non-genetic covariates, we included age and number of education years (EDU) as extra features. In the Control (CN) vs. AD classification, the best performing classifier, XGBoost, trained with gene expression features only and that with extra features included had Area Under Curve (AUC) of 0.64 and 0.65 respectively. However, AUC for the same task using SNP data only and that with extra features included was 0.56 and 0.64 respectively. The just above chance results of classifier trained with SNP features and the improvement when used along with additional covariates indicate low potential of SNP data in AD classification when used alone while also indicating the importance of non-genetic factors associated with AD. Nevertheless, with well above chance performance, gene expression features show great potential especially between groups of AD progression, i.e., CN vs. AD, CN vs. EMCI, EMCI vs. AD and LMCI vs. AD. The source code and manual are available at https://github.com/mvrl/ADNI_Genetics.**

***Index Terms*—Alzheimer's disease, Genetics, Machine Learning, Feature Selection**

I. Introduction

Alzheimer's disease (AD) is the most common form of dementia worldwide. Biologically, AD is defined as the pathological deposition of folded beta-amyloid plagues, and hyperphosphorylated neurofibrillary tau tangles in the brain leading to neurodegeneration [1]. Identification of novel bio-markers for Alzheimer's disease is still an ongoing research [2]–[6].

It has been well established from earlier statistical studies that AD has notable genetic associations [7]–[9]. Encouraged by the discoveries using big data and statistical methods, there has been ongoing efforts in building precise AD predictive tools using Machine Learning (ML) models [10]–[12]. One of the fundamental challenges while utilizing genetic data to build ML classifier is the small sample size (number of subjects) and the large dimension of genetic features [13]. Low sample size coupled with high feature dimension contribute to easy over-fitting of the model, causing poor testing performance [14]. ML models trained on a smaller number of most discriminative features can minimize this problem of "curse-of-dimensionality" and generalize better on the test set [15]. Therefore, to ensure the selection of the most important features, proper care should be taken in this initial step of feature selection.

There is a rich literature on dimensionality reduction techniques (DRT) for high dimensional data. These DRTs can be of different types based on their learning objectives. For example, Principle Component Analysis (PCA) is designed to preserve variance while t-Stochastic neighbor embedding (t-SNE) is designed with the objective of neighborhood preservation. Refer to Ayesha et al. [16] for a detailed review of different DRTs. The primary objective of this study is to fairly assess the potential of genetic data in building ML-based AD classification models while also estimating the contributions of important genes in AD classification. Therefore, instead of using DRT techniques that transform the input features into low dimensional space as done in the recent work [17], we choose to use feature selection strategy which selects the most important features to train the classifiers on. Accordingly, in this work, we adopt a two-step feature selection approach to identify important genetic features. The selected features are then used to train and test a classifier called XGBoost [18]. Also, based on their importance in making classification, the ML classifier produces ranking of the selected genetic features, which can be considered as the potential AD bio-markers.

The main contributions of our study are:

1) To assess the predictive potential of the two forms of genetic data: gene expression and SNP variants, we present a systematic strategy of feature selection and ML model design.
2) We identify important genes for classifying different stages of AD using gene expression data, as well as important SNPs in classifying AD from CN.
3) We highlight the importance of fair feature selection by demonstrating the inflation of model's performance, due to the data leakage issue observed in some of the recent studies.

II. Background

Genetic data used in AD studies to train and test ML models are usually of two forms: gene expression and Single Nucleotide Polymorphism (SNP) variants. While the gene expression data quantifies genome-wide gene expression [19], the SNP data captures the variations found in the nucleotide pairs in the DNA sequences of the subjects [20]. Refer to Mishra and Li [21] for the methods and results reported in literature related to application of ML in genetic study of AD.

A. Gene Expression Data

Gene expression data can be acquired from the specific brain regions or from peripheral blood of subjects. Data from the brain regions help in identifying the deferentially expressed genes in the affected brain cells. ML approaches have been developed using the expression dataset from different brain regions, either to identify genetic bio-markers or for drug re-purposing in AD [22]–[24]. Even though brain region's gene expression data allows direct access to the site of interest and elucidates the role of candidate genes, data acquisition from brain regions is invasive and usually done during postmortem of the subject's brain. As an effective alternative, in the recent years, peripheral blood has been considered as a suitable candidate tissue to acquire gene expression data from. Blood Gene expression data provided by the ADNI database has been used to predict clinical dementia rating [25]. Sethi and Ni [26] have used the blood gene expression data from the ADNI database to predict AD and the highly ranked genes were used to perform various post analyses. Moreover, gene expression data can be integrated with other kinds of data such as DNA methylation to better predict AD [27]. In [17], along with testing multiple feature selection methods, the study trained and tested classifiers using three public datasets. Unlike [17] that uses Variational Auto-Encoder (VAE) for dimensionality reduction, we adopted a rather simpler approach based on statistical tests and model-based feature selection strategy. We achieve similar performance while also producing an easily interpretable curated list of important genes associated with the AD at different stages of progression.

B. SNP Data

Well trained ML models using SNP data may identify potentially novel SNP bio-markers [28]. In the literature, many feature selection and ML approaches have been proposed [29]–[31]. A rich class of feature selection strategies, ML models, evaluation metrics, and result interpretation can be read in a recent review by Nicholls et al. [32].

SNP data, once integrated with phenotypic data such as Magnetic Resonance Imaging (MRI), or other forms of genetic data, may provide more precise results. A deep learning-based approach [33] uses both MRI and genetics data to shed light on the link between SNPs and different brain regions. Transcriptome-Wide Association Study (TWAS), which is a test for correlation between predicted gene expression and traits from Genome Wide Association Study (GWAS) summary data, was recently proposed [34]. Except for the choice of second stage feature selection strategy and the type of classifier used, our overall approach of SNP variants based ML model design is similar to [31]. However, unlike [31], which aims to predict duloxetine response in Major Depressive Disorder (MDD) using SNP data, our goal is to build an AD classifier.

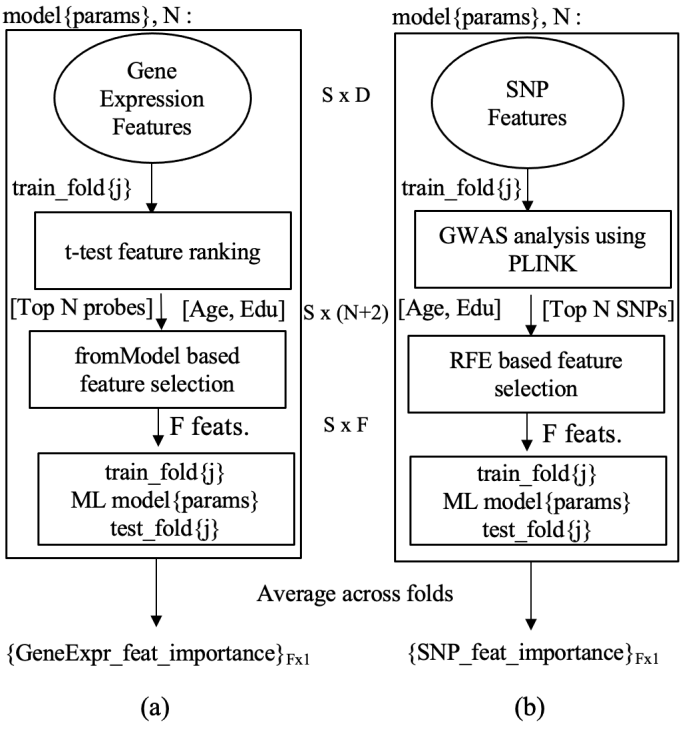

Fig. 1. Methodology of Training and Testing the models for each fold in 5-fold CV using gene expression (a) and SNP (b) features. Although trained independently, for the simplicity of demonstration in this figure, same symbol is used to represent number of samples in training set (S), dimension of gene expression or SNP variants (D), dimension of features after first stage selection (N) and dimension of features after second stage feature selection (F)

From the methods in the recent works [25], [26], we noticed that feature selection, i.e. the selection of genes, on which ML models were later trained on, was often performed using the whole data before the train/test split. This causes the issue of test set information leakage [35] leading to the inflation of testing performance. This performance inflation is demonstrated in Figures 3 and 4 in the Results section.

C. ADNI

One of the most successful initiatives focused on AD research is the Alzheimer's disease Neuroimaging Initiatives (ADNI) [36]. This initiative offers freely accessible Neuroimaging, Neuropsychological, as well as Genetics data for healthy controls and subjects with different stages of AD. The ADNI project has been continuing with its three phases so far: ADNI1, ADNI2/GO, and ADNI3. In this study we focused on two types of genetic data: Gene Expression and SNP variants for GWAS analysis. Gene expression used in our study were produced using the Affymetrix Human Genome U 219 arrays on peripheral blood samples. After pre-processing and Quality

Control (QC) steps ADNI offers gene expression profile of 49,386 probes for 743 subjects. The label distribution of these 743 subjects is: {'CN':260,'EMCI':215, LMCI: 225,'AD':43}. For SNP data, we merge the SNP data from all three phases of ADNI. We choose to focus only on two groups of subjects: CN and AD. Therefore, after merging and QC step during GWAS analysis [37] the label distribution of our overall SNP data is: {'CN': 324, 'AD': 195}. The average count of SNP features across training sets is 294,221. For both gene expression and SNP data, number of subjects with AD is noticeably low as compared to other groups.

III. Methodology

Due to the potential problem of AD data mentioned above, we evaluate ML models for AD classification using stratified 5-fold cross-validation (CV) strategy without a separate held out test set. Instead of adopting the standard approaches of model training, which choose hyper-parameters (parameters of classifiers and the number of input features, etc.) based on the model's performance on the validation set and evaluate the model on a held-out test set, we fix the number of input features to be selected after the first step of feature selection and the classification model. Therefore, for each CV fold, a two-step feature selection is carried on the training set and the testing is done on the validation set. We then report the averaged 5-fold CV performance for the best performing model as the final performance evaluation result.

In this project, we experiment with different tree-based ML classifiers, such as Random Forrest, Gradient Boosting, and XGBoost. Among those, XGBoost constantly achieves the best performance across different classifiers, and hence is chosen as the base classifier in our work.

A. Model Training on Gene Expression Data

For Gene expression data, t-test is carried out between the groups-of-interest, and genes are ranked based on the ascending order of False Discovery Rates (FDR) or the corrected p-values. This is called the first feature selection step.

Using the top N genes selected in the first step as input, we use SelectfromModel as the second feature selection step. For a given model, the SelectfromModel function, as implemented in scikit-learn [38], selects a set of features based on their importance in making prediction. We choose the default parameter settings as in scikit-learn for this selection.

In addition to the genetic features, we are also interested in examining the role of easily available non-genetic features in AD prediction. Earlier studies have shown the association of age [39] and years of education [40] with AD. Therefore, we include age and years of education in the feature space. These two features act as covariates and may improve the performance of ML models. Finally, the features preserved after the second feature selection step are used to train and test the XGBoost classifier. This overall methodology is applied to different combination of model parameters systematically.

B. Model Training on SNP Data

For SNP data, the same procedure is carried out except for the feature selection steps. As the first feature selection step, GWAS is used on the merged ADNI1/2/GO/3 phase data. Standard procedures such as quality control, control of population stratification (using five principal components as covariates), and logistic regression-based association analysis are performed using PLINK [41]. All of these steps, along with the associated hyper-parameters necessary, are implemented by following the tutorial in [37]. Finally, a list of SNPs ranked based on p-value of significance is generated using GWAS.

Top N SNPs are extracted and expanded in one-hot code manner. Among the expanded SNP features, the features coded as missing alleles are dropped. This one-hot representation later allows us to observe not only the SNPs with high importance but also the minor/major allele pair associated with them. Owing to its superior performance than the SelectfromModel feature selector used for gene expression data, we choose Recursive Feature Elimination (RFE) on model training using SNP data. In RFE, for each recursion, the least important features are eliminated leaving only a set of optimum features.

C. Data Imbalance Problem

Large imbalance in the class distribution causes the model to be unfairly biased towards the majority class [26]. To minimize the effect of class imbalance, a synthetic data points interpolation technique called SMOTE [42] has been developed to oversample the minority classes to create a balanced augmented dataset. However, if this over-sampling is done before train-test split, it may lead to the issue of data leakage problem causing unfair inflation of test performance, which can be also observed in the near perfect performance of 0.99 as reported in Sethi and Ni [26]. Therefore, we first split the dataset into train and test set and then use SMOTE only on the train set. Moreover, as fairer metrics for imbalanced dataset, we choose macro-Accuracy and macro-AUC as the evaluation metrics.

IV. Results

A. ML Model Trained using Gene Expression Data

Table I shows the best 5-fold CV performance for different binary groups of AD progression. To make a fair comparison between the classifiers trained on a set of features with and without the second step feature selection, separate grid search of hyper-parameters is performed. The hyper-parameter space for n_estimators is the range $(10, 2 * N, 50)$ and that for max_depths is the range(2,10,2). N is selected from the list $[25, 50, 100, 200, 300, 400, 500]$.

B. ML Model Trained using SNP Data

Manhattan plots in Figure 2 reveal a few SNPs with high association with AD. The top ranked SNP being rs2075650 near gene TOMM40 in chromosome-19. TOMM40, which is believed to be in a linkage disequilibrium (LD) block with the well-known APOE gene, has also been reported as the high-risk gene for the late-onset AD [43].

TABLE I
BEST 5-FOLD CV PERFORMANCE FOR GENE EXPRESSION + [EDU+AGE] FEATURES

Groups	t-test based features selection				t-test +FromModel based feature selection				
	N+2	params*	ACC	AUC	N+2	F	params*	ACC	AUC
CN_AD	27	(10, 2)	0.61	0.63	52	16	(60, 2)	0.6	0.65
CN_EMCI	52	(10, 6)	0.62	0.63	202	64	(160, 8)	0.6	0.63
CN_LMCI	52	(60, 6)	0.55	0.56	27	10	(10, 6)	0.57	0.55
EMCI_LMCI	202	(60, 2)	0.51	0.5	102	41	(60, 6)	0.54	0.51
EMCI_AD	27	(10, 4)	0.65	0.69	52	17	(60, 4)	0.61	0.71
LMCI_AD	27	(10, 4)	0.61	0.67	302	54	(10, 8)	0.59	0.63

* Parameters of the best model (n_estimators, max_depths)

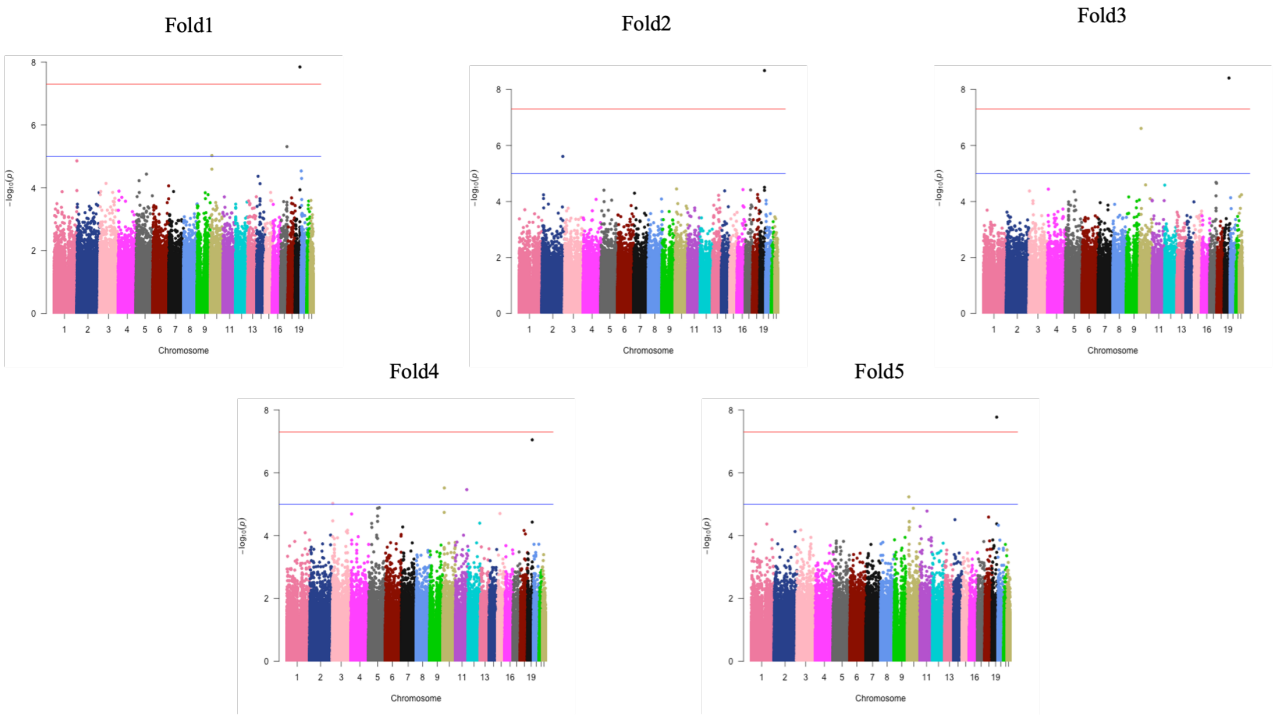

Fig. 2. Manhattan Plot for CN vs. AD GWAS for training data of each CV fold

Table IV presents 5-fold CV performance of the best performing classifier trained with SNP data. As seen for gene expression data, the best performance is achieved for only small number of SNP features. This suggests that there are only a few SNPs that can be considered as having high potential for AD classification. This further also suggests that the GWAS summary-based feature selection is a sufficient strategy when it comes to selecting features to train a ML classifier.

V. DISCUSSION

One of the observations from Table I and III is that the best performance is achieved with relatively small number of features. This suggests that there are only few gene expression features with high association with the AD. This further means that the second feature selection step did not have chance to bring in significant gains as there were only few important genes already selected by t-test based first feature selection step. The best AUC of 0.71 is achieved for the EMCI versus AD group indicating the high potential of gene expression features to classify between these groups. Despite being trained on relatively balanced groups: CN vs. LMCI and EMCI vs. LMCI, the performance of the model is sub-optimal suggesting low potential of the gene expression data in distinguishing between these stages of AD progression. For most of the groups, ACC of the best model trained only using t-test based feature selection is better than that trained using two-stage feature selection strategy. For this reason, we used the best models trained using only t-test based feature selection to produce importance based ranking of features as presented in Table II.

As mentioned before, in addition to the gene expression features; age and education were included in the input feature list. Consistent with the AD research literature, age [39] and years of education [40] were the top ranked features in all the groups. However, to understand how much contribution these non-genetic features had in the performance of the model, we build classifiers using only the gene expression features. The

TABLE II
TOP GENE EXPRESSION FEATURES + [EDU,AGE] RANKED BASED ON AVERAGE FEATURE IMPORTANCE (GREATER THAN 0.01) ACROSS 5 CROSS-VALIDATION FOLDS PRODUCED BY THE BEST PREFORMING MODEL TRAINED WITH ONLY T-TEST BASED FEATURE SELECTION STRATEGY

CN_AD		CN_EMCI		EMCI_AD		LMCI_AD	
features	importance	features	importance	features	importance	features	importance
AGE*	0.0504	AGE*	0.0646	XRCC5	0.0528	AGE*	0.0497
EDU*	0.0443	EDU*	0.0404	AGE*	0.0498	ARMCX5	0.0390
probe(11762535)	0.0366	TDRD1	0.0276	RAB27A	0.0371	EDU*	0.0371
TRIM10	0.0349	CLDND1	0.0172	PIP4K2A	0.0351	GCOM1	0.0343
PIGZ	0.0320	ZSCAN5A	0.0124	PSMF1	0.0332	ZNF428	0.0326
LYSMD1	0.0286	MEFV	0.0110	EDU*	0.0315	ZNF608	0.0322
SUMF1	0.0277			CCDC176	0.0271	SUMF1	0.0262
MT1X	0.0257			BNIP3L	0.0269	METTL7A	0.0261
ARHGEF12	0.0238			OSBP2	0.0250	DRP2	0.0210
CXCL14	0.0235			TTBK2	0.0248	AFFX-PheX-M	0.0206
NOXO1	0.0234			IRAK3	0.0241	USP12	0.0192
METTL7A	0.0228			KAZN	0.0171	FBXO28	0.0186
PIP4K2A	0.0227			NEFH	0.0157	ANXA3	0.0180
TDP2	0.0219			NEK1	0.0153	AFFX-r2-TagO-3	0.0180
OSBPL1A	0.0210			ZNF428	0.0152	MLF1	0.0174
ZNF326	0.0209			HS3ST2	0.0152	AFFX-r2-TagF	0.0172
PRR5	0.0208			probe(11763149)	0.0149	AFFX-r2-TagB	0.0154
SFRP1	0.0206			SPATA5	0.0148	UBE2G2	0.0148
ABHD12B	0.0203			RAP1GAP2	0.0143	C5ORF34	0.0146
IHH	0.0203			EDNRA	0.0119	SNORD16	0.0143
ACOT11	0.0194			PLAC8	0.0118	probe(11762906)	0.0140
FASTKD5	0.0190			MGAM	0.0110	CPLX2	0.0139
HLA-DOA	0.0181			AFFX-r2-TagJ-5	0.0106	CALR	0.0138
OSBPL8	0.0165			LGALS3	0.0103	MB21D1	0.0129
CKM	0.0152			DNAJB2	0.0102	FKRP	0.0115
FAM179B	0.0149			WDR59	0.0102	IPP	0.0111
SMIM5	0.0146			AFFX-r2-TagF	0.0101	ZNF83	0.0110
SESN3	0.0141					DTD1	0.0106
C19ORF77	0.0138					IL20	0.0106
VEGFA	0.0131					EIF4H	0.0106
FBXO38	0.0117					AFFX-Nonspecific-GC13	0.0105
probe(11763125)	0.0117					UPK3BL	0.0102
EIF2AK1	0.0116					SMIM5	0.0101
SULT1B1	0.0115						
CEP170B	0.0110						
AFFX-r2-TagF	0.0109						
ZNF271	0.0107						

* Non-genetic extra feature

TABLE III
BEST 5-FOLD CV PERFORMANCE USING ONLY GENE EXPRESSION FEATURES

Group	N	selection	F	params	ACC	AUC
CN_AD	25	ttest	25	(10, 2)	0.60	0.64
CN_EMCI	500	ttest	500	(10, 6)	0.54	0.55
CN_LMCI	50	ttest	50	(10, 8)	0.53	0.53
EMCI_LMCI	200	both*	65	(60, 2)	0.55	0.53
EMCI_AD	25	ttest	25	(10, 8)	0.65	0.67
LMCI_AD	500	both*	25	(10, 2)	0.59	0.61

* ttest + fromModel based feature selection

results presented in Table III do show drop in performance indicating high association with AGE and EDU. However, for some groups of AD progression: CN vs. AD, EMCI vs. AD and LMCI vs. AD, the performance still remains well above chance indicating good predictability of gene expression features in classifying between these groups. It should be noted that the performance of our best performing model using only gene expression data, classifying CN vs. AD has AUC of 0.64 which is comparable to 0.657 as reported in [17]. However, unlike [17] which focuses only on CN vs. AD classification, we train separate models to classify between different stages of AD progression. Moreover, with simple t-test and importance based feature ranking our curated list is more interpretable with availability of importance score that determines the contribution of each feature in making a prediction.

For each binary classification group with the best performance above 0.60, Table II presents the most important gene expression features. Most of the genes corresponding to the features listed in Table II have been mentioned in literature focusing on genetic association with AD. However, as observed in Table I and Table III, the best performance was seen for classification between EMCI vs. AD. Among the top ranked genes for EMCI vs AD classification with feature importance > 0.02, we cite one representative publication discussing the association of the respective gene with AD for each of the selected genes as follows: XRCC5 [44], RAB27A [45], PIP4K2A [46], PSMF1 [47], CCDC176 [48], BNIP3L [49], OSBP2 [50], TTBK2 [51], and IRAK3 [52].

TABLE IV
BEST 5-FOLD CV PERFORMANCE FOR SNP + [EDU+AGE] FEATURES

GWAS summary-based features selection					RFE based features selection			
N+2	N*	params	ACC	AUC	F	params	ACC	AUC
1002	2814	(100, 6)	0.56	0.59	1407	(200, 6)	0.56	0.59
752	2110	(100, 2)	0.57	0.6	1053	(100, 2)	0.57	0.6
502	1413	(200, 8)	0.57	0.61	706	(300, 8)	0.57	0.61
302	844	(300, 2)	0.57	0.6	421	(300, 2)	0.58	0.6
202	562	(400, 2)	0.58	0.59	281	(400, 2)	0.58	0.59
102	283	(100, 2)	0.57	0.63	141	(100, 2)	0.57	0.63
52	141	(100, 2)	0.59	0.63	70	(100, 2)	0.58	0.64
27	73	(50, 8)	0.62	0.63	36	(60, 8)	0.61	0.63

* Feature dimension after one-hot encoding and dropping missing alleles representation

TABLE V
TOP SNP FEATURES + [EDU,AGE] RANKED BASED ON AVERAGE FEATURE IMPORTANCE (GREATER THAN 0.01) ACROSS 5 CROSS-VALIDATION FOLDS PRODUCED BY THE BEST PREFORMING MODEL TRAINED WITH ONLY GWAS ANALYSIS BASED FEATURE SELECTION

features	importance	Nearest Gene
19_rs2075650_AA	0.0517	TOMM40
10_rs11253696_AA	0.0263	
20_rs6116375_CC	0.0252	
EDU*	0.0201	
10_rs11253696_GG	0.0172	
AGE*	0.0164	
18_rs618236_AA	0.0145	RIT2
5_rs168825_GG	0.0129	PDZD2
14_rs11156803_TT	0.0118	NPAS3
22_rs469120_TC	0.0110	
11_rs12804305_TC	0.0109	
19_rs157580_AA	0.0109	TOMM40
5_rs168825_TT	0.0105	PDZD2
14_rs10151748_TT	0.0104	

* Non-genetic extra feature
feature: CHRno_rsSNPid_minorAllelemajorAllele

TABLE VI
BEST 5-FOLD CV PERFORMANCE USING ONLY SNP FEATURES WITH ONLY GWAS SUMMARY BASED FEATURE SELECTION

N	N'	params	ACC	AUC
500	705	(300, 6)	0.56	0.56
400	562	(100, 4)	0.54	0.53
300	421	(200, 8)	0.52	0.54
200	279	(400, 8)	0.53	0.53
100	140	(200, 6)	0.54	0.55
50	69	(100, 8)	0.55	0.55
25	35	(10, 2)	0.55	0.56

Detailed study of these genes and other genes with high importance in classification between other groups is left for our future work.

As seen with models trained on gene expression features, AGE and EDU appear as the top features when included with a list of GWAS selected SNP features as well. For the nearest gene associated with the top ranked SNP features as listed in V we cite a representative publication discussing about the association of the respective gene and AD as follows: TOMM40 [43], RIT2 [53], PDZD2 [54], and NPAS3 [55].

Data leakage is the problem where information of the test set is directly/indirectly accessible to the model during its

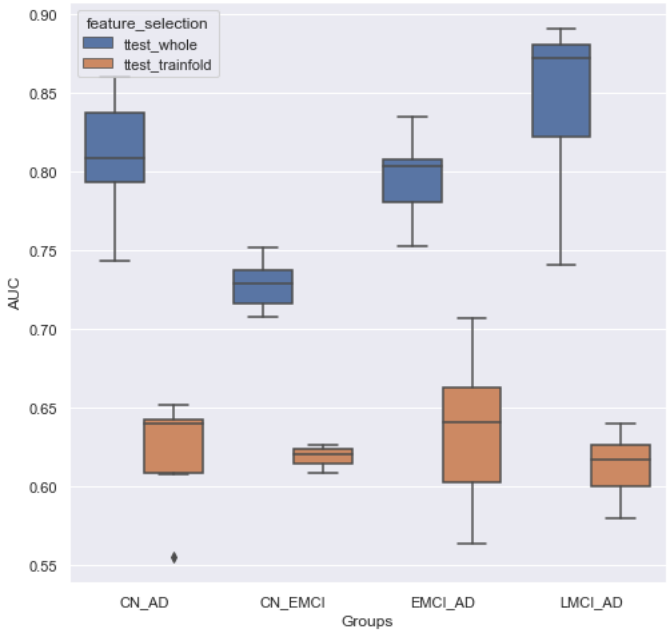

Fig. 3. Box plot for best AUCs of models trained with different values of N after performing ttest based feature selection on whole gene expression data (causing test-set information leakage) vs. only on train set of each CV fold

training. In the case of feature selection, if the feature selection is done on the whole data before splitting the data into train and test set, this feature ranking will also be based on the information in the portion of data which later will be used as a test set to evaluate the model on. This will cause unfair inflation of testing performance. A recent paper [35] has also pointed this issue in earlier studies. As evident in box plots of model performance in figures 3 and 4 for gene expression and SNP data respectively, incorrect way of selecting features causes large inflation of performance leading to probably misleading results. Moreover, we would like to stress that data leakage issue can also occur if the features selected by prior studies using same data are directly used to train ML models for new studies. Data leakage issue therefore can be mitigated by selecting the features based on own training set for each new study. This hinders the advancement of bio-markers discovery research. Therefore, for fair evaluation

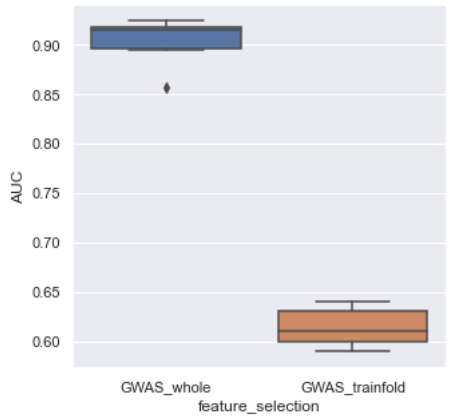

Fig. 4. Box plot for best AUCs of models trained with different values of N after performing GWAS based feature selection on whole SNP data (causing test-set information leakage) vs. only on train set of each CV fold

of models and discovering truly important genetic features associated with a disease, cross-dataset validation as in [17] is an important research direction to move forward.

VI. Conclusion

In this study, we present a systematic approach of feature selection and ML model design using genetic data for AD classification. From the results, it is evident that genetic data, especially gene expressions, have a lot of potential in classifying certain stages of AD. SNP data, however, if used alone had just above chance performance.

The number of genetic features yielding the best performance is low. This allows researchers to focus on a small list of candidate genes with respect to the AD progression. Most of the genes associated with the features ranked by the classifiers in this study are found to have association with AD as reported in previous literature. From a broader perspective, this study may inspire researchers towards considering genetic information as a valuable data within the broader framework of Multi-Modal ML for AD research.

VII. Acknowledgment

This work was funded by the National Institutes of Health (NIH). Grant ID: $R01AG054459$.

References

[1] Anil Kumar, Arti Singh, et al. A review on alzheimer's disease pathophysiology and its management: an update. *Pharmacological reports*, 67(2):195–203, 2015.

[2] Mark S Henry, Anthony P Passmore, Stephen Todd, Bernadette McGuinness, David Craig, and Janet A Johnston. The development of effective biomarkers for alzheimer's disease: a review. *International journal of geriatric psychiatry*, 28(4):331–340, 2013.

[3] Kim Henriksen, Sid E O'Bryant, Harald Hampel, John Q Trojanowski, Thomas J Montine, Andreas Jeromin, Kaj Blennow, Anders Lönneborg, Tony Wyss-Coray, Holly Soares, et al. The future of blood-based biomarkers for alzheimer's disease. *Alzheimer's & Dementia*, 10(1):115–131, 2014.

[4] Bob Olsson, Ronald Lautner, Ulf Andreasson, Annika Öhrfelt, Erik Portelius, Maria Bjerke, Mikko Hölttä, Christoffer Rosén, Caroline Olsson, Gabrielle Strobel, et al. Csf and blood biomarkers for the diagnosis of alzheimer's disease: a systematic review and meta-analysis. *The Lancet Neurology*, 15(7):673–684, 2016.

[5] Giovanni B Frisoni, Marina Boccardi, Frederik Barkhof, Kaj Blennow, Stefano Cappa, Konstantinos Chiotis, Jean-Francois Démonet, Valentina Garibotto, Panteleimon Giannakopoulos, Anton Gietl, et al. Strategic roadmap for an early diagnosis of alzheimer's disease based on biomarkers. *The Lancet Neurology*, 16(8):661–676, 2017.

[6] Samantha Swarbrick, Nick Wragg, Sourav Ghosh, and Alexandra Stolzing. Systematic review of mirna as biomarkers in alzheimer's disease. *Molecular neurobiology*, 56(9):6156–6167, 2019.

[7] Jungsu Kim, Jacob M Basak, and David M Holtzman. The role of apolipoprotein e in alzheimer's disease. *Neuron*, 63(3):287–303, 2009.

[8] Laura Spinney. Alzheimer's disease: The forgetting gene. *Nature News*, 510(7503):26, 2014.

[9] Karolien Bettens, Kristel Sleegers, and Christine Van Broeckhoven. Genetic insights in alzheimer's disease. *The Lancet Neurology*, 12(1):92–104, 2013.

[10] M Tanveer, B Richhariya, RU Khan, AH Rashid, P Khanna, M Prasad, and CT Lin. Machine learning techniques for the diagnosis of alzheimer's disease: A review. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1s):1–35, 2020.

[11] Yudong Zhang, Zhengchao Dong, Preetha Phillips, Shuihua Wang, Genlin Ji, Jiquan Yang, and Ti-Fei Yuan. Detection of subjects and brain regions related to alzheimer's disease using 3d mri scans based on eigenbrain and machine learning. *Frontiers in computational neuroscience*, 9:66, 2015.

[12] Charles K Fisher, Aaron M Smith, and Jonathan R Walsh. Machine learning for comprehensive forecasting of alzheimer's disease progression. *Scientific reports*, 9(1):1–14, 2019.

[13] Papia Ray, S Surender Reddy, and Tuhina Banerjee. Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*, pages 1–43, 2021.

[14] Amrita Chattopadhyay and Tzu-Pin Lu. Gene-gene interaction: the curse of dimensionality. *Annals of translational medicine*, 7(24), 2019.

[15] Alison A Motsinger and Marylyn D Ritchie. Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. *Human genomics*, 2(5):1–11, 2006.

[16] Shaeela Ayesha, Muhammad Kashif Hanif, and Ramzan Talib. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59:44–58, 2020.

[17] Taesic Lee and Hyunju Lee. Prediction of alzheimer's disease using blood gene expression data. *Scientific reports*, 10(1):1–13, 2020.

[18] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[19] Alvis Brazma and Jaak Vilo. Gene expression data analysis. *FEBS letters*, 480(1):17–24, 2000.

[20] Ann-Christine Syvänen. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, 2(12):930–942, 2001.

[21] Rohan Mishra and Bin Li. The application of artificial intelligence in the genetic study of alzheimer's disease. *Aging and disease*, 11(6):1567, 2020.

[22] P Roy Walker, Brandon Smith, Qing Yan Liu, A Fazel Famili, Julio J Valdés, Ziying Liu, and Boleslaw Lach. Data mining of gene expression changes in alzheimer brain. *Artificial intelligence in medicine*, 31(2):137–154, 2004.

[23] Abhibhav Sharma and Pinki Dey. A machine learning approach to unmask novel gene signatures and prediction of alzheimer's disease within different brain regions. *Genomics*, 113(4):1778–1789, 2021.

[24] Steve Rodriguez, Clemens Hug, Petar Todorov, Nienke Moret, Sarah A Boswell, Kyle Evans, George Zhou, Nathan T Johnson, Bradley T Hyman, Peter K Sorger, et al. Machine learning identifies candidates for drug repurposing in alzheimer's disease. *Nature communications*, 12(1):1–13, 2021.

[25] Justin B Miller and John SK Kauwe. Predicting clinical dementia rating using blood rna levels. *Genes*, 11(6):706, 2020.

[26] Amish Sethi and Andrew Warren Ni. Functional genetic biomarkers of alzheimer's disease and gene expression from peripheral blood. *bioRxiv*, 2021.

[27] Chihyun Park, Jihwan Ha, and Sanghyun Park. Prediction of alzheimer's disease based on deep neural network by integrating gene expression and dna methylation dataset. *Expert Systems with Applications*, 140:112873, 2020.

[28] Uday Rangaswamy, S Akila Parvathy Dharshini, Dhanusha Yesudhas, and M Michael Gromiha. Vepad-predicting the effect of variants associated with alzheimer's disease using machine learning. *Computers in Biology and Medicine*, 124:103933, 2020.

[29] Brissa-Lizbeth Romero-Rosales, Jose-Gerardo Tamez-Pena, Humberto Nicolini, Maria-Guadalupe Moreno-Treviño, and Victor Trevino. Improving predictive models for alzheimer's disease using gwas data by incorporating misclassified samples modeling. *PloS one*, 15(4):e0232103, 2020.

[30] Thanh-Tung Nguyen, Joshua Zhexue Huang, Qingyao Wu, Thuy Thi Nguyen, and Mark Junjie Li. Genome-wide association data classification and snps selection using two-stage quality-based random forests. In *BMC genomics*, volume 16, pages 1–11. Springer, 2015.

[31] Malgorzata Maciukiewicz, Victoria S Marshe, Anne-Christin Hauschild, Jane A Foster, Susan Rotzinger, James L Kennedy, Sidney H Kennedy, Daniel J Müller, and Joseph Geraci. Gwas-based machine learning approach to predict duloxetine response in major depressive disorder. *Journal of psychiatric research*, 99:62–68, 2018.

[32] Hannah L Nicholls, Christopher R John, David S Watson, Patricia B Munroe, Michael R Barnes, and Claudia P Cabrera. Reaching the end-game for gwas: machine learning approaches for the prioritization of complex disease loci. *Frontiers in genetics*, 11:350, 2020.

[33] Dan Pan, Yin Huang, An Zeng, Longfei Jia, Xiaowei Song, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. Early diagnosis of alzheimer's disease based on deep learning and gwas. In *International Workshop on Human Brain and Artificial Intelligence*, pages 52–68. Springer, 2019.

[34] Sicheng Hao, Rui Wang, Yu Zhang, and Hui Zhan. Prediction of alzheimer's disease-associated genes by integration of gwas summary data and expression data. *Frontiers in genetics*, 9:653, 2019.

[35] Miseon Shim, Seung-Hwan Lee, and Han-Jeong Hwang. Inflated prediction accuracy of neuropsychiatric biomarkers caused by data leakage in feature selection. *Scientific Reports*, 11(1):1–7, 2021.

[36] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

[37] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, 27(2):e1608, 2018.

[38] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, et al. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013.

[39] Rita Guerreiro and Jose Bras. The age factor in alzheimer's disease. *Genome medicine*, 7(1):1–3, 2015.

[40] Chengxuan Qiu, Lars Bäckman, Bengt Winblad, Hedda Agüero-Torres, and Laura Fratiglioni. The influence of education on clinically diagnosed dementia incidence and mortality data from the kungsholmen project. *Archives of neurology*, 58(12):2034–2039, 2001.

[41] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[42] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[43] Carlos Cruchaga, Petra Nowotny, John SK Kauwe, Perry G Ridge, Kevin Mayo, Sarah Bertelsen, Anthony Hinrichs, Anne M Fagan, David M Holtzman, John C Morris, et al. Association and expression analyses with single-nucleotide polymorphisms in tomm40 in alzheimer disease. *Archives of neurology*, 68(8):1013–1019, 2011.

[44] Zixin Hu, Rong Jiao, Panpan Wang, Yun Zhu, Jinying Zhao, Phil De Jager, David A Bennett, Li Jin, and Momiao Xiong. Shared causal paths underlying alzheimer's dementia and type 2 diabetes. *Scientific reports*, 10(1):1–15, 2020.

[45] Xian Zhang, Timothy Y Huang, Joel Yancey, Hong Luo, and Yun-wu Zhang. Role of rab gtpases in alzheimer's disease. *ACS chemical neuroscience*, 10(2):828–838, 2018.

[46] Padinjat Raghu, Annu Joseph, Harini Krishnan, Pramod Singh, and Sankhanil Saha. Phosphoinositides: regulators of nervous system function in health and disease. *Frontiers in molecular neuroscience*, 12:208, 2019.

[47] Linhai Zhao, Zongxiao He, Di Zhang, Gao T Wang, Alan E Renton, Badri N Vardarajan, Michael Nothnagel, Alison M Goate, Richard Mayeux, and Suzanne M Leal. A rare variant nonparametric linkage method for nuclear and extended pedigrees with application to late-onset alzheimer disease via wgs data. *The American Journal of Human Genetics*, 105(4):822–835, 2019.

[48] Angela K Hodges, Thomas M Piers, David Collier, Oliver Cousins, and Jennifer M Pocock. Pathways linking alzheimer's disease risk genes expressed highly in microglia. *Neuroimmunology and Neuroinflammation*, 8, 2021.

[49] Jesse S Kerr, Bryan A Adriaanse, Nigel H Greig, Mark P Mattson, M Zameel Cader, Vilhelm A Bohr, and Evandro F Fang. Mitophagy and alzheimer's disease: cellular and molecular mechanisms. *Trends in neurosciences*, 40(3):151–166, 2017.

[50] David Cohen, Alexander Pilozzi, and Xudong Huang. Network medicine approach for analysis of alzheimer's disease gene expression data. *International journal of molecular sciences*, 21(1):332, 2020.

[51] Jung-Chi Liao, T Tony Yang, Rueyhung Roc Weng, Ching-Te Kuo, and Chih-Wei Chang. Ttbk2: a tau protein kinase beyond tau phosphorylation. *BioMed research international*, 2015, 2015.

[52] Kwangsik Nho, Kelly Nudelman, Mariet Allen, Angela Hodges, Sungeun Kim, Shannon L Risacher, Liana G Apostolova, Kuang Lin, Katie Lunnon, Xue Wang, et al. Genome-wide transcriptome analysis identifies novel dysregulated genes implicated in alzheimer's pathology. *Alzheimer's & Dementia*, 16(9):1213–1223, 2020.

[53] Pratip Rana, Edian F Franco, Yug Rao, Khajamoinuddin Syed, Debmalya Barh, Vasco Azevedo, Rommel TJ Ramos, and Preetam Ghosh. Evaluation of the common molecular basis in alzheimer's and parkinson's diseases. *International journal of molecular sciences*, 20(15):3730, 2019.

[54] Alison I Bernstein, Yunting Lin, R Craig Street, Li Lin, Qing Dai, Li Yu, Han Bao, Marla Gearing, James J Lah, Peter T Nelson, et al. 5-hydroxymethylation-associated epigenetic modifiers of alzheimer's disease modulate tau-induced neurotoxicity. *Human molecular genetics*, 25(12):2437–2450, 2016.

[55] Charalampos S Floudas, Nara Um, M Ilyas Kamboh, Michael M Barmada, and Shyam Visweswaran. Identifying genetic interactions associated with late-onset alzheimer's disease. *BioData mining*, 7(1):1–19, 2014.