A Machine Learning Approach to Unmask Novel Gene Signatures and Prediction of Alzheimer's Disease Within Different Brain Regions.

Abhibhav Sharma[1], Pinki Dey[2*]

1. School of Computer and System Sciences, Jawaharlal Nehru University, New Delhi110067, India
2. School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney 2033, Australia

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder whose aetiology is currently unknown. Although numerous studies have attempted to identify the genetic risk factor(s) of AD, the interpretability and/or the prediction accuracies achieved by these studies remained unsatisfactory, reducing their clinical significance. Here, we employ the ensemble of random-forest and regularized regression model (LASSO) to the AD-associated microarray datasets from four brain regions - Prefrontal cortex, Middle temporal gyrus, Hippocampus, and Entorhinal cortex- to discover novel genetic biomarkers through a machine learning-based feature-selection classification scheme. The proposed scheme unrevealed the most optimum and biologically significant classifiers within each brain region, which achieved by far the highest prediction accuracy of AD in 5-fold cross-validation (99% average). Interestingly, along with the novel and prominent biomarkers including CORO1C, SLC25A46, RAE1, ANKIB1, CRLF3, PDYN, numerous non-coding RNA genes were also observed as discriminator, of which AK057435 and BC037880 are uncharacterized long non-coding RNA genes.

Keywords: Alzheimer's disease, Machine learning, Biomarkers, Gene expression, Feature Selection, Classification

1. Introduction

Currently, 40-50 million people around the world are living with dementia and this number has doubled from 1990 to 2016[1]. Alzheimer's disease being the most common form of dementia is expected to rise notoriously with the aging population. With the increase in its incidence, the expenses are also rising. It is estimated that in 2010 alone, Alzheimer's disease had cost the world $604 billion[2] and is expected to incur a global AD-associated healthcare cost of $2 trillion by 2030 affecting more than 131 million people by 2050[3]. Hence, Alzheimer's disease is rapidly emerging as critical global health and economic challenge that has prompted vigorous scientific investigations to identify underlying genetic risk factors and regulatory markers, to suppress the estimated healthcare burden by early detection, especially at pre-symptomatic stages. Much research is performed on the late occurring hallmarks of AD[4-6] such as neurofibrillary tangles, amyloid plaques, neuronal tangles, etc. Although these findings hold some important diagnostic values, the overall therapeutic contributions of these late occurring hallmarks of AD remain murky[4]. Moreover, clinical trials indicate that patients with AD show a varied pattern of symptoms and varying responses to a particular therapy that substantiates several pathological causes, making AD even more intricate to investigate[7].

In recent years, data generated through high throughput gene expression profiling has opened new avenues for a better understanding of the complex disease mechanism and pathways at a molecular level[8, 9]. However, the huge dimension, low sample size, and noise in high-throughput gene expression data make it challenging to identify embedded patterns within the dataset. Currently, the methods to identify the most explaining gene subsets by data reduction and feature selection in the context of gene expression profile dataset analysis are broadly classified into two classes[10]: (i) marginal filtering method[11, 12] and (ii) wrapper (embedded) method[13, 14]. The marginal filtering further is subdivided into two types namely, univariate and multivariate. Some examples of univariate filtering methods are paired t-test (TS), F-test (FT), and Pearson Correlation coefficient (PC)[11-13]. Some multivariate filtering approaches are Analysis of variance (ANOVA), F-score, feature selection based on correlation (CFS), and Max-Relevance-Max-Distance (MRMD)[15-18]. Using these methods, weights are assigned to the features (genes), and the genes with higher weights are considered to be the biologically important features. Although the filtering methods are computationally less expensive than the latter approach, they have significant shortcomings i.e. (i) most of the marginal filtering only accounts for the marginal contribution of a gene candidate while completely ignoring the interdependencies among the genes, and (ii) the absence of classification process. The filtering method doesn't corroborate the classification accuracy of the selected features, reducing its clinical credibility[14]. However, the shortcomings of marginal filtering[19, 20] can be overcome by wrapper methods. Wrapper methods are a hybrid of learning algorithms and classifiers that iteratively search for the optimum set of features by corroborating the classification accuracy of each chosen subset of candidate features[10]. Although the wrapper methods are very computationally intensive for high dimensional gene datasets, the classification accuracies obtained by the feature subsets identified by these methods are noticeably high[14]. In addition to this, machine learning models are empowered with efficient dimension reduction and feature selection methodologies to overcome the curse of dimensionality within the gene expression dataset[21]. Over time, many studies have employed machine learning models on microarray datasets to develop robust predictive models for identifying disease onset and prognosis of complex diseases such as cancer[22-25].

Several studies have extensively leveraged machine learning models to identify biomarkers of AD from phenotypic data such as magnetic resonance imaging[26]. However, the identification of molecular signature underlying the mechanism of AD through gene expression profiles of demented patients remains largely unexplored[27]. In this direction, few studies have employed machine learning on gene expression data to delineate the potential differentially expressed genes (DEGs) within the AD-affected brain[28-31]. These studies have successfully used several state-of-the-art machine learning algorithms such as random forest, decision trees, support vector machines, and deep learning models to the feature selection and classification paradigm[32-35]. Although highly innovative, these methods had their own shortcomings such as, (i) the proposed schemes within many of these methods were able to reduce the dimensions (number of features) but they remained mute on demonstrating the discriminative potential of the acquired DEGs, thus fails to vindicate the practical biological relevance of the obtained geneset. (ii) The majority of these studies incorporated only a small set of samples (usually <30), thus the results remained insufficiently descriptive and have low interpretability[32].

Our objective here is to probe the difference in the gene expression levels within different brain regions of AD patients and non-demented controls, to identify the highly discriminating and biologically relevant gene signatures for AD through the wrapper (embedded) approach. We exclusively probe the Prefrontal cortex (PFC), Middle temporal gyrus (MTG), Hippocampus (H), and Entorhinal cortex (EC) as these regions are the most vulnerable to neurodegenerative diseases[36-38]. To retain the most significant and biologically relevant markers, we conceptualized a simple feature-selection and classification scheme based on the ensemble of random forest (RF) and regularized regression model; plugged with the best-configured classifier to obtain maximum classification accuracy in a 5-fold cross validation test (see **Fig 1**). In addition to validating our finding by integrating biological knowledge through systematic literature review, GeneMania[39], and STRING[40] network analysis; we also corroborate the biological relevance of the obtained gene signatures by quantifying their disease discriminative power for the gene expression data obtained from the Visual Cortex (VC) and the Cerebellum (CR) of both AD affected and control brains. Through this work, we attempt to determine the signatures underlying AD and to formulate an efficient disease identification scheme whose clinical applications could further be extended for other diseases of altered expression.

2. Materials and Methods

2.1 Dataset

We extracted the AD-associated gene expression datasets from the public functional genomics data repository NCBI-GEO database (http://www.ncbi.nlm.nih.gov/geo/). "Alzheimer's" was used as a keyword to query all the experimental studies that have probed the gene expression profile within the brain tissues of AD patients against that of the non-demented healthy controls. The brain regions of our interest are the prefrontal cortex (PFC), middle temporal gyrus (MTG), hippocampus (H), and entorhinal cortex (EC). Datasets of only those studies were used that have performed microarray expression profiling and have a sample size of $\geq$15 for each type of brain tissue. This resulted in eight different studies, from which the samples of four brain tissue types (PFC, MTG, H and EC) were separated and grouped accordingly. This way we obtained a large sample size for each brain region. **Table 1** presents a summary of the expression datasets that are finally incorporated in this work. Each of these studies vary in terms of experimental design and measurements, that require special treatment to screen out definite AD and control samples for which we provided a detailed description of each dataset in supplementary **Table S1**.

The computation was carried out on an Intel (R) Core (TM) i5-4310U, 16 GB RAM, and 64-bit OS Win 10 configuration. Method implementation and experiments were conducted using R version 4.0.3. The schematic representation of machine learning workflow to identify potential biomarkers of AD is shown in **Fig 1**.

Brain Region							
Prefrontal Cortex		Medial Temporal Gyrus		Hippocampus		Entorhinal Cortex	
Dataset (Platform)	AD\Control	Dataset (Platform)	AD\Control	Dataset (Platform)	AD\Control	Dataset (Platform)	AD\Control

GSE33000 (GPL4372)	310\157	GSE118553 (GPL10558)	52\31	GSE5281 (GPL570)	10\13	GSE5281 (GPL570)	10\13
GSE44770 (GPL4372)	129\101	GSE132903 (GPL10558)	97\98	GSE48350 (GPL570)	19\43	GSE48350 (GPL570)	15\39
				GSE28146 (GPL570)	7\8	GSE4757 (GPL570)	20\20

Table 1. The gene expression datasets of Alzheimer's Disease for four different brain regions.

2.1.1 Dataset integration and Pre-processing

To increase our sample size for statistically augmented results, we integrated at least two gene expression datasets for each brain region. However, the merging of the expression dataset is challenging because, (i) the platform over which the datasets were originated varies. Each type of platform measures the expression level of a particular set of genes which could be highly different from the gene repertoire of the other platforms; (ii) Due to adopting varying protocols, platforms and processes, different experiments contain various non-biological technical variations in the measurements[41]. These variations can induce a batch effect to the profiles that is potent to confound the true biological variations, thus may indicate misleading conclusions. To overcome these challenges, we essentially chose only those datasets to merge that were generated over a common platform. To subdue the batch effect, we standardized the expression profile of each sample, thus accounting for only the distribution of the gene expression[42]. For each dataset, the probe IDs were mapped to their respective Entrez gene IDs and Genbank Accession IDs that are annotated in the dataset's corresponding platform table. In the case of duplicated gene IDs, the candidate with the maximum interquartile range was kept for further analysis. It was only after this step, we z-score normalized each sample to capture the distribution of the expression. We evaluated the p values for each gene candidate using both paired t-test and Mann Whitney U test, followed by its corresponding FDR correction for PFC and MTG due to their large sample size (> 200). Finally setting $p<0.05$ and FDR < 0.01, we prune our fully merged and pre-processed datasets for feature selection and classification.

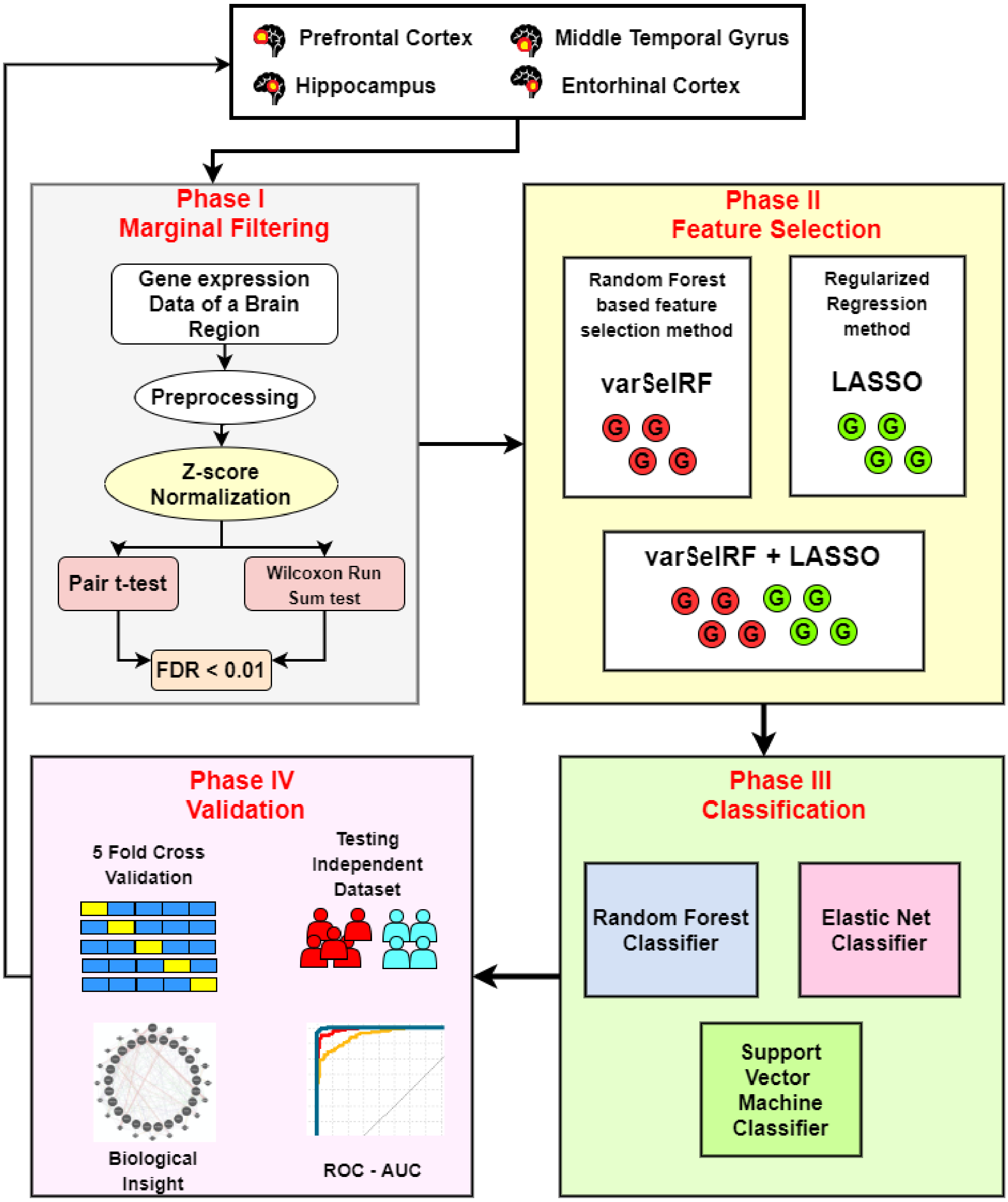


Fig 1. Schematic representation of the Machine Learning workflow to identify potential biomarkers for AD. The gene expression data for a given brain region is processed in the phase I. The features are then identified using wrapper methods (phase II). Subsequently in phase III and phase IV, the discriminative power and the biological relevance of the identified geneset is quantified and validated.

2.2 Feature selection

As aforementioned, the merged gene expression datasets were the compilation of measurements from different samples but were generated from the same brain tissue, thus capturing the crucial biological basis for such expression within that particular brain region. To fetch the important independent players (gene candidates) underlying these expression levels, we employed two highly efficient feature selection methods; (i) Variable selection using Random forest method[43] and (ii) Lasso regression method[44]. The parent models of these methods are probably the most pervasive machine learning algorithms i.e., random forest and generalized regression model respectively. The formalisms and the implementations of these methods are elaborated in the following sub sections.

2.2.1 Variable Selection Using Random Forest (varSelRF)

The random forest algorithm developed by Breiman L.[45, 46] uses the ensemble of regression trees for classification. Employing a bootstrap sample of the data, the classification tree is built. The candidate set of variables at each split of the tree is a random subset of the variables[44, 47]. In this way, RF incorporates bootstrap aggregation (bagging) and feature selection to build trees. To obtain low-bias trees, each tree is grown fully, and then bagging and random selection of variables is performed to facilitate low correlation of the individual trees[43]. For each fitted tree, RF registers a measure of error rate (OOB error) based on the out-of-bag cases (samples that have no contribution in the tree formation) that have very crucial applications in data reduction and feature selection. A detailed description of the algorithm underlying RF is provided in the supplementary text. Based on the characteristics of the RF algorithm, Ramón et al.[43] formularized a feature selection model called varSelRF. This method is available as a package "*varSelRF*" on CRAN repository[48]. varSelRF iteratively fits random forests and selects a set of features (genes) that retains a minimum OOB error rate. Exploiting the embedded classification process, varSelRF returns a small subset of important genes while augmenting the predictive performance. This approach has already been incorporated in several literatures and has shown promising results[49-52]. The rationale to employ varSelRF in our framework is (i) the method returns a small set of gene candidates that has low correlation and high predictive power[52] and (ii) RF based approach requires a less fine-tuning of parameters as the default parameter values often deliver the best performance[53].

2.2.2 Regularized regression models

Least Absolute Shrinkage and Selection Operator (LASSO) is a type of regularization regression method to fit a generalized linear model. Based on the idea of penalizing the regression model (L1-norm), LASSO squashes the regression coefficient to zero for the variable that has the least contribution to the model. This way the LASSO regression model has an optimal feature selection capability. LASSO regression is an alternative regression approach to Ridge regression that too is based on penalizing the model but follows a L2-norm[44].

For a given population X, let x_{ij} be the $i^{th}(1 \leq i \leq n)$ observation of the $j^{th}(1 \leq i \leq p)$ variable and let y_i be the corresponding label of the i^{th} instance. For each p variable, the regularized regression model estimates the regression coefficient $\beta_j(1 \leq i \leq p)$ by minimizing the sum of squared error (eq. 1) along with a constraint on the coefficients $\sum J(\beta_j) \leq t$[44, 54, 55].

$$\beta = argmin_{\beta} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \qquad \text{eq. 1}$$

For LASSO the coefficient $J(\beta_j) = |\beta_j|$ and in the Ridge $J(\beta_j) = \beta_j^{\,2}$ [42, 51]. This way LASSO regression tranculates the coefficient of the non-contributing variable to zero while Ridge shrinks the coefficient close to zero, delineating LASSO as an efficient feature selection model. The LASSO obtains the β_j estimate by minimizing eq. 2.

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad \text{eq.2}$$

where λ is the penalty parameter that determines the shrinkage proportion and is often determined using cross-validation[44]. LASSO retains an excellent performance for the situations when (i) the data has very high dimension and low sample and (ii) few variables explain the majority of data (have large coefficient) and the remaining variable has very low predictive potentials[44]. Moreover, LASSO has some significant advantages such as (i) LASSO efficiently handles the multicollinearity within the features and returns highly independent features and (ii) Being computationally less expensive, LASSO retains the optimal gene candidates faster. These characteristics of LASSO befit the gene expression data as a feature selection model. LASSO has elucidated excellent performance in numerous studies[55-58], delineating as a very promising feature selection model. The variables with relative scaled importance >10 was considered significantly important.

However, studies have indicated that L1-norm (LASSO) is not universally dominant over the L2 norm (Ridge). However, to improve computational tractability Zou et al.[59] proposed a relatively new penalty called the Elastic Net, built as an intelligent compromise between LASSO and Ridge penalty. For Elastic Net, the $J(\beta_j)$ (coefficient constrain) is:

$$j(\beta_j) = \lambda \sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha\,)|\beta_j|) \qquad \text{eq.3}$$

where the new α constant is introduced that regulates the intensity of LASSO and Ridge penalties. Elastic Net handles multicollinearity more efficiently than LASSO by accounting for every correlated pair during training[44, 59]. Elastic Net has better performance on many occasions, however, there are only a few studies that corroborate the same[60]. Although we have employed LASSO as a feature selection, we leverage the Elastic Net classifier to test the determinative power of all the selected features combined due to its high efficiency towards multicollinearity. The R package “caret” was used to implement LASSO and Elastic Net[61].

2.2.3 Multiplicity Problem

For microarray dataset, the problem of multiplicity can cast a false sense of trust in the genes identified by wrapper approach. The basis of multiplicity problem has been explained in detail in the supplementary text. Subscribing to the notion of the studies investigating the problem of multiplicity[62-65], we lend credence to the combined set of genes that were obtained by both the methods (varSelRF and LASSO); and exclusively probed the biological significance of the common and repeatedly selected gene candidates.

2.3 Classification model

Classification modeling led by feature selection is a crucial phase of the paradigm, that depicts the clinical application of the selected gene candidates. Although the embedded classifier within the wrapper method leverages the classification accuracy to quantify the importance of a gene subset, but in the context of therapeutic application it is very crucial to corroborate the best suiting classification model that improves the prediction accuracy. In this work, we employed Support Vector Machines (SVM), Random forest classifier, and Elastic Net classifier and performed a comparison study. We also probed the classification efficacy of these models for the gene candidates obtained by (i) varSelRF, (ii) LASSO and (iii) combined gene subsets retained by both varSelRF and LASSO. An overview of RF and SVM classifiers is provided in the supplementary text. The R package "random Forest" and "e1071" were used to implement the RF[53] and SVM[66] respectively.

2.4 Assessment

2.4.1 Model Assessment

We assess the prediction power of the selected gene candidates through SVM, RF, and Elastic Net classifiers. Exploiting the relatively large sample size due to the merging of gene datasets, we perform a 5-fold cross validation method to judge the external prediction power of the gene set as well as of the classification model with a high level of certainty. To compare the efficacy of the models we measure the following metrics.

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TN}+\text{FN}+\text{FP}+\text{TP})}$$

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN}+\text{FN})}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})}$$

$$\text{Matthew's Correlation (MCC)} = \frac{(\text{TP}\times\text{TN}-\text{FP}\times\text{FN})}{\sqrt{(\text{TP}+\text{FN})\times(\text{TP}+\text{FP})\times(\text{TN}+\text{FN})\times(\text{TN}+\text{FP})}}$$

Here TP, TN, FN and FP represent true positive, true negative, false negative and false positive predictions respectively made by the classification model for each brain region (AD state is denoted positive and non-demented healthy state is denoted negative). For further comparative analysis, we plot the receiver operating characteristics (ROC) curve and compared the area under the curve (AUC) obtained by the model for different brain tissue.

2.4.2 Feature Assessment

We assess the biological significance of the feature set obtained by our framework by integrating the biological knowledge through a systematic literature review. We have used GeneMania[39] and STRING[40] network analysis to identify the co-expression, genetic and physical interactions among the obtained biomarkers of AD and also with the previously well-known AD genes. Using the same, we also delineated the networks (hub genes) associated with our obtained molecular signatures to deliver deeper insight into the mechanism of AD in different brain tissues. To further corroborate the biological meaningfulness of the obtained markers, we tested the discriminative power of these markers to classify AD patients from non-demented controls for two different brain regions, Visual Cortex (VC) and the Cerebellum

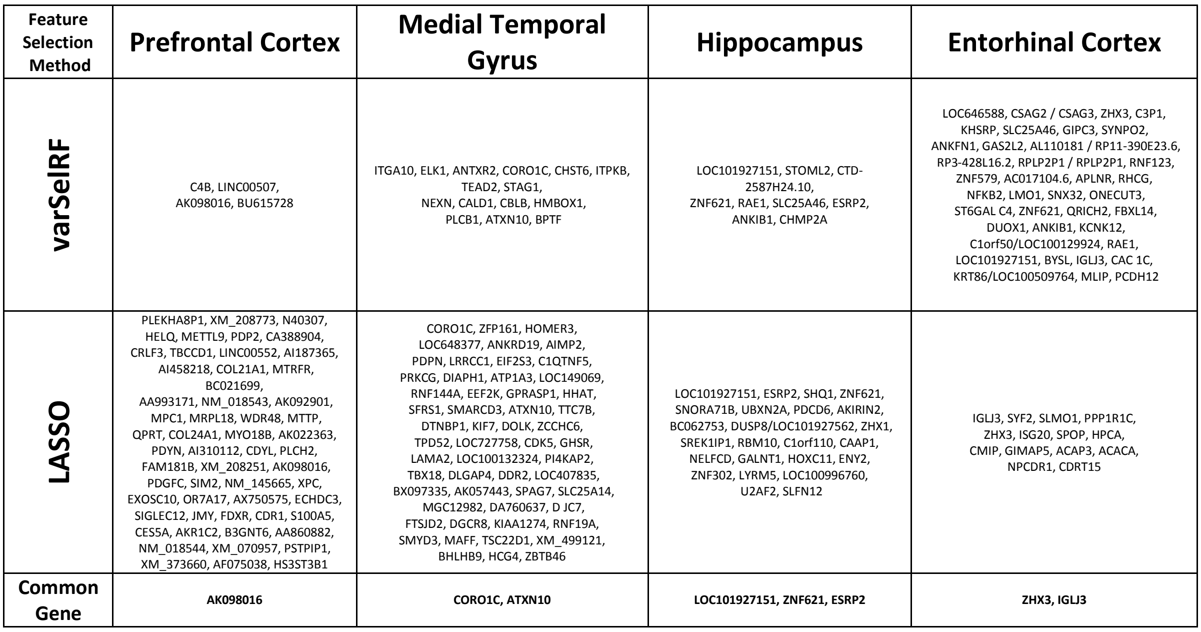
(CR). This way, not only the biological relevance is unmasked quantitatively, the therapeutic application of the proposed framework is also depicted.

3. Results

After marginal filtering in the first phase (Phase I), we obtained 26,593, 13,037, 3,268 and 10,029 genes for PFC, MTG, H and EC respectively, that was further processed to identify DEGs using the varSelRF and LASSO method (Phase II). In the following section, we shed light on the obtained features as well as their discriminative power when treated with different benchmark classifiers.

3.1 The Obtained Features

For each brain region, we employed varSelRF and LASSO on the gene candidates from phase I. varSelRF is based on RF that has the inner nature of being purely random and performs random sampling within the algorithm. This leads to slightly varying results when implemented multiple times. Therefore, for each brain region, we implemented varSelRF for five times and considered each selected candidate as important feature. The tuned hyperparameter sets for varSelRF and LASSO are provided in the supplementary **Table S2**. The features obtained are summarized in **Table 2**. We found that LASSO obtained a higher number of candidates than varSelRF for the brain region with a large sample size and vice versa.



Feature Selection Method	Prefrontal Cortex	Medial Temporal Gyrus	Hippocampus	Entorhinal Cortex
varSelRF	C4B, LINC00507, AK098016, BU615728	ITGA10, ELK1, ANTXR2, CORO1C, CHST6, ITPKB, TEAD2, STAG1, NEXN, CALD1, CBLB, HMBOX1, PLCB1, ATXN10, BPTF	LOC101927151, STOML2, CTD-2587H24.10, ZNF621, RAE1, SLC25A46, ESRP2, ANKIB1, CHMP2A	LOC646588, CSAG2 / CSAG3, ZHX3, C3P1, KHSRP, SLC25A46, GIPC3, SYNPO2, ANKFN1, GAS2L2, AL110181 / RP11-390E23.6, RP3-428L16.2, RPLP2P1 / RPLP2P1, RNF123, ZNF579, AC017104.6, APLNR, RHCG, NFKB2, LMO1, SNX32, ONECUT3, ST6GAL C4, ZNF621, QRICH2, FBXL14, DUOX1, ANKIB1, KCNK12, C1orf50/LOC100129924, RAE1, LOC101927151, BYSL, IGLJ3, CAC 1C, KRT86/LOC100509764, MLIP, PCDH12
LASSO	PLEKHA8P1, XM_208773, N40307, HELQ, METTL9, PDP2, CA388904, CRLF3, TBCCD1, LINC00552, AI187365, AI458218, COL21A1, MTRFR, BC021699, AA993171, NM_018543, AK092901, MPC1, MRPL18, WDR48, MTTP, QPRT, COL24A1, MYO18B, AK022363, PDYN, AI310112, CDYL, PLCH2, FAM181B, XM_208251, AK098016, PDGFC, SIM2, NM_145665, XPC, EXOSC10, OR7A17, AX750575, ECHDC3, SIGLEC12, JMY, FDXR, CDR1, S100A5, CES5A, AKR1C2, B3GNT6, AA860882, NM_018544, XM_070957, PSTPIP1, XM_373660, AF075038, HS3ST3B1	CORO1C, ZFP161, HOMER3, LOC648377, ANKRD19, AIMP2, PDPN, LRRCC1, EIF2S3, C1QTNF5, PRKCG, DIAPH1, ATP1A3, LOC149069, RNF144A, EEF2K, GPRASP1, HHAT, SFRS1, SMARCD3, ATXN10, TTC7B, DTNBP1, KIF7, DOLK, ZCCHC6, TPD52, LOC727758, CDK5, GHSR, LAMA2, LOC100132324, PI4KAP2, TBX18, DLGAP4, DDR2, LOC407835, BX097335, AK057443, SPAG7, SLC25A14, MGC12982, DA760637, D JC7, FTSJD2, DGCR8, KIAA1274, RNF19A, SMYD3, MAFF, TSC22D1, XM_499121, BHLHB9, HCG4, ZBTB46	LOC101927151, ESRP2, SHQ1, ZNF621, SNORA71B, UBXN2A, PDCD6, AKIRIN2, BC062753, DUSP8/LOC101927562, ZHX1, SREK1IP1, RBM10, C1orf110, CAAP1, NELFCD, GALNT1, HOXC11, ENY2, ZNF302, LYRM5, LOC100996760, U2AF2, SLFN12	IGLJ3, SYF2, SLMO1, PPP1R1C, ZHX3, ISG20, SPOP, HPCA, CMIP, GIMAP5, ACAP3, ACACA, NPCDR1, CDRT15
Common Gene	**AK098016**	**CORO1C, ATXN10**	**LOC101927151, ZNF621, ESRP2**	**ZHX3, IGLJ3**

Table 2. The gene biomarkers obtained for different brain regions using varSelRF and LASSO methods.

Both the models largely identified a varying set of markers; however few gene candidates were commonly identified by both methods. Of interest, the majority of these commonly identified markers are closely associated with neurodegenerative disorders, depicting the biological significance of the models. In addition to the common genes identified by the models, there were common regulatory gene candidates within the brain regions **(see Fig S1)**. The common biomarkers found within the H and EC region are ZNF621, SLC25A46, RAE1, and ANKIB1. Among these biomarkers, RAE1, ANKIB1, and SLC25A46 have been reported to be prominently involved in several neurodegenerative disorders. The RAE1 protein is found to be the interacting partner of Huntingtin protein aggregates[67] and experimental evidence of early ageing associated phenotypes is reported in Rae1 haplo-insufficient mice[68]. ANKIB1 is also found to be associated with Cerebral cavernous malformations[69]. Another potential biomarker that has been associated with neurodegenerative disorders is SLC25A46. A study by Abram's et al. has experimentally shown that the mutations in the SLC25A46 genes can lead to the degeneration of optic and peripheral nerve fibers[70]. Also, loss of function in the SLC25A46 gene leads to lethal congenital and peripheral neuropathy[71, 72]. Although these genes have been extensively studied for different neurological disorders, their role in Alzheimer's disease is yet to be exclusively explored. Our models were also able to unravel the participation of non-coding RNAs, identifying 9 non-coding RNAs within the brain regions. Among the non-coding RNAs, we found two long non-coding RNAs, AK057435 and BC037880 in the prefrontal cortex and the hippocampus region respectively that are classified as potential biomarkers. Since long non-coding RNAs are known to play an important role in human neurological development and cognition, experimental characterization of these biomarkers can help to elucidate the role of long non-coding RNAs in Alzheimer's disease.

3.2 Classification

To determine the classification potential of the obtained gene set for each brain region, we built three benchmark classification models (SVM, random forest and Elastic Net). Performing extensive machine learning experiments, we made an attempt to identify the best pair of feature-selection and classification models in the context of disease class prediction. For each of the four brain regions, we applied three different best-configured classification model to the gene set obtained through varSelRF, LASSO and finally to the combine pool of gene set (varSelRF + LASSO), depicting a total of 9 scenarios to identify the best performing combination. The classification performance was assessed through a 5-fold cross validation method. **Table 3** represents a complete summary of the assessment metrics obtained for each possible scenario. In our study, the proposed framework has obtained foremost the highest AD prediction accuracy than any previous studies in a similar paradigm to our knowledge to date. For the prefrontal cortex and hippocampus, the scheme has even obtained 100% prediction accuracy.

Feature Selection Method	Model	Brain Region																			
		Prefrontal cortex					Middle temporal gyrus					Hippocampus					Entorhinal cortex				
		Acc	Sen	Spe	Pre	Mcc	Acc	Sen	Spe	Pre	Mcc	Acc	Sen	Spe	Pre	Mcc	Acc	Sen	Spe	Pre	Mcc
VarSelRF	SVM	0.93	0.95	0.90	0.91	0.85	0.86	0.85	0.87	0.89	0.73	0.90	1.00	0.80	0.84	0.82	0.91	0.95	0.80	0.80	0.75
	RF	**0.95**	**0.95**	**0.94**	**0.94**	**0.89**	0.87	0.88	0.87	0.88	0.74	**0.94**	**1.00**	**0.88**	**0.91**	**0.89**	**0.95**	**1.00**	**0.85**	**0.87**	**0.84**
	ElasticN	0.93	0.97	0.90	0.90	0.87	**0.88**	**0.86**	**0.89**	**0.90**	**0.75**	0.93	0.97	0.88	0.91	0.87	0.94	1.00	0.82	0.84	0.81
LASSO	SVM	0.99	1.00	0.99	0.99	0.99	0.96	0.96	0.95	0.96	0.91	0.79	0.81	0.77	0.78	0.59	0.91	0.92	0.90	0.95	0.79
	RF	0.97	0.97	0.96	0.96	0.93	0.91	0.91	0.91	0.92	0.81	0.83	0.83	0.83	0.84	0.66	0.92	0.91	0.92	0.88	0.78
	ElasticN	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.86**	**0.86**	**0.87**	**0.87**	**0.74**	**0.97**	**0.98**	**0.97**	**0.97**	**0.93**
VarSelRF + LASSO	SVM	0.99	1.00	0.99	0.99	0.98	0.95	0.95	0.95	0.96	0.91	0.99	1.00	0.97	0.98	0.97	0.92	0.95	0.82	0.84	0.78
	RF	0.97	0.98	0.96	0.96	0.94	0.88	0.87	0.88	0.90	0.75	0.94	0.98	0.91	0.93	0.90	**0.95**	**1.00**	**0.85**	**0.87**	**0.84**
	ElasticN	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.99**	**0.98**	**0.98**	**0.96**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.94	1.00	0.82	0.84	0.81

Table 3. Performance comparison of the three different classification models (SVM, RF, Elastic Net) applied to the gene set obtained through varSelRF, LASSO and varSelRF + LASSO for the four brain regions, namely Prefrontal cortex, Middle Temporal Gyrus, Hippocampus and Entorhinal Cortex.

3.3 Performance Evaluation

It was observed that in the majority of the scenarios, the Elastic Net classifier obtained excellent performance, followed by the random forest classifier, while SVM performance remained low (**Fig 2**). Substantiating the parent algorithms, both RF and Elastic Net classifier has performed higher for the gene sets obtain through their respective allied feature selection model i.e., varSelRF and LASSO respectively. Considering the problem of multiplicity, we substantiate the combined gene markers of varSelRF and LASSO over the gene set obtained by these individual methods. The ROC-AUC plot elucidates the superiority of Elastic Net over RF and SVM for three brain regions (PFC, MTG and H) while remaining slightly lower but highly competitive for the EC region (**Fig 3**). The one explanation of low performance of Elastic Net for EC region is possible due to the very small sample to gene ratio.

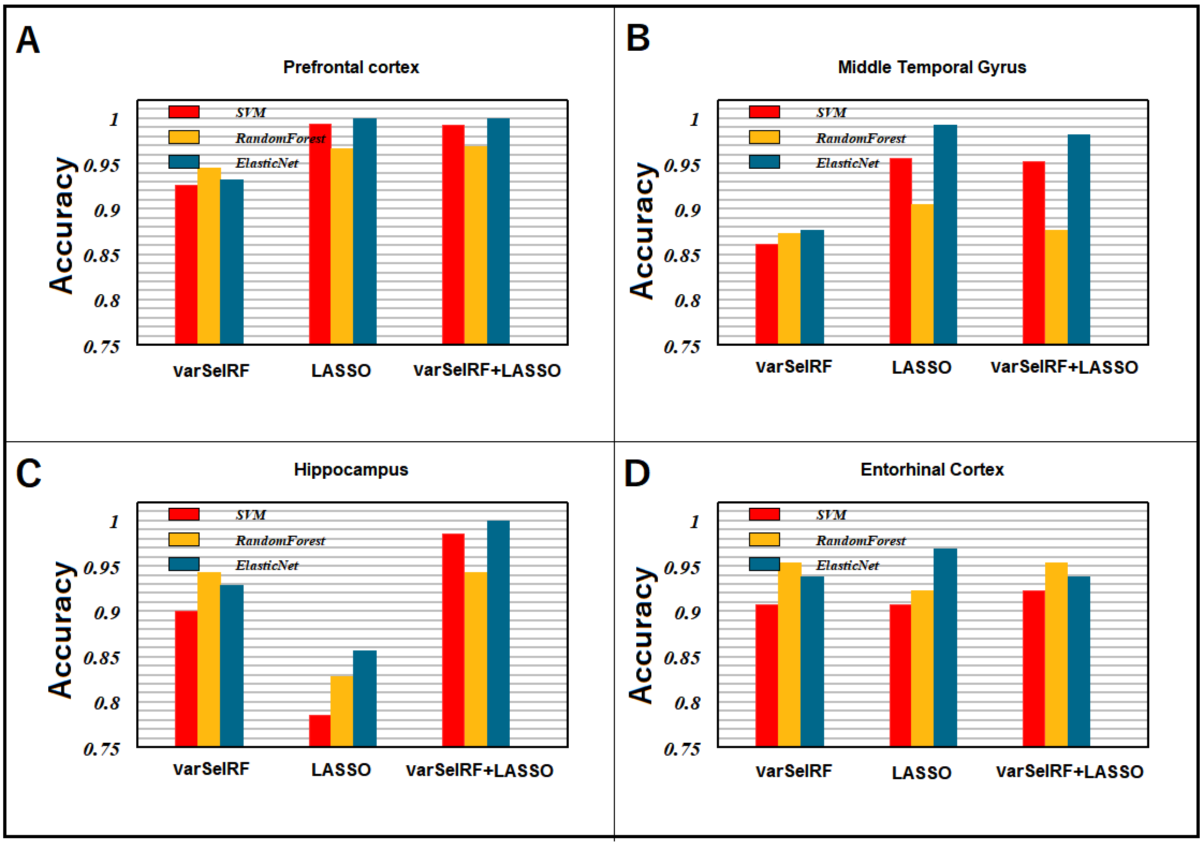


Fig 2. Prediction accuracy obtained by the SVM, Random Forest, Elastic Net classifier employed in the varSelRF, LASSO and varSelRF + LASSO for **(A)** Prefrontal cortex, **(B)** Middle Temporal Gyrus, **(C)** Hippocampus and **(D)** Entorhinal Cortex. The Elastic Net classifier obtained excellent performance in the majority of scenarios, followed by the random forest classifier and SVM. Genes obtained through LASSO with Elastic net classifier performed higher in PFC, MTG and EC region.

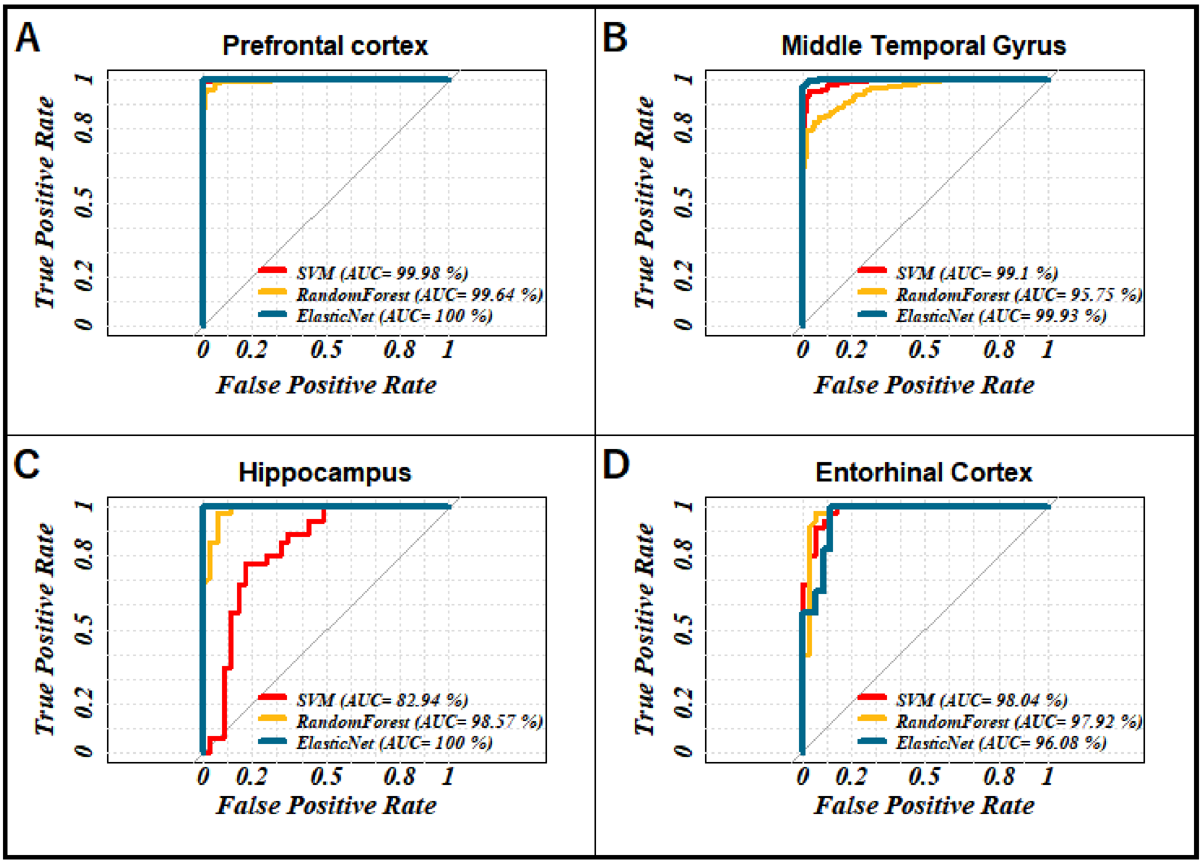


Fig 3. The classification performances to discover potential biomarkers in four brain regions. The ROC-AUC curves of Elastic Net, Random Forest and SVM classifiers for **(A)** Prefrontal cortex, **(B)** Middle Temporal Gyrus, **(C)** Hippocampus and **(D)** Entorhinal Cortex.

In addition to adopting a 5-fold cross validation method, we also took several other measures to establish the biological credibility of the identified gene candidates. We hypothesize that the gene markers obtained for one brain region hold some biological relevance for the adjacent brain region. We therefore evaluated the AD prediction potential of the gene subset of PFC (60 genes) for the gene expression data obtained from Virtual Cortex (VC) and Cerebellum (CR). VC and CR data were extracted from GEO NCBI database (GSE44771 and GSE44768). The sample size for VC and CR are both 230 with AD to control ratio of 129:101. We employed LASSO feature-selection only for the expression level of those 60 gene candidates that were identified as PFC markers on the VC and CR datasets. We find that the biomarkers of PFC displayed an excellent AD classification performance (5-fold CV) of 92% and 91% on VC and CR datasets respectively (see supplementary **Table S3**). The complete assessment metric obtained for VC and CR is provided in the supplementary **Table S4.** This quantitatively validates the biological meaningfulness of gene candidates obtained in our study.

4. Discussion

The formalism of the proposed framework has two integrated components (i) Identification of the AD associated crucial gene markers within each brain region and (ii) the disease class prediction. After carrying out an extensive comparative analysis and corroborating the problem of multiplicity, it is apparent that Elastic Net classifier has a remarkable potential for disease prediction when employed over the gene subset identified by multiple varieties of gene selection models (LASSO and varSelRF in this case). In addition to having outstanding AD predictive potentials, the markers identified through this framework are of high calibre in terms of explaining the expression level and multicollinearity.

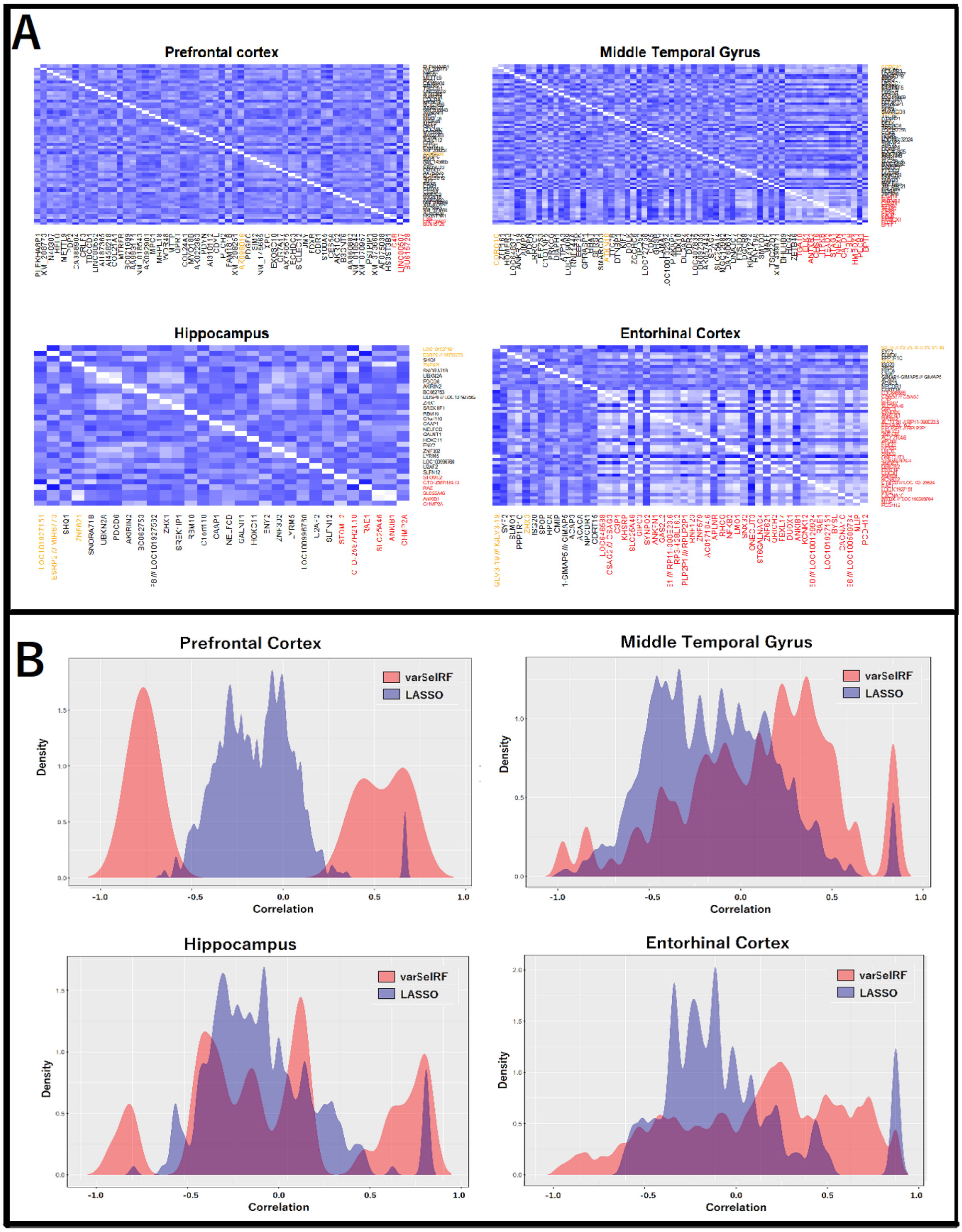


Fig 4. (A) The correlation heatmap (n x n, where n is number of biomarkers) for the expression level of the biomarkers obtained by LASSO and varSelRF method for each brain region. Every

block in a heatmap plot represents correlation between the gene on each axis. Correlation ranges from -1 to +1. The shade corresponding to the values closer to zero indicate low linear trend between the two markers. The red labelled markers are the one that are obtained by varSelRF. The black labelled markers are the one that are obtained by LASSO. The orange labelled are the markers that were identified by both the models. **(B)** The density plot for the correlation values among the gene subset obtained by each type of feature selection model within different brain region. Density plot of correlation value for the markers obtained through varSelRF is shown in red. Density plot of correlation value for the markers obtained through LASSO is shown in blue. The density for the correlation value near to zero remained higher for LASSO comparative to varSelRF in every brain region.

Fig 4A illustrates the correlation heatmap for the expression level of the biomarkers obtained by LASSO and varSelRF for each brain region. We see that the biomarkers elucidated very low correlation, thus together they are of great relevance in the context of depicting the biological basis for the observed expression level. Although both feature selection models are immune to multicollinearity, the LASSO obtained significantly lower correlated markers than that of varSelRF, especially for the EC region. This is also apparent in the correlation density plot for the regions, where the density remained high near the centre for the geneset obtained through LASSO, while it remains inflated on the tails for the varSelRF obtained geneset (**Fig 4B**).

4.1 Biological Insight

We performed a combination of biological network analysis and a comprehensive literature review to validate the biomarkers obtained in our study. We started with bioinformatics analysis of all the biomarkers obtained from our models and are listed in supplementary **Table S5**. We find the presence of potential biomarkers in all the chromosomes, except Chromosome Y. This may point towards the higher prevalence of AD in woman than in man[73]. The Chromosomes 1, 6, 17, 19 are found to contain the maximum number of biomarkers (**Fig S2**). Although most of the genes that are classified as biomarkers in our study are protein coding genes, some non-coding genes, such as LINC00552, LINC00507, MGC12982, HCG4, LOC101927151, NPCDR1, LOC646588 are also found to be the biomarkers of AD. These non-coding genes are novel and mostly uncharacterised.

Moreover, we identified 7 up-regulated and 6 down-regulated genes in the AD samples with respect to the normal ones by employing the GSE5281 expression data due to the availability of raw count. We considered $p < 0.01$ and $|\log2FC| \geq 0.6$ (FC, fold change) as cut-off criterion on different samples of H and EC brain regions from the GSE5281 dataset. Using this information, we identified the biomarkers that are up and down regulated (**Fig S3**). We find that some of the biomarkers are significantly downregulated in AD such as MLIP and STOML2. While the down regulation of STOML2 gene has been reported previously in AD patient's samples[74], the EC biomarker, MLIP can be clinically tested as a novel possible biomarker of AD.

We also performed GeneMania network analysis for all the biomarkers of each brain region (**Fig. S4-7**) and found that the biomarkers are not only co-expressed but share both physical and genetic interactions. Some of the highly interacting genes in the PFC are ECHDC3, PDGFC, MPC1, CRLF3, CDYL, FDXR that are also found to be co-expressed (**Fig S4**). Among these genes, the expression of ECHDC3 is found to be significantly higher in AD patients than non-AD patients from genome-wide association studies of more than 200,000 individuals[75]. Also, CRLF3 has been studied in neuronal aging rates in human brain regions[76]. In the EC region, we see that the biomarkers interact with each other by largely physical and genetic interactions **(Fig S5)**. In particular, ZNF621 and ISG20 are found to genetically interact with many of the other biomarkers. The ZNF621 gene has been recently reported as an upregulated gene in AD patients[77]. From the network analysis of the H region, we find extensive interactions of the biomarkers with each other, where the biomarkers not only are involved in physical and genetic interactions as well as co-expression and co-localization (**Fig S6**). Some of the highly interacting biomarkers of the H region are RBM10, SLC25A46, STOML2. It is interesting to note that both the SLC25A46, STOML2 protein are involved in mitochondrial dynamics and it has been proposed that mitochondrial dysfunction due to oxidative stress may be one of the earliest and prominent features of AD; and it has been experimentally shown that slower mitochondrial dynamics is correlated with reduced expression of STOML2 and MFN2[74]. The network analysis of MTG region shows that most of the biomarkers in the region genetically interact with each other, however, co-expression is also seen for some of the biomarkers such as CALD1, DNAJC7, TSC22D1, CMTR1, CORO1C (**Fig S7**). TSC22D1 is one of the most studied transcription factors that has also been reported as the potential new target for treating AD[78]. Hence, the biomarkers found by our models have not only been studied for different neuropathies but some of them are also reported as potential targets against AD. Also, we see that our biomarkers extensively interact with each other and thus, careful targeting of a potential biomarker can also help to regulate the biological functions of other biomarkers involved in various neuropathies.

4.2 Relationship between the biomarkers and AD genes

The most well-known genes that have the largest effect on the risk of developing AD are APOE, APP, PSEN1, and PSEN2[79]. Although we have not identified these genes in our study, the relationship between these AD genes and our biomarkers is worth analysing. To seek the potential interactions between the biomarker genes and the AD genes according to different brain regions, the STRING[40] (*Search Tool for the Retrieval of Interacting Genes/Proteins*) tool was employed. Active interaction sources such as experimental data, public databases, text mining, computational prediction methods, and species limited to "*Homo sapiens*" are applied to construct the protein-protein interaction (PPI) networks. From the interaction networks shown in **Fig 5**, we see that the biomarkers of all the brain regions, except the hippocampus have interactions with the AD genes. In the prefrontal cortex, the biomarkers showing significant interactions with the AD genes are C4A, SIM2 and PDYN (**Fig 5A**). The complement pathway protein, C4A is found to be present in higher levels in patients with AD and represents the inflammation generally associated with neurodegenerative diseases[80]. The biomarker SIM2 is also supposed to serve as a noble target for Down's Syndrome-related AD[81]. Although the PDYN gene is extensively studied in Huntington's Disease[82], its role in AD is yet to be explored. The interacting biomarkers with AD genes in the MTG region are CDK5, GHSR, PLCB1, ITPKB, HOMER3 (**Fig 5B**). CDK5 is gradually emerging as an obvious therapeutic target for AD because Cdk5/p25 is involved in two most important pathological hallmarks of AD, the formation of Aβ plaques and NFTs[83]. Also, in the current scenario, we see GHSR, PLCB1, ITPKB genes are considered to be promising therapeutic targets for AD[84-87]. Similarly, in the EC region, the interacting biomarkers are NFKB2, CACNA1C, APLNR (**Fig 5D**) . The transcription factor NFKB2 has emerged as a potential target for AD prevention by targeted anti-inflammatory treatment to increase the time of disease onset[88]. Moreover, by targeting the calcium voltage-gated channel subunit alpha-1 C gene, CACNA1C by miRNA, studies have reported the inhibition of tau protein hyperphosphorylation in AD[89]. The apelin receptor protein, APLNR is also been recently studied as a potential target for several neurodegenerative diseases including AD as expression level alterations in apelin significantly affects the neuronal structure, calcium signalling, apoptosis, and autophagy etc[90]. From the analysis, we see that some of our biomarkers that closely interact with the well-known AD genes are also closely associated with various neurological disorders including AD. Future work requires the experimental testing of these gene biomarkers found in our study to identify the potential signature biomarker for efficient early diagnosis and treatment of AD.

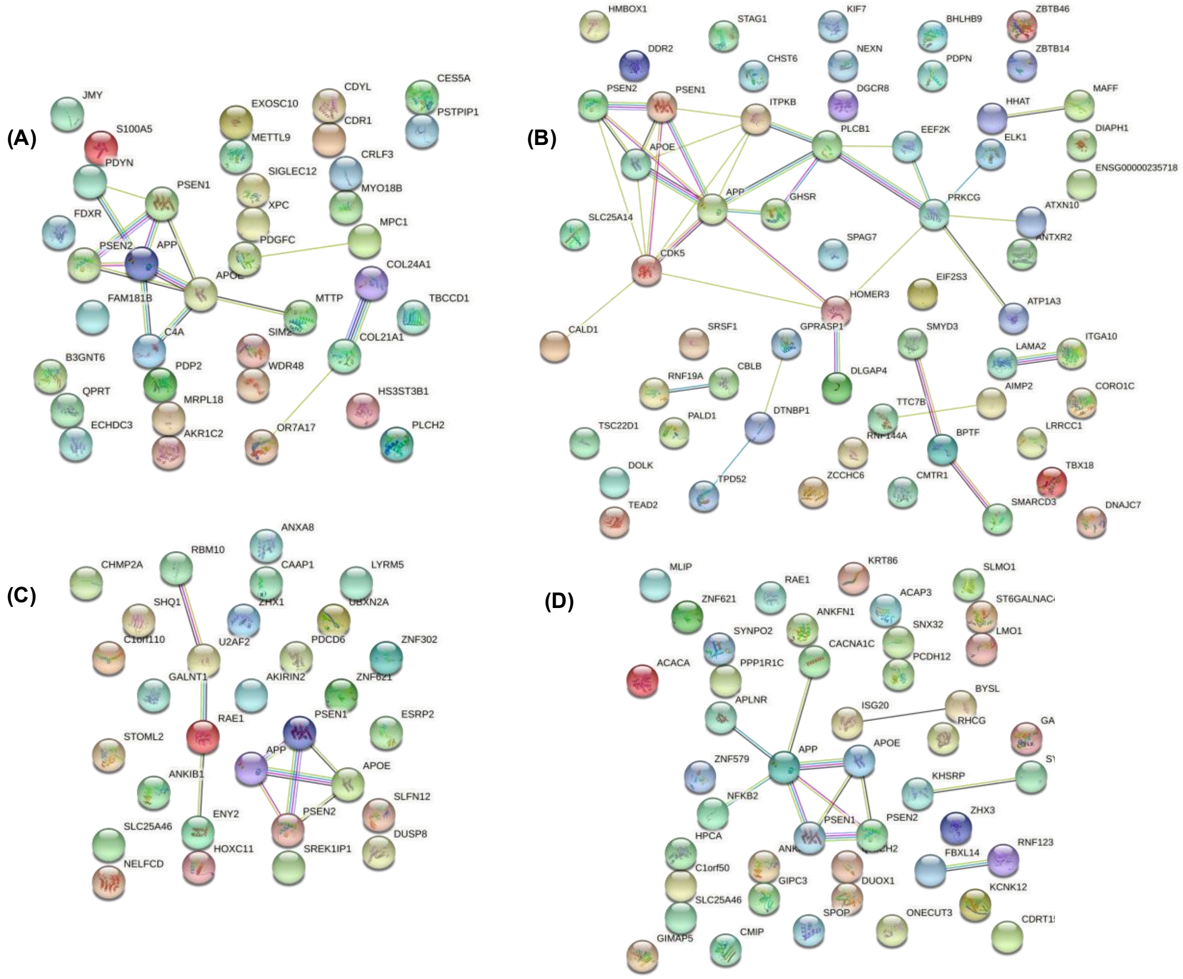


Fig 5. Protein-protein interaction (PPI) networks of the gene biomarkers for **(A)** Prefrontal cortex, **(B)** Middle Temporal Gyrus, **(C)** Hippocampus and **(D)** Entorhinal Cortex. The coloured nodes represent the proteins with first shell of interactions whereas the white nodes represent second shell of interactions. The proteins whose 3D structure are not known is shown by empty nodes. The coloured edges represent protein-protein interactions[40].

5. Conclusion

The use of comprehensive machine learning models to identify potential gene biomarkers for Alzheimer's disease is a significant step to determine the early treatment of AD patients. In this work, we propose a simple and robust framework to identify biologically important genes in the context of AD. There are three crucial aspects that corroborate the strength of the framework, (i) To identify the potential genetic markers of AD, probing the gene expression data from different brain tissue is more effective than analysing the combined profiles of expression level from all the regions together. In addition to that, incorporating a large sample size augments the credibility of the findings. (ii) The use of the best configured benchmark machine learning based feature selection model (wrapper approach) provided the most explaining gene subsets with the highest AD predictive power. (iii) To explain the biological significance, a strong validation is a must. Alongside conducting an extensive literature survey, the biological relevance is elucidated quantitatively by testing the biological significance of the obtained gene for two independent brain regions (Visual Cortex and Cerebellum). By employing the gene expression data of diseased vs. normal patients for four different brain regions to identify the biomarkers and incorporating them, our study has achieved, by far the highest prediction accuracy through optimally configured classification models.

In summary, we found several potential biomarkers, some of which are previously linked to AD such as ECHDC3, ZNF621, STOML2, TSC22D1, SIM2, CDK5, C4A, GHSR, PLCB1, ITPKB, NFKB2, CACNA1C, etc. and some novel biomarkers such as CORO1C, SLC25A46, RAE1, ANKIB1 CRLF3, PDYN, AK057435, and BC037880. Future work requires clinical and experimental testing of these gene candidates to identify potential prognostic biomarkers that can support the early diagnosis of Alzheimer's disease or can be targeted at the gene level to prevent the disease. We will also extend the application of the proposed paradigm to discover novel potential markers for other complex diseases in future.

6. Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

7. Conflict of Interests

The authors declare that they have no conflict of interests.

8. References:

1. Nichols, E. et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology* **18**, 88-106 (2019).
2. Wimo, A. & Prince, M. World Alzheimer Report 2015, The Global Impact of Dementia. *Alzheimer's Dis. Int.* (2015).
3. Prince, M., Comas-Herrera, A., Knapp, M., Guerchet, M. & Karagiannidou, M. World Alzheimer Report 2016. Improving healthcare for people living with dementia. *Alzheimer's Dis. Int.* (2016).
4. Huang, Y. & Mucke, L. Alzheimer mechanisms and therapeutic strategies. *Cell* **148**, 1204-1222 (2012).
5. Putcha, D. et al. Hippocampal hyperactivation associated with cortical thinning in Alzheimer's disease signature regions in non-demented elderly adults. *J. Neurosci.* **31**, 17680–17688 (2011).
6. Palop, J.J., and Mucke, L. Amyloid-beta-induced neuronal dysfunction in Alzheimer's disease: from synapses toward neural networks. *Nat. Neurosci.* **13**, 812–818 (2010).

7. Oxford, A.E., Stewart, E.S. & Rohn, T.T. Clinical Trials in Alzheimer's Disease: A Hurdle in the Path of Remedy. *Int J Alzheimers Dis.* , 5380346. (2020).

8. Lee, T. & Lee, H. Prediction of Alzheimer's disease using blood gene expression data. . *Sci Rep* **10**, 3485 (2020).

9. Karczewski, K. & Snyder, M. Integrative omics for health and disease. *Nat Rev Genet* **19**, 299–310 (2018).

10. Hira, Z.M. & Gillies, D.F. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinformatics.*, 198363 (2015).

11. Li, L., Li, X. & Guo, Z. Efficiency of two filters for feature gene selection. *Life Sci. Res.*, 372 – 396 (2003).

12. Park, P.J., Pagano, M. & Bonetti, M. A nonparametric scoring algorithm for identifying informative genes from microarray data. . *Pac Symp Biocomput.*, 52-63 (2001).

13. Kohavi, R. & John, G.H. Wrappers for feature subset selection. *Artif. Intell* **97**, 273 – 324 (1997).

14. Pirgazi, J., Alimoradi, M., Esmaeili Abharian, T. & Hossein Olyaee, M. An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets. *Sci Rep* **9**, 18580 (2019).

15. Ding, H. & Li, D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino acids* **47**, 329–333 (2015).

16. Zou, Q., Zeng, J., Cao, L. & Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* **173**, 346–354 (2016).

17. Le, N.-Q.-K., Nguyen, T.-T.-D. & Ou, Y.-Y. Identifying the molecular functions of electron transport proteins using radial basis function networks and biochemical properties. *Journal of Molecular Graphics and Modelling* **73**, 166-178 (2017).

18. Hall, M.A. Correlation-based feature selection for machine learning. *The University of Waikato* (1999).

19. Bermejo, P., Gámez, J.A. & Puerta, J.M. A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets. *Pattern Recognition Letters* **32**, 701–711 (2011).

20. Shukla, A.K., Singh, P. & Vardhan, M. A hybrid framework for optimal feature subset selection. *Journal of Intelligent & Fuzzy Systems* **36**, 2247–2259 (2019).

21. 11 Machine learning approaches to genomics. Nature (2019). *Nature* (2019).

22. Choi, J., Park, S., Yoon, Y. & Ahn, J. Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers. *Bioinformatics* **33**, 3619-3626 (2017).

23. Tabl, A.A., Alkhateeb, A., ElMaraghy, W., Rueda, L. & Ngom, A. A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. *Frontiers in Genetics* **10** (2019).

24. Koumakis, L. Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal* **18**, 1466-1473 (2020).

25. Libbrecht, M. & Noble, W. Machine learning applications in genetics and genomics. . *Nat Rev Genet* **16**, 321–332 (2015).

26. Bayram, E., Caldwell, J.Z.K. & Banks, S.J. Current understanding of magnetic resonance imaging biomarkers and memory in Alzheimer's disease. *Alzheimers Dement (N Y)* **4**, 395-413 (2018).

27. Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A. & Davatzikos, C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *Neuroimage* **155**, 530-548 (2017).

28. Scheubert, L., Luštrek, M., Schmidt, R., Repsilber, D. & Fuellen, G. Tissue-based Alzheimer gene expression markers–comparison of multiple machine learning

approaches and investigation of redundancy in small biomarker sets. *BMC Bioinformatics* **13**, 266 (2012).

29. Ricciarelli, R. et al. Microarray analysis in Alzheimer's disease and normal aging. *IUBMB Life* **56**, 349-354 (2004).
30. Bringay, S. et al. Discovering novelty in sequential patterns: application for analysis of microarray data on Alzheimer disease. *Stud Health Technol Inform* **160**, 1314-1318 (2010).
31. Kong, W. et al. Independent component analysis of Alzheimer's DNA microarray gene expression data. *Molecular Neurodegeneration* **4**, 5 (2009).
32. Martínez-Ballesteros, M., García-Heredia, J.M., Nepomuceno-Chamorro, I.A. & Riquelme-Santos, J.C. Machine learning techniques to discover genes with potential prognosis role in Alzheimer’s disease using different biological sources. *Information Fusion* **36**, 114-129 (2017).
33. Pirooznia, M., Yang, J.Y., Yang, M.Q. & Deng, Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* **9**, S13 (2008).
34. Park, C., Ha, J. & Park, S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. . *Expert Systems with Applications* **140**, 112873–112873 (2020).
35. Chen, H., He, Y., Ji, J. & Shi, Y. A Machine Learning Method for Identifying Critical Interactions Between Gene Pairs in Alzheimer's Disease Prediction. *Frontiers in Neurology* **10** (2019).
36. Salat, D.H., Kaye, J.A. & Janowsky, J.S. Selective Preservation and Degeneration Within the Prefrontal Cortex in Aging and Alzheimer Disease. *Archives of Neurology* **58**, 1403-1408 (2001).
37. Van Someren, E.J.W. et al. Medial temporal lobe atrophy relates more strongly to sleep-wake rhythm fragmentation than to age or any other known risk. *Neurobiology of Learning and Memory* **160**, 132-138 (2019).
38. Mu, Y. & Gage, F.H. Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Mol Neurodegener* **6**, 85 (2011).
39. Warde-Farley, D. et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* **38**, W214-220 (2010).
40. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-d613 (2019).
41. Fajarda, O., Duarte-Pereira, S., Silva, R.M. & Oliveira, J.L. Merging microarray studies to identify a common gene expression signature to several structural heart diseases. *BioData Mining* **13**, 8 (2020).
42. Cheadle, C., Vawter, M.P., Freed, W.J. & Becker, K.G. Analysis of microarray data using Z score transformation. *J Mol Diagn* **5**, 73-81 (2003).
43. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3 (2006).
44. Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning. New York: Springer. (2001).
45. Breiman, L. Random forests. Machine Learning. **45**, 5–32 (2001).
46. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* **16**, 199-231, 133 (2001).
47. Breiman, L. Bagging predictors. Machine Learning. **24**, 123–140 (1996).

48. Diaz-Uriarte, R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics* **8** (2007).

49. Man, M.Z., Dyson, G., Johnson, K. & Liao, B. Evaluating Methods for Classifying Expression Data. *Journal of Biopharmaceutical Statistics* **14**, 1065-1084 (2004).

50. Wu, B. et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* **19**, 1636-1643 (2003).

51. Izmirlian, G. Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. *Ann N Y Acad Sci* **1020**, 154-174 (2004).

52. Alvarez, S. et al. A predictor based on the somatic genomic changes of the BRCA1/BRCA2 breast cancer tumors identifies the non-BRCA1/BRCA2 tumors with BRCA1 promoter hypermethylation. *Clin Cancer Res* **11**, 1146-1153 (2005).

53. Liaw, A. & Wiener, M. Classification and regression by randomForest. *Rnews* **2**, 18–22 (2002).

54. Robert, T. Regression shrinkage and selection via the lasso: a retrospective. *J. R. Statist. Soc. B* **73**, 273–282 (2011).

55. Klau, S., Jurinovic, V., Hornung, R., Herold, T. & Boulesteix, A.-L. Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics* **19**, 322 (2018).

56. Deutelmoser, H. et al. Robust Huber-LASSO for improved prediction of protein, metabolite and gene expression levels relying on individual genotype data. *Briefings in Bioinformatics* (2020).

57. Ghosh Roy, G., Geard, N., Verspoor, K. & He, S. PoLoBag: Polynomial Lasso Bagging for signed gene regulatory network inference from expression data. *Bioinformatics* **36**, 5187-5193 (2020).

58. Hua, J., Liu, H., Zhang, B. & Jin, S. LAK: Lasso and K-Means Based Single-Cell RNA-Seq Data Clustering Analysis. *IEEE Access* **8**, 129679-129688 (2020).

59. Hui, Z. & Trevor, H. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, . 301–320 (2005).

60. Ma, S., Song, X. & Huang, J. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* **8**, 60 (2007).

61. Kuhn, M. caret: Classification and Regression Training. (2020).

62. Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488-492 (2005).

63. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171-178 (2005).

64. Somorjai, R.L., Dolenko, B. & Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* **19**, 1484-1491 (2003).

65. Pan, K.H., Lih, C.J. & Cohen, S.N. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc Natl Acad Sci U S A* **102**, 8961-8965 (2005).

66. Meyer , D., Dimitriadou , E., Hornik , K., Weingessel, A. & Leisch, F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. (2020).

67. Diez, L. & Wegmann, S. Nuclear Transport Deficits in Tau-Related Neurodegenerative Diseases. *Frontiers in Neurology* **11** (2020).

68. Baker , D.J. et al. Early aging–associated phenotypes in Bub3/Rae1 haploinsufficient mice. *Journal of Cell Biology* **172**, 529-540 (2006).

69. Muscarella, L.A. et al. Small deletion at the 7q21.2 locus in a CCM family detected by real-time quantitative PCR. *J Biomed Biotechnol* **2010** (2010).

70. Abrams, A.J. et al. Mutations in SLC25A46, encoding a UGO1-like protein, cause an optic atrophy spectrum disorder. *Nat Genet* **47**, 926-932 (2015).

71. Bitetto, G. et al. SLC25A46 mutations in patients with Parkinson's Disease and optic atrophy. *Parkinsonism Relat Disord* **74**, 1-5 (2020).

72. Wan, J. et al. Loss of function of SLC25A46 causes lethal congenital pontocerebellar hypoplasia. *Brain* **139**, 2877-2890 (2016).

73. Schmidt, R. et al. [Sex differences in Alzheimer's disease]. *Neuropsychiatr* **22**, 1-15 (2008).

74. Martín-Maestro, P. et al. Slower Dynamics and Aged Mitochondria in Sporadic Alzheimer's Disease. *Oxid Med Cell Longev* **2017**, 9302761 (2017).

75. Desikan, R.S. et al. Polygenic Overlap Between C-Reactive Protein, Plasma Lipids, and Alzheimer Disease. *Circulation* **131**, 2061-2069 (2015).

76. Lu, A.T. et al. Genetic architecture of epigenetic and neuronal ageing rates in human brain regions. *Nat Commun* **8**, 15353 (2017).

77. Yan, T., Ding, F. & Zhao, Y. Integrated identification of key genes and pathways in Alzheimer’s disease via comprehensive bioinformatical analyses. *Hereditas* **156**, 25 (2019).

78. Vargas, D.M., De Bastiani, M.A., Zimmer, E.R. & Klamt, F. Alzheimer’s disease master regulators analysis: search for potential molecular targets and drug repositioning candidates. *Alzheimer's Research & Therapy* **10**, 59 (2018).

79. Tanzi, R.E. The genetics of Alzheimer disease. *Cold Spring Harb Perspect Med* **2** (2012).

80. Simonsen, A.H., Hagnelius, N.O., Waldemar, G., Nilsson, T.K. & McGuire, J. Protein markers for the differential diagnosis of vascular dementia and Alzheimer's disease. *Int J Proteomics* **2012**, 824024 (2012).

81. Jagadeesh, A., Maroun, L.E., Van Es, L.M. & Millis, R.M. Autoimmune Mechanisms of Interferon Hypersensitivity and Neurodegenerative Diseases: Down Syndrome. *Autoimmune Diseases* **2020**, 6876920 (2020).

82. Al Shweiki, M.R. et al. Cerebrospinal Fluid Levels of Prodynorphin-Derived Peptides are Decreased in Huntington's Disease. *Movement Disorders* **36**, 492-497 (2021).

83. Shukla, V., Skuntz, S. & Pant, H.C. Deregulated Cdk5 activity is involved in inducing Alzheimer's disease. *Arch Med Res* **43**, 655-662 (2012).

84. Hullinger, R. & Puglielli, L. Molecular and cellular aspects of age-related cognitive decline and Alzheimer's disease. *Behav Brain Res* **322**, 191-205 (2017).

85. Stygelbout, V. et al. Inositol trisphosphate 3-kinase B is increased in human Alzheimer brain and exacerbates mouse Alzheimer pathology. *Brain* **137**, 537-552 (2014).

86. Garwain, O., Valla, K. & Scarlata, S. Phospholipase Cb1 regulates proliferation of neuronal cells. *The FASEB Journal* **32**, 2891-2898 (2018).

87. Seminara, R.S. et al. The Neurocognitive Effects of Ghrelin-induced Signaling on the Hippocampus: A Promising Approach to Alzheimer's Disease. *Cureus* **10**, e3285 (2018).

88. Jones, S.V. & Kounatidis, I. Nuclear Factor-Kappa B and Alzheimer Disease, Unifying Genetic and Environmental Risk Factors from Cell to Humans. *Front Immunol* **8**, 1805 (2017).

89. Jiang, Y. et al. in Med Sci Monit, Vol. 24 5635-5644 (2018).

90. Luo, H., Han, L. & Xu, J. Apelin/APJ system: A novel promising target for neurodegenerative diseases. *Journal of Cellular Physiology* **235**, 638-657 (2020).

Supplementary for:

A Machine Learning Approach to Unmask Novel Gene Signatures and Prediction of Alzheimer's Disease Within Different Brain Regions.

Abhibhav Sharma[1], Pinki Dey[2*]

1. School of Computer and System Sciences, Jawaharlal Nehru University, New Delhi110067, India
2. School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney 2033, Australia

Contents

Machine Learning Model descriptions:

Random Forest (RF)

Given a training dataset, $L = \{(X_i, Y_i)_{i=1}^{N} \mid X_i \in \mathrm{R}^M, Y \in \{1,2,\dots,c\}\}$, where X_i represents the variables or the feature set and Y denotes the corresponding label (class response variable). The number of training samples and features are denoted as N and M respectively. The random forest model (RF) is delineated below. For a given input X , let the prediction of the tree T_k is denoted by $\widehat{Y^k}$. The random forest amalgamating K trees have the prediction given as:

$$\hat{Y} = majority\ vote\ \{\widehat{Y^k}\}_1^K$$

Algorithm [1]
Input: The training dataset $L = \{(X_i, Y_i)_{i=1}^{N} \mid X_i \in \mathrm{R}^M, Y \in \{1,2,\dots,c\}\}$
K: $the\ number\ of\ tress,$
$mtry$: $the\ size\ of\ the\ subspace$
Output: A random Forest

```
a) For $k \rightarrow 1\ to\ K\ do$
b) $\mathrm{L}_k$ samples as a bagged subset are drawn from L
c) While (stopping condition is not met) do
```

d) $mtry$ features are randomly selected.
e) For $m \rightarrow 1\ to\ \|mtry\|$ do
f) The decline in the node impurity is computed
g) The features that contribute the most in decreasing the impurity is chosen.
h) The node is then branched/divided into two children nodes
i) K trees are combined to produce a random forest

As the trees are grown from a bagged sample set, only a proportion of samples were leveraged to grow the tree also called *in-bag* samples. A small proportion of instance that is left out is called *out-of-bag* (OOB) samples that are employed to estimate the rate of prediction error called OOB error rate.

The OOB predicted value is given as:
$\hat{Y}^{OOB} = (\frac{1}{\|\theta_{i'}\|}) \sum_{k \epsilon \theta_{i'}} \hat{Y}^{k}$, where $\theta_{i'} = \frac{\text{L}}{\theta_i}$, i’ and *i* denotes the out-of-bag and in-bag sampled instances, $\|\theta_{i'}\|$ is the cardinality/size of OOB instances, and the OOB prediction error is

$$\widehat{Err}^{OOB} = \frac{1}{N_{OOB}} \sum_{i=1}^{N_{OOB}} \Psi\left(\text{Y}, \hat{Y}^{OOB}\right)$$

Here $\Psi(.)$ is the error function and N_{OOB} is OOB sample’s size.

Support Vector Machine (SVM)

An SVM classifier identifies and maximizes the most optimal hyperplane that separates the data points of each type of label (category). In a simple SVM model, the optimal hyperplane is evaluated on the basis of the distance between the support vectors [8]–[10] Once the hyperplane is evaluated using train data points, SVM allocates the new instances to a class based on its relative nearness from the trained data points [11]. For a given set of data points $(x_i, y_i), i = 1,2, \ldots.., m$ where $x\ \epsilon\ \text{R}^n$, $y\ \epsilon\ \text{R}$. Given a set of weight $\boldsymbol{w}$**,** The optimal hyperplane H is:

$$(\boldsymbol{w}.x) + b = 0$$

SVM classifier follows the constraints:

$$y_i[\boldsymbol{w}.x_i + b] \geq 1$$

The optimization problem to minimize $\boldsymbol{w}$ (or maximize $2/\|\boldsymbol{w}\|$) is solved using a Lagrange function eq3:

$$\text{L}(\boldsymbol{w}, b, a) = \frac{1}{2\|\boldsymbol{w}\|} - \lambda\Big(y\big((\boldsymbol{w}.x) + b\big) - 1\Big);\ \lambda_i > 0$$

Here the λ is a Lagrange multiplier. Solving the partial derivatives for w and b to 0, the optimal hyperplane is built as:

$$y(x) = sign\left[\sum_{i=1}^{m} \lambda_i y_i x_i^T x + b\right]$$

Several genomics studies have employed variations of SVM models as a classifier and retained excellent performance [12][13].

Multiplicity Problem

For microarray dataset, different wrapper feature selection models identify a varying set of candidate genes as the important signature based on the prediction accuracy attained by the gene subset [2][3] This leads to the problem of multiplicity, especially for the case when the motivation is not only the prediction but also the identification of biologically relevant gene signatures[4][5] . This variation or the lack of uniqueness could be reasoned as different in the patient batch, differing analysis and varying technologies. Studies indicate that the difference in the gene subset is strongly influenced by the cohort that have been used for gene selection[3]. This problem has also been elaborated and discussed extensively in recent studies that too indicated the extremely small ratio of samples to genes in the microarray dataset is the most likely cause of this problem[6][7]. Unfortunately, this issue casts a false sense of trust in the results obtained by most studies falling under this paradigm of gene identification through wrapper approach. Subscribing to the notion of the studies investigating the problem of multiplicity, we lend credence to the combined set of genes that were obtained by both the methods (varSelRF and LASSO); and exclusively probed the biological significance of the common and repeatedly selected gene candidates.

Figures

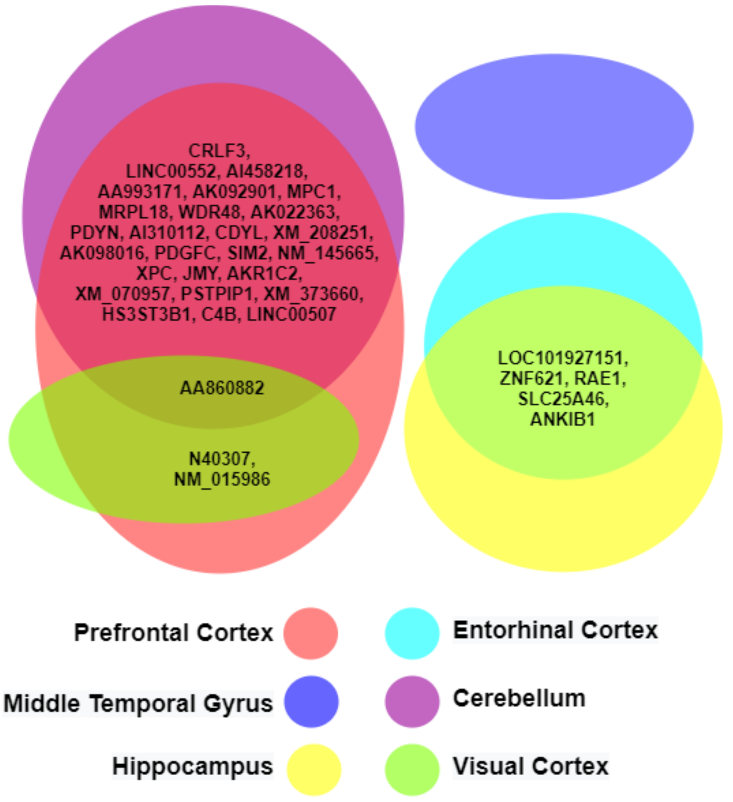


Figure S1. The common genes identified by the machine learning models within the different brain regions.

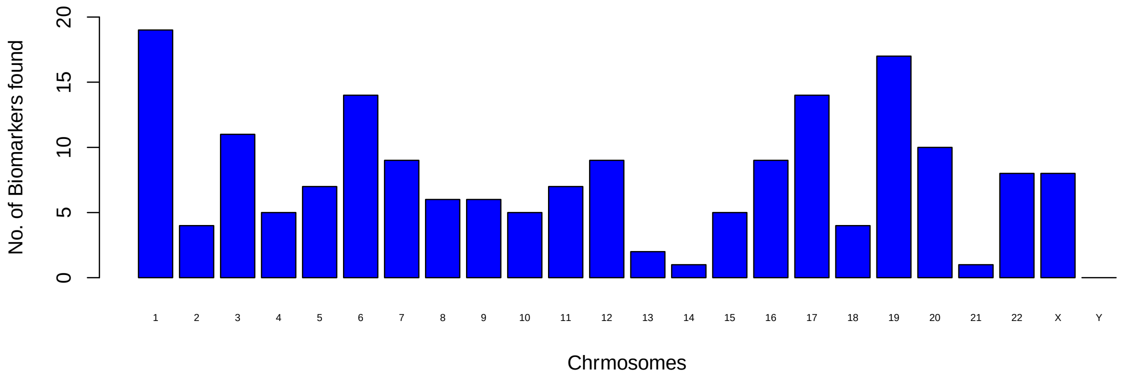


Figure S2. The number of biomarkers found in all the human chromosomes by our models.

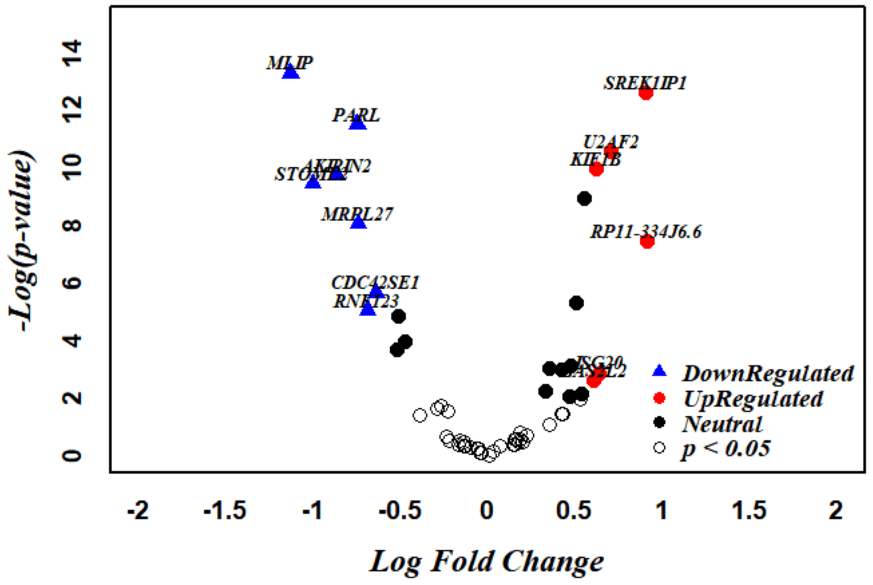


Figure S3. The up and down regulated biomarkers in the AD samples with respect to the normal ones identified from the GSE5281 expression data.

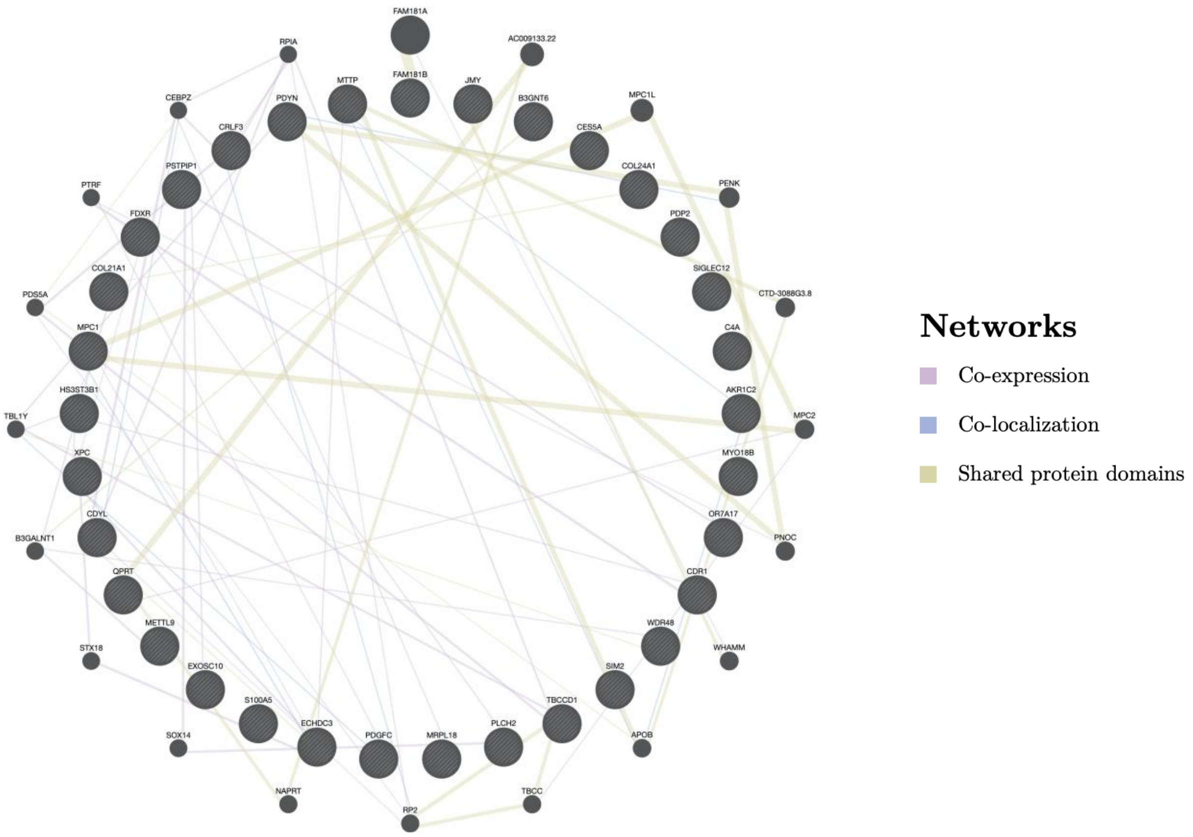


Figure S4. The GeneMania network analysis for all the biomarkers of PFC brain region. The biomarkers are shown by the stripped black circles. The solid-coloured lines represent the type of interactions in the biomarkers.

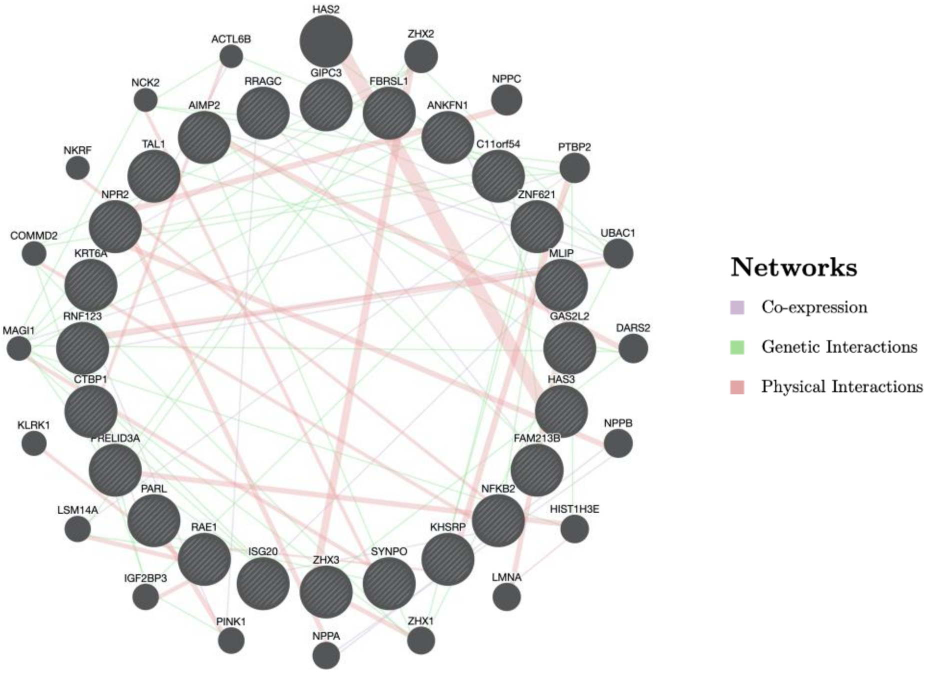


Figure S5. The GeneMania network analysis for all the biomarkers of EC brain region. The biomarkers are shown by the stripped black circles. The solid-coloured lines represent the type of interactions found within the biomarkers.

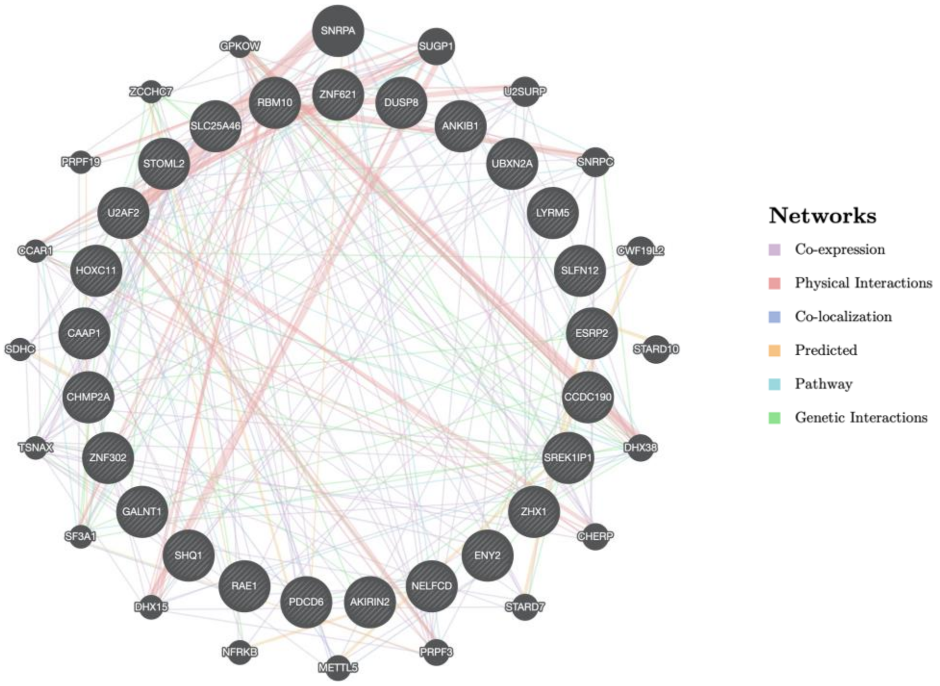


Figure S6. The GeneMania network analysis for all the biomarkers of H brain region. The biomarkers are shown by the stripped black circles. The solid-coloured lines represent the type of interactions found in the biomarkers.

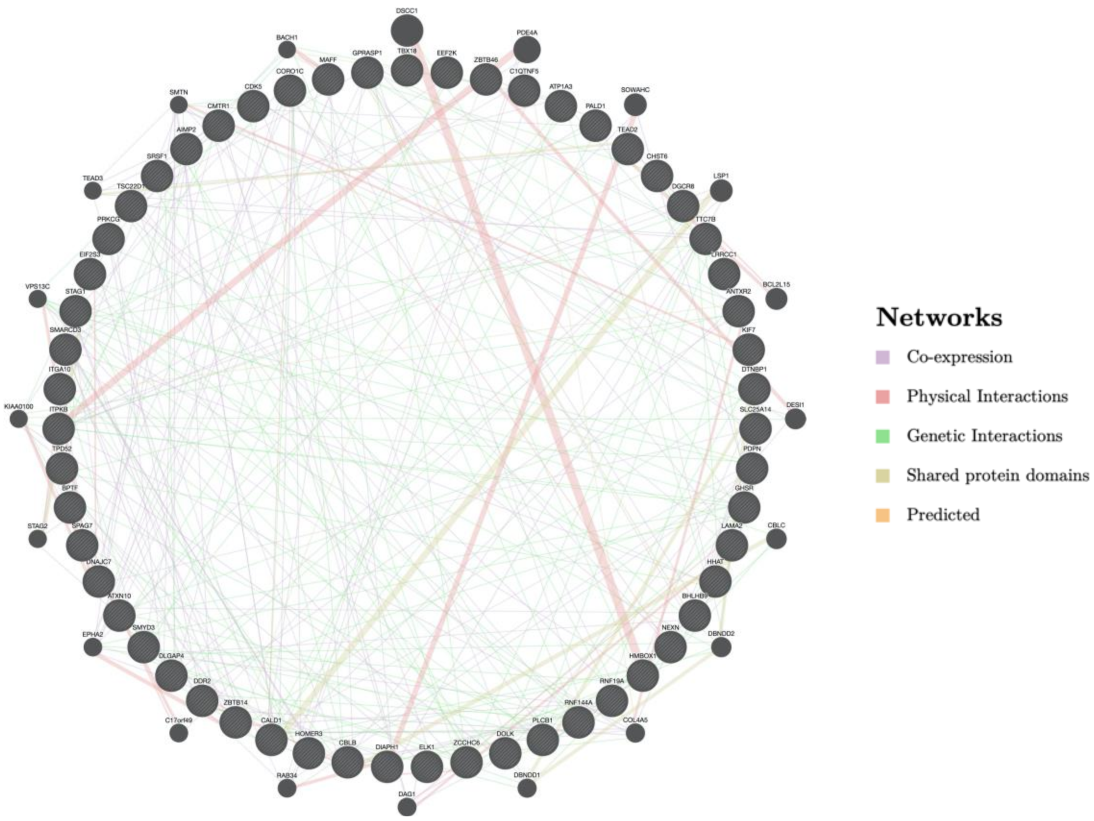


Figure S7. The GeneMania network analysis for all the biomarkers of MTG brain region. The biomarkers are shown by the stripped black circles. The solid-coloured lines represent the type of interactions found in the biomarkers.

References

[1] T. T. Nguyen, J. Z. Huang, and T. T. Nguyen, "Unbiased feature selection in learning random forests for high-dimensional data," *Sci. World J.*, 2015, doi: 10.1155/2015/471371.

[2] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: A multiple random validation strategy," *Lancet*, 2005, doi: 10.1016/S0140-6736(05)17866-0.

[3] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: Is there a unique set?," *Bioinformatics*, 2005, doi: 10.1093/bioinformatics/bth469.

[4] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: Curses, caveats, cautions," *Bioinformatics*, 2003, doi: 10.1093/bioinformatics/btg182.

[5] K. H. Pan, C. J. Lih, and S. N. Cohen, "Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays," *Proc. Natl. Acad. Sci. U. S. A.*, 2005, doi: 10.1073/pnas.0502674102.

[6] E. Walker, "Regression Modeling Strategies," *Technometrics*, 2003, doi: 10.1198/tech.2003.s158.

[7] L. Breiman, "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," *Stat. Sci.*, 2001, doi: 10.1214/ss/1009213726.

[8] V. Vapnik, S. E. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in *Advances in Neural Information Processing Systems*, 1997.

[9] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Trans. Neural Networks*, 1999, doi: 10.1109/72.788645.

[10] V. Vapnik and O. Chapelle, "Bounds on error expectation for support vector machines," *Neural Comput.*, 2000, doi: 10.1162/089976600300015042.

[11] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*. 1999, doi: 10.1109/72.788640.

[12] C. Devi Arockia Vanitha, D. Devaraj, and M. Venkatesulu, "Gene expression data classification using Support Vector Machine and mutual information-based gene selection," in *Procedia Computer Science*, 2014, doi: 10.1016/j.procs.2015.03.178.

[13] M. P. S. Brown *et al.*, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Natl. Acad. Sci. U. S. A.*, 2000, doi: 10.1073/pnas.97.1.262.

Tables

Dataset	Dataset
GSE5281	The signal value calculated by MAS 5 or GCOS software.
GSE48350	GC-RMA normalized expression values
GSE4757	The signal value calculated by MAS 5 or GCOS software.
GSE28146	MAS5-calculated Signal intensity
GSE118553	Normalized signal
GSE132903	Normalized (log2 scale) with the lumiExpresso function (R-package Lumi)
GSE33000	Normalized log10 ratio (Cy5/Cy3) representing test/reference
GSE44770	Normalized log10 ratio (Cy5/Cy3) representing test/reference
GSE44771	Normalized log10 ratio (Cy5/Cy3) representing test/reference
GSE44768	Normalized log10 ratio (Cy5/Cy3) representing test/reference

Table S1: Summary of experimental designs and measurements of the gene expression datasets used in our study.

Model	Critical Parameters	Note

varSelRF	ntree = 5000,
	ntreeIterat = 2000,
	vars.drop.frac = 0.2
		After tuning the default values remained the best value

LASSO	method = "glmnet",
	lambda= seq(0.0001, 1, length = 5)
		The alpha is not declared, setting 0 by default thus performing LASSO

RandomForest	ntree=500,
	mtry= max((number of gene) / 3 , 1)
		ntree employed here is the default value which was tested against 250, 300 and 400.

Elastic Net	method = "glmnet",
	alpha= seq(0,1,length=10),
	lambda= seq(0.0001, 1, length = 5)
		Both aplha and beta were searched within the given possible set

SVM	kernel= "radial",
	degree = 7
		Degree of 7 is compared against 2 to 6 and remained the best suiting for the profiles

Table S2: The tuned hyperparameter sets for the different models used in our machine learning workflow.

5Fold CV	**Visual Cortex**					**Cerebellum**				
	Acc	**Sen**	**Spe**	**Pre**	**Mcc**	**Acc**	**Sen**	**Spe**	**Pre**	**Mcc**
Fold1_SVM	0.92	0.91	0.94	0.95	0.84	0.87	0.75	1.00	1.00	0.77
Fold1_RF	0.97	1.00	0.94	0.96	0.95	0.79	0.80	0.79	0.80	0.59
Fold1_EN	0.97	1.00	0.94	0.96	0.95	0.92	0.85	1.00	1.00	0.86
Fold2_SVM	0.97	1.00	0.94	0.96	0.95	0.92	0.95	0.88	0.91	0.84
Fold2_RF	0.92	0.95	0.88	0.91	0.84	0.90	0.91	0.88	0.91	0.79
Fold2_EN	0.97	0.95	1.00	1.00	0.95	0.97	0.95	1.00	1.00	0.95
Fold3_SVM	0.95	0.91	1.00	1.00	0.90	0.90	0.94	0.86	0.85	0.80
Fold3_RF	0.95	0.91	1.00	1.00	0.90	0.79	0.89	0.71	0.73	0.61
Fold3_EN	0.95	0.91	1.00	1.00	0.90	0.92	0.89	0.95	0.94	0.85
Fold4_SVM	0.85	0.87	0.81	0.87	0.68	0.90	0.90	0.89	0.90	0.79
Fold4_RF	0.85	0.87	0.81	0.87	0.68	0.85	0.86	0.83	0.86	0.69
Fold4_EN	0.85	0.87	0.81	0.87	0.68	0.92	0.95	0.89	0.91	0.85
Fold5_SVM	0.87	0.94	0.81	0.81	0.75	0.85	0.85	0.84	0.85	0.69
Fold5_RF	0.90	0.94	0.86	0.85	0.80	0.82	0.90	0.74	0.78	0.65
Fold5_EN	0.87	0.94	0.81	0.81	0.75	0.79	0.80	0.79	0.80	0.59
Avg_SVM	0.91	0.93	0.90	0.92	0.83	0.89	0.88	0.89	0.90	0.78
Avg_RF	0.92	0.94	0.90	0.92	0.83	0.83	0.87	0.79	0.82	0.66
Avg_EN	0.92	0.94	0.91	0.93	0.85	0.91	0.89	0.93	0.93	0.82

Table S3: Alzheimer's disease classification performance of the PFC biomarkers on VC and CR gene expression datasets.

Visual Cortex	**Cerebellum**
N40307, NM_015986, AA860882	CRLF3, LINC00552, AI458218, AA993171, AK092901, MPC1, MRPL18, WDR48, AK022363, PDYN, AI310112, CDYL, XM_208251, AK098016, PDGFC, SIM2, NM_145665, XPC, JMY, AKR1C2, AA860882, XM_070957, PSTPIP1, XM_373660, HS3ST3B1, C4B, LINC00507
Common Marker	CRLF3, AA860882

Table S4: The assessment metric obtained for VC and CR validating the AD prediction potential of the gene biomarker subset of PFC.

GeneBank Accession No	Gene Symbol	Gene name	Genomic Sequence Region	Gene type
NM_015899	PLEKHA8P1	pleckstrin homology domain containing A8 pseudogene 1	**Chromosome 12 - NC_000012.12**	pseudo
XM_208773	LOC283664			
N40307				
AL359211	METTL9	methyltransferase like 9	**Chromosome 16 - NC_000016.10**	protein coding
AB037769	PDP2	pyruvate dehyrogenase phosphatase catalytic subunit 2	**Chromosome 16 - NC_000016.10**	protein coding
CA388904				
NM_015986	CRLF3	cytokine receptor like factor 3	**Chromosome 17 - NC_000017.11**	protein coding
NM_018138	TBCCD1	TBCC domain containing 1	**Chromosome 3 - NC_000003.12**	protein coding
AK057435	LINC00552	long intergenic non-protein coding RNA 552	**Chromosome 13 - NC_000013.11**	ncRNA
AI187365				
AI458218				
NM_030820	COL21A1	collagen type XXI alpha 1 chain	**Chromosome 6 - NC_000006.12**	protein coding
NM_152269	MTRFR	mitochondrial translation release factor in rescue	**Chromosome 12 - NC_000012.12**	protein coding
BC021699				
AA993171				
NM_018543				
AK092901				
NM_016098	MPC1	mitochondrial pyruvate carrier 1	**Chromosome 6 - NC_000006.12**	protein coding
NM_014161	MRPL18	mitochondrial ribosomal protein L18	**Chromosome 6 - NC_000006.12**	protein coding
NM_020839	WDR48	WD repeat domain 48	**Chromosome 3 - NC_000003.12**	protein coding
NM_000253	MTTP	microsomal triglyceride transfer protein	**Chromosome 4 - NC_000004.12**	protein coding
NM_014298	QPRT	quinolinate phosphoribosyltransferase	**Chromosome 16 - NC_000016.10**	protein coding
NM_152890	COL24A1	collagen type XXIV alpha 1 chain	**Chromosome 1 - NC_000001.11**	protein coding
NM_032608	MYO18B	myosin XVIIIB	**Chromosome 22 - NC_000022.11**	protein coding
AK022363				
NM_024411	PDYN	prodynorphin	**Chromosome 20 - NC_000020.11**	protein coding
AI310112				
NM_004824	CDYL	chromodomain Y like	**Chromosome 6 - NC_000006.12**	protein coding
NM_014638	PLCH2	phospholipase C eta 2	**Chromosome 1 - NC_000001.11**	protein coding
NM_175885	FAM181B	family with sequence similarity 181 member B	**Chromosome 11 - NC_000011.10**	protein coding
XM_208251				
AK098016				
NM_016205	PDGFC	platelet derived growth factor C	**Chromosome 4 - NC_000004.12**	protein coding
NM_005069	SIM2	SIM bHLH transcription factor 2	**Chromosome 21 - NC_000021.9**	protein coding
NM_145665				
NM_004628	XPC	XPC complex subunit, DNA damage recognition and repair factor	**Chromosome 3 - NC_000003.12**	protein coding
NM_002685	EXOSC10	exosome component 10	**Chromosome 1 - NC_000001.11**	protein coding
NM_030901	OR7A17	olfactory receptor family 7 subfamily A member 17	**Chromosome 19 - NC_000019.10**	protein coding
AX750575				
NM_024693	ECHDC3	enoyl-CoA hydratase domain containing 3	**Chromosome 10 - NC_000010.11**	protein coding
NM_053003	SIGLEC12	sialic acid binding Ig like lectin 12	**Chromosome 19 - NC_000019.10**	protein coding
NM_152405	JMY	junction mediating and regulatory protein, p53 cofactor	**Chromosome 5 - NC_000005.10**	protein coding

NM_004110	FDXR	ferredoxin reductase	**Chromosome 17 - NC_000017.11**	protein coding
NM_004065	CDR1	cerebellar degeneration related protein 1	**Chromosome X - NC_000023.11**	protein coding
NM_002962	S100A5	S100 calcium binding protein A5	**Chromosome 1 - NC_000001.11**	protein coding
NM_145024	CES5A	carboxylesterase 5A	**Chromosome 16 - NC_000016.10**	protein coding
NM_001354	AKR1C2	aldo-keto reductase family 1 member C2	**Chromosome 10 - NC_000010.11**	protein coding
NM_138706	B3GNT6	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 6	**Chromosome 11 - NC_000011.10**	protein coding
AA860882				
NM_018544				
XM_070957				
NM_003978	PSTPIP1	proline-serine-threonine phosphatase interacting protein 1	**Chromosome 15 - NC_000015.10**	protein coding
XM_373660				
AF075038				
NM_006041	HS3ST3B1	heparan sulfate-glucosamine 3-sulfotransferase 3B1	**Chromosome 17 - NC_000017.11**	protein coding
NM_000592	C4A	complement C4A (Rodgers blood group)	**Chromosome 6 - NC_000006.12**	protein coding
BC037880	LINC00507	long intergenic non-protein coding RNA 507	**Chromosome 12 - NC_000012.12**	ncRNA
AK098016				
BU615728				
NM_014325.2	CORO1C	coronin 1C	**Chromosome 12 - NC_000012.12**	protein coding
NM_003409.2	ZFP161	zinc finger and BTB domain containing 14	**Chromosome 18 - NC_000018.10**	protein coding
NM_004838.2	HOMER3	homer scaffold protein 3	**Chromosome 19 - NC_000019.10**	protein coding
XM_937430.2	LOC648377	TERF1 pseudogene 3	**Chromosome 4 - NC_000004.12**	protein coding
NM_001010925.2	ANKRD19	ankyrin repeat domain 19, pseudogene	**Chromosome 9 - NC_000009.12**	pseudo
NM_006303.3	AIMP2	aminoacyl tRNA synthetase complex interacting multifunctional protein 2	**Chromosome 7 - NC_000007.14**	protein coding
NM_001006625.1	PDPN	podoplanin	**Chromosome 1 - NC_000001.11**	protein coding
NM_033402.3	LRRCC1	leucine rich repeat and coiled-coil centrosomal protein 1	**Chromosome 8 - NC_000008.11**	protein coding
NM_001415.3	EIF2S3	eukaryotic translation initiation factor 2 subunit gamma	**Chromosome X - NC_000023.11**	protein coding
NM_015645.2	C1QTNF5	C1q and TNF related 5	**Chromosome 11 - NC_000011.10**	protein coding
NM_002739.3	PRKCG	protein kinase C gamma	**Chromosome 19 - NC_000019.10**	protein coding
NM_005219.3	DIAPH1	diaphanous related formin 1	**Chromosome 5 - NC_000005.10**	protein coding
NM_152296.3	ATP1A3	ATPase Na+/K+ transporting subunit alpha 3	**Chromosome 19 - NC_000019.10**	protein coding
XM_940631.1	LOC149069	doublecortin domain containing 2B	**Chromosome 1 - NC_000001.11**	protein coding
NM_014746.3	RNF144A	ring finger protein 144A	**Chromosome 2 - NC_000002.12**	protein coding
NM_013302.3	EEF2K	eukaryotic elongation factor 2 kinase	**Chromosome 16 - NC_000016.10**	protein coding
NM_001099411.1	GPRASP1	G protein-coupled receptor associated sorting protein 1	**Chromosome X - NC_000023.11**	protein coding
NM_018194.2	HHAT	hedgehog acyltransferase	**Chromosome 1 - NC_000001.11**	protein coding
NM_001078166.1	SFRS1	serine and arginine rich splicing factor 1	**Chromosome 17 - NC_000017.11**	protein coding
NM_001003802.1	SMARCD3	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 3	**Chromosome 7 - NC_000007.14**	protein coding
NM_013236.2	ATXN10	ataxin 10	**Chromosome 22 - NC_000022.11**	protein coding
NM_001010854.1	TTC7B	tetratricopeptide repeat domain 7B	**Chromosome 14 - NC_000014.9**	protein coding

NM_183041.1	DTNBP1	dystrobrevin binding protein 1	**Chromosome 6 - NC_000006.12**	protein coding
NM_198525.1	KIF7	kinesin family member 7	**Chromosome 15 - NC_000015.10**	protein coding
NM_014908.3	DOLK	dolichol kinase	**Chromosome 9 - NC_000009.12**	protein coding
NM_024617.2	ZCCHC6	terminal uridylyl transferase 7	**Chromosome 9 - NC_000009.12**	protein coding
NM_001025253.1	TPD52	tumor protein D52	**Chromosome 8 - NC_000008.11**	protein coding
XM_001125808.2	LOC727758	Rho associated coiled-coil containing protein kinase 1 pseudogene 1	**Chromosome 18 - NC_000018.10**	pseudo
NM_004935.2	CDK5	cyclin dependent kinase 5	**Chromosome 7 - NC_000007.14**	protein coding
NM_004122.1	GHSR	growth hormone secretagogue receptor	**Chromosome 3 - NC_000003.12**	protein coding
NM_001079823.1	LAMA2	laminin subunit alpha 2	**Chromosome 6 - NC_000006.12**	protein coding
XR_039314.1	LOC100132324	hypothetical LOC100132324	Chromosome: 20; NC_000020.10	pseudo
NM_199345.3	PI4KAP2	phosphatidylinositol 4-kinase alpha pseudogene 2	**Chromosome 22 - NC_000022.11**	pseudo
NM_001080508.1	TBX18	T-box transcription factor 18	**Chromosome 6 - NC_000006.12**	protein coding
NM_183006.2	DLGAP4	DLG associated protein 4	**Chromosome 20 - NC_000020.11**	protein coding
NM_006182.2	DDR2	discoidin domain receptor tyrosine kinase 2	**Chromosome 1 - NC_000001.11**	protein coding
NR_002144.1	LOC407835	mitogen-activated protein kinase kinase 2 pseudogene	**Chromosome 7 - NC_000007.14**	pseudo
BX097335				
AK057443				
NM_004890.2	SPAG7	sperm associated antigen 7	**Chromosome 17 - NC_000017.11**	protein coding
NM_003951.2	SLC25A14	solute carrier family 25 member 14	**Chromosome X - NC_000023.11**	protein coding
NR_026878.1	MGC12982	FOXD2 adjacent opposite strand RNA 1	**Chromosome 1 - NC_000001.11**	ncRNA
DA760637				
NM_003315.1	DNAJC7	DnaJ heat shock protein family (Hsp40) member C7	**Chromosome 17 - NC_000017.11**	protein coding
NM_015050.2	FTSJD2	cap methyltransferase 1	**Chromosome 6 - NC_000006.12**	protein coding
NM_022720.5	DGCR8	DGCR8 microprocessor complex subunit	**Chromosome 22 - NC_000022.11**	protein coding
NM_014431.1	KIAA1274	phosphatase domain containing paladin 1	**Chromosome 10 - NC_000010.11**	protein coding
NM_183419.1	RNF19A	ring finger protein 19A, RBR E3 ubiquitin protein ligase	**Chromosome 8 - NC_000008.11**	protein coding
NM_022743.1	SMYD3	SET and MYND domain containing 3	**Chromosome 1 - NC_000001.11**	protein coding
NM_012323.2	MAFF	MAF bZIP transcription factor F	**Chromosome 22 - NC_000022.11**	protein coding
NM_183422.1	TSC22D1	TSC22 domain family member 1	**Chromosome 13 - NC_000013.11**	protein coding
XM_499121				
NM_030639.1	BHLHB9	basic helix-loop-helix family member b9	**Chromosome X - NC_000023.11**	protein coding
NR_002139.1	HCG4	HLA complex group 4	**Chromosome 6 - NC_000006.12**	ncRNA
NM_025224.2	ZBTB46	zinc finger and BTB domain containing 46	**Chromosome 20 - NC_000020.11**	protein coding
NM_003637.3	ITGA10	integrin subunit alpha 10	**Chromosome 1 - NC_000001.11**	protein coding

NM_005229.2	ELK1	ETS transcription factor ELK1	**Chromosome X - NC_000023.11**	protein coding
NM_058172.3	ANTXR2	ANTXR cell adhesion molecule 2	**Chromosome 4 - NC_000004.12**	protein coding
NM_014325.2	CORO1C	coronin 1C	**Chromosome 12 - NC_000012.12**	protein coding
NM_021615.4	CHST6	carbohydrate sulfotransferase 6	**Chromosome 16 - NC_000016.10**	protein coding
NM_002221.2	ITPKB	inositol-trisphosphate 3-kinase B	**Chromosome 1 - NC_000001.11**	protein coding
NM_003598.1	TEAD2	TEA domain transcription factor 2	**Chromosome 19 - NC_000019.10**	protein coding
NM_005862.2	STAG1	stromal antigen 1	**Chromosome 3 - NC_000003.12**	protein coding
NM_144573.3	NEXN	nexilin F-actin binding protein	**Chromosome 1 - NC_000001.11**	protein coding
NM_033157.2	CALD1	caldesmon 1	**Chromosome 7 - NC_000007.14**	protein coding
NM_170662.3	CBLB	Cbl proto-oncogene B	**Chromosome 3 - NC_000003.12**	protein coding
NM_024567.2	HMBOX1	homeobox containing 1	**Chromosome 8 - NC_000008.11**	protein coding
NM_015192.2	PLCB1	phospholipase C beta 1	**Chromosome 20 - NC_000020.11**	protein coding
NM_013236.2	ATXN10	ataxin 10	**Chromosome 22 - NC_000022.11**	protein coding
NM_004459.6	BPTF	bromodomain PHD finger transcription factor	**Chromosome 17 - NC_000017.11**	protein coding
BC024732	LOC101927151	uncharacterized LOC101927151	**Chromosome 19 - NC_000019.10**	ncRNA
NM_024939	ESRP2	epithelial splicing regulatory protein 2	**Chromosome 16 - NC_000016.10**	protein coding
NM_018130	SHQ1	SHQ1, H/ACA ribonucleoprotein assembly factor	**Chromosome 3 - NC_000003.12**	protein coding
AK074366	ZNF621	zinc finger protein 621	**Chromosome 3 - NC_000003.12**	protein coding
Y11166	SNORA71B	small nucleolar RNA, H/ACA box 71B	**Chromosome 20 - NC_000020.11**	snoRNA
BG111015	UBXN2A	UBX domain protein 2A	**Chromosome 2 - NC_000002.12**	protein coding
AI907083	PDCD6	programmed cell death 6	**Chromosome 5 - NC_000005.10**	protein coding
BC000764	AKIRIN2	akirin 2	**Chromosome 6 - NC_000006.12**	protein coding
BE968576	BC062753			
NM_004420	DUSP8	dual specificity phosphatase 8	**Chromosome 11 - NC_000011.10**	protein coding
AI123518	ZHX1	zinc fingers and homeoboxes 1	**Chromosome 8 - NC_000008.11**	protein coding
NM_173829	SREK1IP1	SREK1 interacting protein 1	**Chromosome 5 - NC_000005.10**	protein coding
AW409974	RBM10	RNA binding motif protein 10	**Chromosome X - NC_000023.11**	protein coding
BC040018	C1orf110	coiled-coil domain containing 190	**Chromosome 1 - NC_000001.11**	protein coding
NM_024828	CAAP1	caspase activity and apoptosis inhibitor 1	**Chromosome 9 - NC_000009.12**	protein coding
AJ238379	NELFCD	negative elongation factor complex member C/D	**Chromosome 20 - NC_000020.11**	protein coding
BC038440	GALNT1	polypeptide N-acetylgalactosaminyltransferase 1	**Chromosome 18 - NC_000018.10**	protein coding
NM_014212	HOXC11	homeobox C11	**Chromosome 12 - NC_000012.12**	protein coding

NM_020189	ENY2	ENY2 transcription and export complex 2 subunit	**Chromosome 8 - NC_000008.11**	protein coding
BF508739	ZNF302	zinc finger protein 302	**Chromosome 19 - NC_000019.10**	protein coding
AA015609	LYRM5			
BC029890	LOC100996760	uncharacterized LOC100996760	**Chromosome 10**	protein coding
NM_007279	U2AF2	U2 small nuclear RNA auxiliary factor 2	**Chromosome 19 - NC_000019.10**	protein coding
NM_018042	SLFN12	schlafen family member 12	**Chromosome 17 - NC_000017.11**	protein coding
BC024732	LOC101927151	uncharacterized LOC101927151	**Chromosome 19 - NC_000019.10**	ncRNA
AC004472	STOML2	stomatin like 2	**Chromosome 9 - NC_000009.12**	protein coding
AW207712	CTD-2587H24.10			
AK074366	ZNF621	zinc finger protein 621	**Chromosome 3 - NC_000003.12**	protein coding
U85943	RAE1	ribonucleic acid export 1	**Chromosome 20 - NC_000020.11**	protein coding
M74089	SLC25A46	solute carrier family 25 member 46	**Chromosome 5 - NC_000005.10**	protein coding
NM_024939	ESRP2	epithelial splicing regulatory protein 2	**Chromosome 16 - NC_000016.10**	protein coding
AB037807	ANKIB1	ankyrin repeat and IBR domain containing 1	**Chromosome 7 - NC_000007.14**	protein coding
NM_014453	CHMP2A	charged multivesicular body protein 2A	**Chromosome 19 - NC_000019.10**	protein coding
AF234255	IGLJ3	immunoglobulin lambda joining 3	**Chromosome 22 - NC_000022.11**	
NM_015484	SYF2	SYF2 pre-mRNA splicing factor	**Chromosome 1 - NC_000001.11**	protein coding
NM_006553	SLMO1	PRELI domain containing 3A	**Chromosome 18 - NC_000018.10**	protein coding
AI806944	PPP1R1C	protein phosphatase 1 regulatory inhibitor subunit 1C	**Chromosome 2 - NC_000002.12**	protein coding
AB007855	ZHX3	zinc fingers and homeoboxes 3	**Chromosome 20 - NC_000020.11**	protein coding
U88964	ISG20	interferon stimulated exonuclease gene 20	**Chromosome 15 - NC_000015.10**	protein coding
NM_003563	SPOP	speckle type BTB/POZ protein	**Chromosome 17 - NC_000017.11**	protein coding
BC001777	HPCA	hippocalcin	**Chromosome 1 - NC_000001.11**	protein coding
AI363061	CMIP	c-Maf inducing protein	**Chromosome 16 - NC_000016.10**	protein coding
AI435089	GIMAP1	GTPase, IMAP family member 1	**Chromosome 7 - NC_000007.14**	protein coding
AI492175	ACAP3	ArfGAP with coiled-coil, ankyrin repeat and PH domains 3	**Chromosome 1 - NC_000001.11**	protein coding
BE855983	ACACA	acetyl-CoA carboxylase alpha	**Chromosome 17 - NC_000017.11**	protein coding
AF134979	NPCDR1	nasopharyngeal carcinoma, down-regulated 1	Chromosome: 3; NC_000003.11	ncRNA
AW183187	CDRT15	CMT1A duplicated region transcript 15	**Chromosome 17 - NC_000017.11**	protein coding
AW026465	LOC646588	uncharacterized LOC646588	**Chromosome 7 - NC_000007.14**	ncRNA
NM_004909	CSAG2	CSAG family member 2	**Chromosome X - NC_000023.11**	protein coding

AB007855	ZHX3	zinc fingers and homeoboxes 3	**Chromosome 20 - NC_000020.11**	protein coding
AV700829	C3P1	complement component 3 precursor pseudogene	**Chromosome 19 - NC_000019.10**	pseudo
BF594164	KHSRP	KH-type splicing regulatory protein	**Chromosome 19 - NC_000019.10**	protein coding
M74089	SLC25A46	solute carrier family 25 member 46	**Chromosome 5 - NC_000005.10**	protein coding
BF510709	GIPC3	GIPC PDZ domain containing family member 3	**Chromosome 19 - NC_000019.10**	protein coding
AA541622	SYNPO2	synaptopodin 2	**Chromosome 4 - NC_000004.12**	protein coding
AW593028	ANKFN1	ankyrin repeat and fibronectin type III domain containing 1	**Chromosome 17 - NC_000017.11**	protein coding
NM_139285	GAS2L2	growth arrest specific 2 like 2	**Chromosome 17 - NC_000017.11**	protein coding
BF438330	AL110181			
AI633559	RP3-428L16.2			
AL133267	RPLP2P1	ribosomal protein lateral stalk subunit P2 pseudogene 1	**Chromosome 6 - NC_000006.12**	pseudo
AL136729	RNF123	ring finger protein 123	**Chromosome 3 - NC_000003.12**	protein coding
AI689676	ZNF579	zinc finger protein 579	**Chromosome 19 - NC_000019.10**	protein coding
BC014149	AC017104.6			
X89271	APLNR	apelin receptor	**Chromosome 11 - NC_000011.10**	protein coding
NM_016321	RHCG	Rh family C glycoprotein	**Chromosome 15 - NC_000015.10**	protein coding
BC002844	NFKB2	nuclear factor kappa B subunit 2	**Chromosome 10 - NC_000010.11**	protein coding
NM_002315	LMO1	LIM domain only 1	**Chromosome 11 - NC_000011.10**	protein coding
BC040981	SNX32	sorting nexin 32	**Chromosome 11 - NC_000011.10**	protein coding
BE259137	ONECUT3	one cut homeobox 3	**Chromosome 19 - NC_000019.10**	protein coding
BE858453	ST6GALNAC4	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 4	**Chromosome 9 - NC_000009.12**	protein coding
AK074366	ZNF621	zinc finger protein 621	**Chromosome 3 - NC_000003.12**	protein coding
AL136774	QRICH2	glutamine rich 2	**Chromosome 17 - NC_000017.11**	protein coding
NM_152441	FBXL14	F-box and leucine rich repeat protein 14	**Chromosome 12 - NC_000012.12**	protein coding
BI768821	DUOX1	dual oxidase 1	**Chromosome 15 - NC_000015.10**	protein coding
AB037807	ANKIB1	ankyrin repeat and IBR domain containing 1	**Chromosome 7 - NC_000007.14**	protein coding
NM_022055	KCNK12	potassium two pore domain channel subfamily K member 12	**Chromosome 2 - NC_000002.12**	protein coding
W37846	C1orf50	chromosome 1 open reading frame 50	**Chromosome 1 - NC_000001.11**	protein coding
U85943	RAE1	ribonucleic acid export 1	**Chromosome 20 - NC_000020.11**	protein coding
BC024732	LOC101927151	uncharacterized LOC101927151	**Chromosome 19 - NC_000019.10**	ncRNA
NM_004053	BYSL	bystin like	**Chromosome 6 - NC_000006.12**	protein coding
AF234255	IGLJ3	immunoglobulin lambda joining 3	**Chromosome 22 - NC_000022.11**	

BG430061	CACNA1C	calcium voltage-gated channel subunit alpha1 C	**Chromosome 12 - NC_000012.12**	protein coding
X99142	KRT86	keratin 86	**Chromosome 12 - NC_000012.12**	protein coding
AI242549	MLIP	muscular LMNA interacting protein	**Chromosome 6 - NC_000006.12**	protein coding
NM_016580	PCDH12	protocadherin 12	**Chromosome 5 - NC_000005.10**	protein coding

Table S5: Bioinformatics analysis of the biomarkers found by the machine learning models.