Integrated Higher-Order Evidence-Based Framework for Prediction of Higher-Order Epistasis Interactions in Alzheimer's Disease

Fayroz F. Sherif[1], Nourhan Zayed[1], Mahmoud Fakhr[1], Manal Abdel Wahed[2], and Yasser M. Kadah[3]

Abstract— Alzheimer's disease (AD) is the most common form of dementia with strong genetic factors in which a combination of genetic variants contributes to AD risk. Discovering epistasis interactions among genetic variants is key to identifying valuable AD predictive models that allow earlier diagnosis and better prognosis for patient. Presently, AD predictive models are derived using either statistical or biological feature selection methods. Unfortunately, both approaches suffer from inherent limitations in their generalization and prediction power. This study presents a new hybrid method between these two approaches based on integrated higher-order evidence-based (IHOEB) framework. This method combines statistical and biological feature selection methods and allow computationally-efficient detection of up to 4-way epistasis models associated with AD. The new processing framework was applied to data obtained from the Alzheimer's Disease Neuroimaging Initiative database (ADNI). The classification accuracies of IHOEB 4-way models varied between (0.7410-0.7860) whereas the accuracies of statistical and biological 2-way models varied between (0.6450-0.6760) and (0.5300-0.5750) respectively. This new IHOEB framework offers a promising alternative for epistasis interactions in genome wide association studies where it allows identification of AD models that are supported by both statistical and biological analyses efficiently and at higher accuracy.

.

Keywords— Feature selection, Epistasis, SNPs, Statistical, Biological interactions, Multifactor dimensionality reduction, Alzheimer's disease, Biofilter, Genome wide association studies.

[1]*Computers and Systems Department, Electronics Research Institute, Giza 12622, Egypt*
[2]*Biomedical Engineering and Systems Department, Cairo University, Giza 12613, Egypt*
[3]*Electrical and Computer Engineering Department, King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Correspondence:
Fayroz F. Sherif
Computers and Systems Department
Electronics Research Institute (ERI)
Giza 12622, Egypt
E-mail: Fayroz_Farouk@eri.sci.eg

I. Introduction

The accessibility of high-throughput genotyping data has greatly boosted biomedical research by more accurately identifying genetic variations associated with disease risk [1]. In recent years, the search for genetic variations that control common complex diseases and description of the effects of those variations has been a challenging goal [2]. The most common type of DNA variation is the single nucleotide polymorphism (SNP) that results when a single nucleotide changes to another in the genome sequence [1]. SNPs are estimated to be presented every 300 base pair of DNA, and human genome is estimated to include roughly 11 million SNPs. SNPs are spread throughout the human genome and their effect on phenotype depends on the genomic regions where they are located. It is believed that SNPs play a critical role in the biological mechanisms underlying many complex diseases such as Alzheimer's Disease (AD) [3].

Alzheimer's disease is a complex genetic brain disease with little understood etiology identified by gradual progressive memory loss, confusion, and cognitive impairment [4]. The absence of relevant biomarkers that can identify the disease risk and progression poses a major limitation for diagnosis and prognosis of the disease. Late Onset Alzheimer's disease (LOAD) is the most common form of AD accounting for approximately 95% of all cases [5] and is considered the most prevalent cause of dementia associated with aging. LOAD is a complex disease with both genetic and environmental risks and hence, identifying the influence of genetic risks in the development of LOAD has been the focus of many recent studies. A well-known genetic risk factor for LOAD is Apolipoprotein E (APOE) genotypes at loci rs429358 and rs7412 [3]. The search for AD treatment to slow or stop the disease progression depends on identifying a set of effective biomarkers in early stages.

Genome-wide association studies (GWAS) approach offers a powerful tool to analyze the genetic profile of human disease. The main objectives of GWAS include using causal variants in early prediction of phenotype and understanding underlying biological mechanisms of the disease [6]. With this approach, it was well-established that SNP profiles describe a variety of diseases [2]. In GWAS, data from a large number of subjects are collected in a case-control study with genotyping of each with up to millions of SNPs. Many studies proposed different methods for analyzing the individual effect of each SNP to

identify the small subset of SNPs associated with a particular disease [7].

Most complex diseases like AD are caused by intricate interactions among multiple genes known as epistasis interactions. Epistasis interactions are commonly studied using two feature selection approaches. The first is a statistical approach that detects independent main and joint effects of SNPs based on the population variations in patients and control [8]. The second approach relies on the biological interactions that occur at the cellular level [9]. The latter approach detects the physical interaction between two or more biological components (e.g. DNA, RNA, proteins, enzymes, etc.) [10]. Methods based on either of the two approaches inherently introduce bias into the analysis based on their assumptions and hence will not be optimal for all cases [9]. For example, using the statistical approach alone as a filter limits SNP selection to SNPs that have a significant main effect on the phenotype, whereas applying a statistical model for prediction may be proper in the same sampled population. Also, application of that model to other populations cannot be generalized unless it is based on biological function [9]. On the other hand, using biological knowledge alone to filter SNPs will bias the analysis toward previously known interactions. Consequently, it is not possible to discover new interactions using this approach [11]. Furthermore, reporting two physically interacting molecules does not indicate which phenotypes will be affected. So, it is preferred to use biological interactions together with statistical analysis to cross-examine a given interaction's role in the phenotype.

Epistasis interaction in human genome data poses four major challenges for identifying the associations between SNPs (more than 700 thousands) and disease. The first challenge is the noisy nature of data and the similarity between nearby SNPs. The second is the expanded search space and the combinatorial nature of the analysis where the search for all SNP-SNP interactions is computationally prohibitive [12]. Furthermore, 3- and 4-way interactions are impractical to examine due to the exponential relationship between the number of tests and the order of interactions sought [6]. The third challenge is the model bias related to the statistical or biological filters used. Even when both are combined, such approaches are applied in a sequential manner where bias comes into the analysis in each of these steps. The fourth and last challenge is the missing replication or verification. The replication of statistical epistasis between studies and the biological evidence of them strongly suggest that these predictive models contribute to the disease risk. Consequently, most of AD predictive models suffer a limited success in replication, generalization, and prediction.

A study based on the statistical approach used Random Forest (RF) to identify new genes associated with neurological disorders [13]. It utilized a two-stage quality-based strategy for SNP selection within 188 neurological controls and 176 AD patients from National Alzheimer's Coordinating Center (NACC) brain bank. The prediction performance achieved with RF and two-stage RF models on AD data did not exceed 0.6320 and 0.7100 respectively. Another study obtained poor results when identifying the potential genetic variants in the Alzheimer's Disease Neuroimaging Initiative (ADNI) genotype dataset using RF [14]. An alternative decision tree based statistical method was proposed in [15] to predict AD from SNP biomarkers and clinical data with sensitivity of 0.5753 and specificity 0.5462. Label propagation (LP) method was also reported to rank SNPs in two different AD datasets with classification accuracies of 0.6039 ± 0.0300 and 0.7023 ± 0.0272 using the top two ranked SNPs and the top five ranked SNPs respectively [16]. The search for epistasis role in the disease process using biological approaches was attempted in [17] where an interaction between rs6455128(KHDRBS2) and rs7989332 (CRYL1) was identified using GWAS data (2,259 cases and 6,017 controls). Another study found that epistatic interaction between rs1049296 (P589S) in the transferring gene (TF) and rs1800562 (C282Y) in the hemochromatosis gene (HFE) has a significant association with AD risk [18].

Other studies applied both statistical and biological approaches in a sequential way to identify genetic variation with AD. A study based on ADNI dataset used Bayesian networks to identify AD genetic biomarkers within the top ten genes associated with AD as identified by GWAS [19]. They observed seven new SNP biomarkers that had significant association with AD. Another genome association study for patients carrying APOE-ε4 used random forest and enrichment analysis to detect 1058 SNPs associated with AD [20]. The study identified 27 significant functional annotations that were associated with AD in APOE-ε4 carriers. One further study used Bayesian combinatorial method (BCM) to identify pairs of SNPs that are significantly related to AD in two different datasets [7]. They restricted the analysis to SNP pairs with one of them identified from prior knowledge to be associated with AD. The comparison between BCM and logistic regression results revealed that presence of 5 and 8 SNP pairs in common between the two methods in ADRC and TGEN datasets respectively. In [17], an exhaustive genome-wide epistasis analysis was developed for AD using statistical methods followed by biological validation. This study did not explore higher order interactions in their analysis but confirmed the association of (rs6455128 and rs7989332) with AD.

In this work, we present a new processing framework that integrates both the statistical and biological approaches to derive high-order interactions in a computationally efficient manner. The new approach tries to combine the advantages of both approaches and address their limitations. The developed method is implemented to process real data from the ADNI database [21] to demonstrate the practical potential of the new approach.

II. Methodology

A. Dataset

The genome sequencing data of 434 individuals between case and control were obtained from ADNI database [21]. The ADNI database was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), and other biological markers can be combined for early prediction of AD. ADNI contained total genotypes of 730,525 SNPs for both 125 neurologically human controls and 306 Alzheimer's disease patients (cases). Among all SNPs, we limited our selection for those reference

SNPs (starting with 'rs' identifiers), cataloged on the Single Nucleotide Polymorphism database (dbSNP) [22].

B. Approach

The goal of this work is to develop a new approach to predict higher-order models called Integrated Higher-Order Evidence-Based (IHOEB) framework to overcome the challenges outlined above. First. the noisy nature of data and similarity between nearby SNPs are addressed by applying SNP quality control and linkage disequilibrium (LD) pruning. Then, the massive computational requirements for predicting higher-order interactions are significantly reduced by applying SNP-disease association ranking in which each SNP is tested against disease separately then discarding statistically insignificant SNPs. The relevant and statistically significant subset of SNPs obtained from the above two steps is then used to identify multi-SNP interaction models up to fourth-order associated with AD. Finally, the model bias including limited prediction accuracy and missing replication between statistical and biological models is minimized by integrating statistical and biological SNP filtering to detect higher-order interactions associated with AD. In the statistical analysis, statistical filters were applied to reach significant consensus dataset then a carefully selected classification technique was used to assess AD predictive models. In the biological analysis, prior biological information was utilized that filter an extensive number of SNP data to find a biologically relevant subset. Then, all possible pairs of SNPs with prior knowledge of putative epistasis are identified. Within IHOEB framework, integrated statistical and biological SNP filters provide a combined dataset. This combined dataset is used by a statistical classifier to detect higher-order AD models. A block diagram of the statistical, biological and IHOEB analyses is shown in Fig. (1).

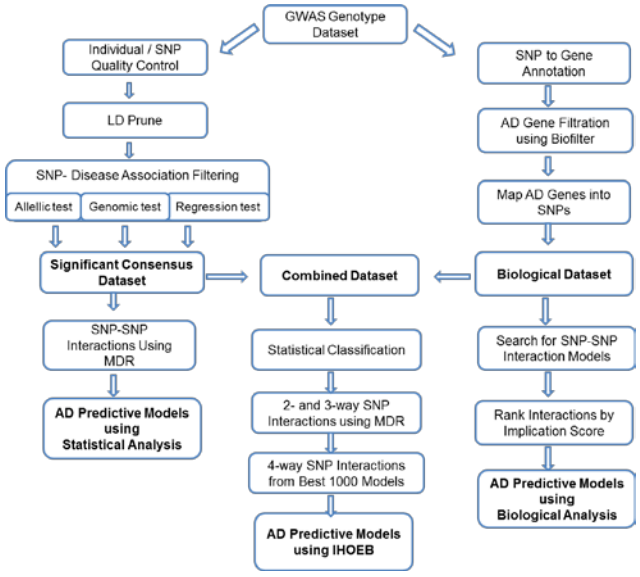

Figure (1): A block diagram of the Integrated Higher-Order Evidence-Based (IHOEB) proposed framework.

C. Statistical feature selection method

The selection of informative SNPs was done using statistical filters to reach significant consensus dataset [25]. A two-stage SNP filtering was utilized to differentiate informative SNPs from noisy one. In the first stage, SNP based quality control and LD pruning are used. PLINK, an open source package allowing a full range of basic large-scale GWAS analyses, is used to filter and exclude individual or SNP datasets that have missing rate more than 10% [23-24]. Also, SNPs with minor allele frequency less than 10% are excluded. These filtering parameters were chosen in this study due to their wide use in the literature [10]. In the second stage, SNP disease association ranking based on three statistical tests was utilized to prioritize SNPs. We compared allele or genotype frequencies at each SNP loci in the given case-control population to decide whether there is a statistical association between AD and that SNP loci. Allelic, genomic and regression tests were all used to test each SNP against disease separately and SNP with the strongest association to disease was selected. In each test, a p-value threshold of 0.05 was taken as the significance level to detect SNP associations. By intersecting the results of the three tests, a significant consensus SNP dataset is determined. In the statistical epistasis analysis, statistical filters were applied to reach significant consensus dataset then MDR used this dataset to identify 2-way interactions models associated with AD. We refer to these models here as AD predictive models using statistical analysis.

D. Biological feature selection method

Prior biological information was utilized to filter the extensive number of GWAS data records into a biological dataset with relevant biological basis without applying statistical analysis. In the last decade a huge amount of biological information became available via public databases such as gene–gene interaction databases, gene ontology annotation, pathways, and gene networks [26]. To investigate the association between genetic variations and disease, Biofilter (version 2.0) software tool, whose main functions include annotation and filtering, was used [27]. This software consults a local database called the Library of Knowledge Integration (LOKI), which stores data collected from public resources and databases including the National Center for Biotechnology (NCBI) dbSNP [28] and gene Entrez database [29], Gene Ontology (GO) [30], Protein Families database (Pfam) [31], and others [32]. LOKI utilizes three main terms for representing the different types of data. Position and region terms refer to SNP data and gene region respectively, and the term "group" is used to represent a set of regions that are linked in some way [33]. GWAS SNPs were annotated with gene and group information based on the relationships stored in the LOKI database. We filtered out all genes that do not contain at least one evidence-supported relation with AD. Then, by mapping the associated genes back to SNPs, we formed a new data subset, which we will refer to here as the biological dataset. Since combinations supported in more than one data source are more likely to be relevant, the confidence in biological interaction of multi-SNP model was assessed here using its implication score [27].

Table 1. The performance metrics of the classification methods

Classifier	Accuracy	Sensitivity	Specificity
Naive Bayes	0.6172	0.6200	0.6140
Random Forest	0.7031	0.7130	0.6930
KNN	0.6032	0.6547	0.5530
Logistic Regression	0.7100	0.7020	0.7180
SVM	0.7070	0.6810	0.7340
MDR	0.7816	0.7455	0.8214

The implication score measures of strength of evidence supporting a given model based on prior knowledge. It provides the number of sources followed by the number of groups which support this gene model [27]. We call the models obtained using the above procedure as AD predictive models using biological analysis.

E. Higher-Order Model Prediction

The developed IHOEB framework works to identify AD models that were replicated in both statistical and biological analyses with higher accuracy. Multiple statistical and data mining methods were used to investigate multi-locus interactions in AD. In this framework, we integrated statistical and biological SNP filters to form the combined dataset. A total of six popular classifiers were tested for utilization to build AD prediction models based on this combined dataset. These classification techniques include Support vector machine (SVM) [34], naïve Bayes [19, 35], k-nearest neighbor (KNN), random forest (RF) [36], logistic regression (LR) [37], and multifactor dimensionality reduction (MDR) [38] classifiers. These algorithms (except MDR) were implemented in Python whereas the open-source MDR software package was used for MDR [39]. In our experiments, KNN used 5 nearest neighbors and Euclidean distance metric, SVM classifier used radial basis function kernel, RF included 100 trees and a minimal number of instances in a leaf of 5, and LR used least squares for regularization. The predictive performance of these algorithms were compared in terms of classification accuracy, sensitivity, and specificity to select the best method that would be utilized in IHOEB proposed framework. We evaluated the performance of using 10-fold cross validation. The performance of these classifiers shown in Table 1 indicated that MDR achieved the highest accuracy, sensitivity and specificity compared to the other techniques and hence will be the classification method of choice in this work.

The selected classification method (MDR) will be used to evaluate the predictive accuracy of 2-, 3-, and 4-SNP interaction models. Since searching all 4-way models within the combined dataset is computationally prohibitive, we limit our search to the interaction of the 1000 best 3-way models with any other of the SNPs included in the combined set to evaluate 4-way combinations. This procedure significantly reduces the computational complexity while maintaining potential to find good models. The deduced 4-way interaction models are called AD predictive models using IHOEB.

To account for data imbalance, the assessment metric was chosen here to be the balanced accuracy (BA) metric, which is the average of the sensitivity and specificity, shown to outperform the traditional measure of classification accuracy [42].

III. Results and Discussion

The new approach outlined in the block diagram of Fig. (1) was implemented and applied to the ADNI database. In the first stage of statistical feature selection, the whole dataset (730,525 SNPs) was examined for quality control using PLINK. Two subsequent quality control procedures were used for individual and marker. A total of 431 individuals passed the quality control filtering. About 9.3% reduction in the count of markers was obtained, leaving 662,630 SNPs after quality control. LD pruning further reduced the redundancy of signals and decreased the level of similarity between one or more nearby SNPs. These highly correlated SNPs were excluded from the dataset leaving number of markers at 464,168 SNPs. Then, we applied three different SNP–disease association tests implemented in PLINK to assess the statistical association of each SNP with the disease. Each of the total of 464,168 SNPs was examined for association with the disease in the three tests independently. The results of each of the three tests were ranked and the non-significant SNPs with p-value greater than 0.05 were discarded. Subsequently, the SNP markers were reduced to 31,333, 35,275 and 20,197 SNPs using allelic, genomic and regression tests respectively.

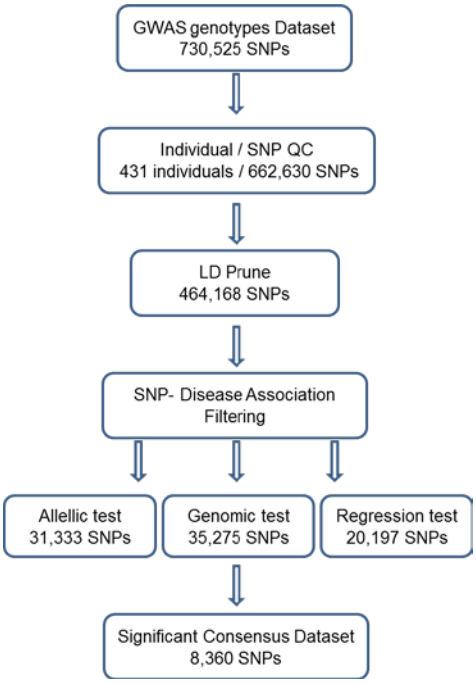

Figure (2): Results of statistical epistasis with two-stage SNP filtering and SNP-disease association ranking, indicating the number of SNPs in each stage.

Table 2. The top ten 2-way interaction models using MDR in statistical analysis

	Model	BA Training	BA Testing	p-Value
1	rs1542176, rs4955669	0.6763	0.6763	3.66E-3
2	rs7650925, rs3765121	0.6726	0.6645	6.70E-3
3	rs7650925, rs3800908	0.6690	0.6690	3.00E-3
4	rs7650925, rs9862078	0.6663	0.6712	2.97E-3
5	rs17021105, rs11155266	0.6602	0.6530	2.56E-3
6	rs182009, rs17021105	0.6587	0.6547	4.00E-3
7	rs17073734, rs942666	0.6547	0.6634	7.68E-3
8	rs878698, rs7799696	0.6491	0.6519	8.20E-3
9	rs1317902, rs13013095	0.6463	0.6427	1.57E-3
10	rs6546056, rs2798224	0.6450	0.6450	2.00E-3

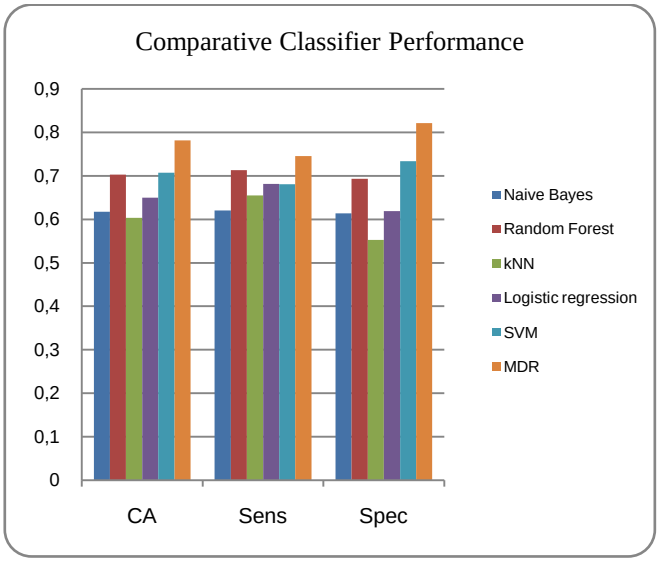

Figure (3): The Comparative performance of all classification methods with respect to Classification accuracy (CA), sensitivity (Sens), specificity (Spec)

By intersecting the significant SNP results from these tests, a total of 8,360 significant consensus SNPs were obtained. Using this set of significant consensus SNPs help reduce the false-positive association with disease. Fig. (2) shows the processing steps and summarizes their results of the two-stage SNP filtering and SNP-disease association ranking including the number of SNPs in each stage.

For biological feature selection, a total of 730,525 SNPs were annotated with gene and group information using Biofilter based on the relationships stored in the LOKI database. Then, the genes that do not contain at least one supported relation with AD are excluded. After mapping AD-associated genes back to SNPs, the new biological dataset of 3,131 SNPs was obtained.

The results of comparing classification accuracy, sensitivity, and specificity of the six classifiers are shown in Fig. (3). MDR achieved the highest classification accuracy, therefore it would be utilized in IHOEB proposed framework.

Three AD models were obtained from statistical analysis alone, biological analysis alone and the new IHOEB framework to illustrate its performance advantage.

In the statistical model, the significant consensus dataset is used to identify 2-way models that are associated with AD. The best ten models for 2-way interaction are presented in Table 2. The predictive performance of the models was estimated from a 10-fold cross-validation along with the training and testing datasets. The metric of model fit was BA averaged for all cross-validation experiments. Our results showed that the training and testing accuracies are close to each other, which means less over fitting and more generalizability [39]. However, such indicated statistical associations lack evidence for biological mechanism in the disease.

The biological model used Biofilter whereby all possible pairs of genes with prior knowledge of putative epistasis were identified. The generated gene-gene models were ranked by their implication score then mapped into SNP-SNP interaction models. The ranking by the implication score was done by sorting those models by the number of sources then number of groups in a descending manner. A total of 8,192 SNP-SNP interaction models were generated from the biological dataset of 3,131 AD-associated SNPs. The resultant top ten SNP-SNP and gene-gene models are listed in Table 3. As can be noticed, these models show less accuracy than statistical models due to the absence of significant effects between factors in the model.

The models derived from the new IHOEB framework start by integrating statistical and biological feature selection methods to form a combined dataset. The combined dataset helped identify more reliable models with underlying biological mechanism of AD in a statistically significant manner to reduce the likelihood of false-positive results. This was evident by observing many 2-way interaction models from IHOEB to be replicated in both the statistical and biological models, which serve as validation and confirmation of the potential of the new method. We used the combined dataset to identify 3-way models associated with AD using MDR as the classifier of choice based on its superior performance relative to other classifiers. The best ten models for 3-way interaction are shown in Table 4.

Given the size of the combined dataset of 11491 SNPs, performing a search within all 4-way models within this dataset is computationally prohibitive with estimated operations of more than 200 trillion comparisons. So, we used the strategy to limit our search to the interaction of the 1000 best 3-way models with other SNPs included in the combined set. The top ten results of the 4-way interaction models are presented in Table 5.

Five significant pairs of interacting models in the results of biological analysis were repeated in IHOEB 4-way models with BA greater than 0.74. The most repeated interaction is between CALM1 and CALM3 in models 2, 7 and 10, in addition to the interaction between FAS and FADD repeated in models 3 and 6 as observed from Table 5.

Table 3. The top 10 SNP-SNP and corresponding gene-gene models using biological analysis

	SNP1	SNP2	Gene1	Gene2	Implication Score	BA
1	rs2268433	rs710889	CALM1	CALM3	3-104	0.574
2	rs17036325	rs10113	CALM2	CALM3	3-104	0.572
3	rs2300496	rs17036325	CALM1	CALM2	3-104	0.563
4	rs10931934	rs1131715	CASP8	FADD	3-19	0.579
5	rs6948	rs4647698	CASP3	CASP9	3-6	0.557
6	rs11899004	rs17860418	CASP7	CASP8	3-4	0.561
7	rs1143634	rs3093662	IL1B	TNF	2-15	0.532
8	rs3218615	rs1131715	FAS	FADD	2-13	0.557
9	rs3093665	rs10793035	TNF	FADD	2-5	0.541
10	rs165932	rs6669689	PSEN1	NCSTN	2-4	0.557

Table 4. The top ten 3-way interaction models using IHOEB analysis (BA: balancing accuracy)

	Model	BA Training	BA Testing	p-Value
1	rs1542176, rs4955669, rs17021105	0.7219	0.7155	1.30E-4
2	rs1542176, rs7650925, rs3800908	0.7195	0.7066	2.78E-3
3	rs7650925, rs3765121, rs3800908	0.7116	0.6981	6.39E-3
4	rs1825503, rs556322, rs173644	0.7073	0.6947	3.00E-3
5	rs7591175, rs9878318, rs2309949	0.7054	0.7119	1.66E-3
6	rs2300496, rs10498633, rs710889	0.7063	0.7162	9.40E-3
7	rs2289319, rs347984, rs7786289	0.6988	0.6750	9.40E-3
8	rs759050, rs4547755, rs412657	0.7018	0.6809	7.15E-3
9	rs6882, rs10230371, rs235390	0.6898	0.6443	2.20E-3
10	rs7650925, rs3800908, rs2300496	0.6856	0.6767	5.10E-3

These 4-way models had a pair of SNPs with prior evidence of being related with AD and interacting with two other SNPs in a significant manner. So there is adequate evidence to suggest a possible biological relationship between these SNPs and AD.

New 4-way interaction models were also reported in Table 5 models 3, 5, 8 and 9. These 4-way models were significantly associated with AD with BA values ranging from 0.7646 to 0.7502. All the SNPs in these defined models were recognized in the genetic region. Models 1 and 4 are two additional new models with better accuracy of 0.7864 and 0.7622 respectively. However, the last SNP in each model of them is not recognized in the genetic region.

The obtained predictive values are higher than those reported in literature [13,15,16]. In [13], the study used Random Forest (RF) classifier to identify new genes associated with neurological disorders using a two-stage quality-based approach in RF for SNP selection within 188 neurological controls and 176 AD patients from NACC brain bank. However, the prediction performance of the RF and two stage RF models on AD data set did not exceed 0.6320 and 0.7100 respectively with 20 trees. In [15] the study predicted AD from SNP biomarkers and clinical data where they identified 958 biologically and statistically significant SNPs associated with late onset AD. These are later used to construct a decision tree model for differential diagnosis with an accuracy of classification of 0.5608. In [16], the study used Label Propagation (LP) method to rank SNPs in two different AD datasets with classification accuracies of 0.6039 ± 0.0300 and 0.7023 ± 0.0272 using the top two ranked SNPs and the top five ranked SNPs respectively. Therefore, the proposed technique offers significant improvement over previous methods.

The strategy employed in this study of building higher dimensional models starting from interactions involving one SNPs/genes previously identified to be related to AD and using that to identify other SNPs while making the search far more efficient is a key aspect of the proposed method. In order to illustrate the effectiveness of such strategy, we studied the correlation between pairs of SNPs with one of them having well-known association with AD from the literature whereas the other was identified from our study. The results are shown in Table 6 where the highest correlation had r^2 value of 0.183266 and was found between rs11771145 (AD gene EPHA1) and rs2300496 (CALM3), which is significantly better than a recent study [43] where the highest correlation had an r^2 value of 0.0730273 and was found between rs72508453 (AD gene HLA-DRB5) and SNP on chromosome 16 at position 6110138 (non-AD gene RP11-509E10.1).

As expected, IHOEB models performed significantly better than using either statistical or biological models alone in terms of classification accuracies. This confirms the advantage of combining both statistical and biological filters and also the potential of using such combination to reduce the search space for higher-order interactions without exhaustive evaluation of all possible higher order models. This helps to efficiently identify accurate and reliable higher-order AD predictive models.

IV. Conclusions

Large In this study, we presented a new IHOEP framework that integrates both the statistical and biological predictive modeling approaches to derive high-order interactions in a computationally efficient manner. The new approach combines the advantages of statistical significance and clarity of underlying biological mechanism. The developed method was implemented and applied to the ADNI database [21] and the results of best 2-, 3-, and 4-way interactions are presented.

Table 5. The top ten 4-way interaction models using IHOEB analysis (BA: balancing accuracy)

	Model (SNP / GENE)	BA Training	BA Testing	p-Value
1	rs7650925, rs3765121, rs3800908, rs1542176 SRPRB, KCNIP4, MAD1L1, NON	0.7671	0.7315	1.00E-3
2	rs1862710, rs11136000, rs2300496, rs10113 CASP9, CLU, CALM1, CALM3	0.7710	0.7340	4.90E-3
3	rs1571011, rs1131715, rs7650925, rs3800908 FAS, FADD, MAD1L1, SRPRB	0.7550	0.7270	4.80E-3
4	rs7650925, rs6546056, rs2300496, rs878698 SRPRB, LINC00309, CALM1, NON	0.7451	0.7240	5.28E-3
5	rs7650925, rs3765121, rs3800908, rs6546056 SRPRB, KCNIP4, MAD1L1, LINC00309	0.7362	0.7439	7.10E-3
6	rs7650925, rs3765121, rs1571011, rs1131715 SRPRB, KCNIP4, FAS, FADD	0.7271	0.7366	2.00E-3
7	rs3800908, rs1131715, rs2268433, rs10113 MAD1L1, FADD, CALM1, CALM3	0.7339	0.6995	2.34E-3
8	rs7650925, rs3800908, rs7799696, rs11136000 SRPRB, MAD1L1, DGKB, CLU	0.7257	0.7324	2.66E-3
9	rs3800908, rs11155266, rs6948, rs10113 MAD1L1, HIVEP2, CASP3, CALM3	0.7206	0.6914	1.50E-3
10	rs7650925, rs2268433, rs710889, rs1542176 SRPRB, CALM1, CALM3, NON	0.7288	0.7057	5.05E-3

Table 6. Correlation between pairs of SNPs with one of them has well-known association with AD from the literature whereas the other was identified from our study

	Well-known AD SNPs/ Genes				
	rs7412 APOE	**rs3865444 CD33**	**rs11771145 EPHA1**	**rs190982 MEF2C**	**rs8093731 DSG2**
rs10113	r^2 0.0020164 BA 0.0605735	r^2 0.0112447 BA 0.0532283	r^2 0.0201508 BA 0.0587378	r^2 0.00515073 BA 0.0279893	r^2 0.0724922 BA 0.591871
rs2300496	r^2 0.129783 BA 0.516324	r^2 0.0020073 BA 0.023835	r^2 0.183266 BA 0.167158	r^2 0.0172606 BA 0.055826	r^2 0.0161796 BA 0.25663
rs1131715	r^2 0.0147742 BA 0.142851	r^2 0.0375487 BA 0.0851675	r^2 0.00091567 BA 0.0112233	r^2 0.13238 BA 0.127189	r^2 0.00118773 BA 0.0845198
rs6948	r^2 0.095152 BA 0.429449	r^2 0.0074773 BA 0.0404584	r^2 0.072470 BA 0.112244	r^2 0.0724706 BA 0.112244	r^2 0.0122947 BA 0.252147
rs11899004	r^2 0.00027884 BA 0.0512519	r^2 0.0174752 BA 0.150235	r^2 0.0095853 BA 0.0535214	r^2 0.00650571 BA 0.0467828	r^2 0.133124 BA 0.344864
rs1862710	r^2 0.0089435 BA 0.137457	r^2 0.0310021 BA 0.0954368	r^2 0.0529091 BA 0.105407	r^2 0.0397716 BA 0.0722064	r^2 0.0351871 BA 0.444162

The IHOEB framework identified AD models that were replicated in both statistical and biological analyses with higher accuracy. The two gene pairs of CALM1/CALM3 and FAS/FADD were found to be the most replicated and validated models in our results. New 4-way interaction models that were significantly associated with AD were also identified with accuracies varying from 0.7464 to 0.7813. The ability to combine statistical analysis with support from prior biological knowledge in the new IHOEB framework makes it rather flexible, reliable and applicable to derive models for other complex diseases.

ACKNOWLEDGMENT

Data collection and sharing of this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 BG024904) and DOD ADNI (Department of Defense Bward no. W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: BbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Braclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org/). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuroimaging at the University of Southern California.

APPENDIX I: SUMMARY OF STATISTICAL CLASSIFIERS TESTED

In medical data analysis, data models differ significantly and it is usually not possible to determine which classifier would perform best without experimentation. It is therefore necessary to try and compare the performance of various classifiers to determine which one works well on the given dataset. The performance of a classifier is assessed using quantitative performance metrics such as accuracy, sensitivity and specificity.

Five popular classification techniques in addition to multifactor dimensionality reduction (MDR) [38,40-41] were applied to identify the model that best fit the significant consensus dataset and correctly predict the case-control status of test set. A summary of the classifier methods used is as follows:

Support vector machine (SVM) is a supervised learning algorithm designed to find the optimal separating hyper plane between the two groups of data. The corresponding hyperplane permits SVM to predict class label of unlabeled samples [34].

Naïve Bayes is a Bayesian Network that utilizes the Bayes theorem. NB classifier assume that the values of particular variables are conditionally independent of any other variable given the class (target). In NB structure all variables are children of the target variable. It means ignoring interactions between variables of the same class, which is violated in practice [19, 35]. However, NB is widely used due to its competitive classification accuracy.

K-nearest neighbor (KNN) method is a simple method for classifying objects based on closest training points in the feature space. KNN assumes that objects, which are close together, are probable to have the same classification. The chance that a point x belongs to a class can be estimated by the majority voting for the training data sets. in a specified neighborhood of x that belong to that class. The Euclidean distance that calculate the distances from x to all points in the training set is the most common distance metric used in K-nearest neighbor.

Random forest (RF) classifier is a classification method based on a collection of decision trees CART classifiers. RF uses bootstrap samples from the dataset to build a set of trees. To classify a new sample, a majority vote method is utilized to make a decision about class label. RF has better performance over the single (CART) [36].

Logistic regression (LR) is commonly used to analyze interactions between variables [37]. Logistic Regression is a nonlinear model that is particularly useful when the output variable is binary as in case and control studies.

Multifactor dimensionality reduction (MDR) [38] is a method that can increase the power of detecting interactions and be able to detect high-order interactions even in the absence of statistical main effects [9]. Using MDR for high-order interactions may provide an optimal approach to solve this problem. In principle, MDR is a machine learning approach designed for detecting and evaluating SNP-SNP interactions associated with disease. An advantage of MDR over conventional statistical methods like logistic regression lies in that MDR is a nonparametric method and model-free method that does not need a genetic model [39]. Several studies applied MDR to assess gene–gene interactions in different human diseases such as bladder cancer, multiple sclerosis, and AD [38,40-41]. MDR pools the genotypes from two or more SNPs into one attribute that has high risk or low risk groups. The binary attribute is considered high risk if the ratio of cases to controls in that group is higher than the original ratio of cases to controls in the dataset [40]. Otherwise, it is considered low risk. This change in the space representation leads to dimensionality reduction that aids in discovering higher interactions among the SNPs. MDR uses cross-validation to evaluate the predictive accuracy of all exhaustive 2-, 3-, 4-, up to n-SNP combination models. The best model is the model with the highest classification accuracy that is subsequently evaluated by the test set to assess its prediction error. The genetic interactions have been successfully carried out using the open-source MDR [39] software package.

REFERENCES

[1] P. Bailey, "Biological markers in Alzheimer's disease," *Can J Neurol Sci*, vol. 34, Suppl. 1, pp. S72-6, 2007.

[2] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, and L. Shen, "From phenotype to genotype: an association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs," *Bioinformatics*, vol. 28, pp. i619-i625, 2012.

[3] C. M. Karch and A. M. Goate, "Alzheimer's Disease Risk Genes and Mechanisms of Disease Pathogenesis," *Biological Psychiatry*, 2014/10/16 2014.

[4] Alzheimer's Association, "2015 Alzheimer's disease facts and figures," *Alzheimers Dement*, vol. 11, no. 3, pp. 332-84, 2015.

[5] S. L. Rosenthal and M. I. Kamboh, "Late-Onset Alzheimer's Disease Genes and the Potentially Implicated Pathways," *Curr Genet Med Rep*, vol. 2, pp. 85-101, 2014.

[6] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature genetics*, vol. 37, pp. 413 - 417, 2005.

[7] C. Floudas, N. Um, M. Kamboh, M. Barmada, and S. Visweswaran, "Identifying genetic interactions associated with late-onset Alzheimer's disease," *BioData Mining*, vol. 7, no. 35, pp. 1-19, 2014.

[8] X. Guo, N. Yu, F. Gu, X. Ding, J. Wang, and Y. Pan, "Genome-wide interaction-based association of human diseases-a survey," *Tsinghua Science and Technology*, vol. 19, pp. 596-616, 2014.

[9] M. T. W. Ebbert, P. G. Ridge, and J. S. K. Kauwe, "Bridging the Gap between Statistical and Biological Epistasis in Alzheimer's Disease," *BioMed Research International*, vol. 2015, Article ID 870123, pp. 1-7, 2015.

[10] X. Sun, Q. Lu, S. Mukherjee, P. K. Crane, R. Elston, and M. D. Ritchie, "Analysis pipeline for the epistasis search - statistical versus biological filtering," *Front Genet*, vol. 5, no. 106, pp. 1-7, 2014.

[11] T. Honkela, W. O. A. Duch, M. Girolami, S. Kaski, I. Braenne, J. Erdmann, and A. Mamlouk, "SNPboost: Interaction Analysis and Risk Prediction on GWA Data," in *Artificial Neural Networks and Machine Learning ICANN 2011*, vol. 6792, Springer, pp. 111-118, 2011.

[12] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, pp. 781 - 791, 2006.

[13] T.-T. Nguyen, J. Huang, Q. Wu, T. Nguyen, and M. Li, "Genome-wide association data classification and SNPs selection using two-stage

quality-based Random Forests," *BMC Genomics*, vol. 16 (Suppl. 2), no. S5, pp.1-11, 2015.
[14] J. O. Setubal, N. Almeida, G. Araujo, M. B. Souza, J. R. Oliveira, and I. Costa, "Random forest and gene networks for association of SNPs to Alzheimer's disease," in *Advances in Bioinformatics and Computational Biology*, vol. 8213 of the series Lecture Notes in Computer Science, Springer, pp. 104-115, 2013.
[15] O. Erdogan and Y. Aydin Son, "Predicting the disease of Alzheimer with SNP biomarkers and clinical data using data mining classification approach: decision tree," *Stud Health Technol Inform*, vol. 205, pp. 511-5, 2014.
[16] M. E. Stokes, M. M. Barmada, M. I. Kamboh, and S. Visweswaran, "The application of network label propagation to rank biomarkers in genome-wide Alzheimer's data," *BMC Genomics*, vol. 15, no. 282, pp. 1-13, 2014.
[17] E. S. Gusareva, M. M. Carrasquillo, J. Williams, P. Amouyel, K. Sleegers, N. Ertekin-Taner, J. C. Lambert, and K. Van Steen, "Genome-wide association interaction analysis for Alzheimer's disease," *Neurobiol Aging*, vol. 35, pp. 2436-43, 2014.
[18] J. S. K. Kauwe, S. Bertelsen, K. Mayo, C. Cruchaga, R. Abraham, P. Hollingworth, D. Harold, M. J. Owen, J. Williams, S. Lovestone, J. C. Morris, A. M. Goate, Alzheimer's Disease Neuroimaging Initiative, "Suggestive synergy between genetic variants in TF and HFE as risk factors for Alzheimer's disease," *Am J Med Genet Part B*, vol. 153B, pp. 955-959, 2010.
[19] F. F. Sherif, N. Zayed, and M. Fakhr, "Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks," *Advances in Bioinformatics*, vol. 2015, p. 8, 2015.
[20] L. Zou, Q. Huang, A. Li, and M. Wang, "A genome-wide association study of Alzheimer's disease using random forests and enrichment analysis," *Science China Life Sciences*, vol. 55, pp. 618-625, 2012.
[21] M. C. Carrillo, L. J. Bain, G. B. Frisoni, and M. W. Weiner, "Worldwide Alzheimer's disease neuroimaging initiative," *Alzheimers Dement*, vol. 8, pp. 337-42, 2012.
[22] E. M. Smigielski, K. Sirotkin, M. Ward, and S. T. Sherry, "dbSNP: a database of single nucleotide polymorphisms," *Nucleic Acids Res*, vol. 28, pp. 352-5, 2000.
[23] S. Purcell, "PLINK-1.07," in *http://pngu.mgh.harvard.edu/purcell/plink/*.
[24] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, pp. 559-75, 2007.
[25] C. Gondro, J. Van der Werf, and B. Hayes, *Genome-wide association studies and genomic prediction*: Humana Press, 2013.
[26] M. D. Ritchie, "Using Biological Knowledge to Uncover the Mystery in the Search for Epistasis in Genome Wide Association Studies," *Annals of human genetics*, vol. 75, pp. 172-182, 2011.
[27] S. A. Pendergrass, A. T. Frase, J. R. Wallace, D. Wolfe, N. Katiyar, C. Moore, and M. D. Ritchie, "Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development," *BioData Mining*, vol. 6, 2013.
[28] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res*, vol. 29, pp. 308-11, 2001.
[29] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res*, vol. 35, pp. D26-31, 2007.
[30] G. R. Brown, V. Hem, K. S. Katz, M. Ovetsky, C. Wallin, O. Ermolaeva, I. Tolstoy, T. Tatusova, K. D. Pruitt, D. R. Maglott, and T. D. Murphy, "Gene: a gene-centered information resource at NCBI," *Nucleic Acids Res*, vol. 43, pp. D36-42, 2014.
[31] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Res*, vol. 36, pp. D281-8, Jan 2008.
[32] E. W. Sayers, T. Barrett, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res*, vol. 37, pp. D5-15, 2009.
[33] T. Hu, C. Darabos, M. E. Cricco, E. Kong, and J. H. Moore, "Genome-wide genetic interaction analysis of glaucoma using expert knowledge derived from human phenotype networks," *Pac Symp Biocomput*, pp. 207-18, 2015.
[34] S. C. Yücebaş, Y. A. Son, "A Prostate Cancer Model Build by a Novel SVM-ID3 Hybrid Feature Selection Method Using Both Genotyping and Phenotype Data from dbGaP," *PLoS ONE*, vol. 9, p. e91404, 2014.
[35] E. N. Richard, *Learning Bayesian Networks*: Prentice-Hall, Inc., 2003.
[36] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5 - 32, 2001.
[37] W.-H. Wei, G. Hemani, and C. S. Haley, "Detecting epistasis in human complex traits," *Nature Reviews Genetics*, 2014.
[38] Y. Wu, L. Zhang, L. Liu, Y. Zhang, Z. Zhao, X. Liu, and D. Yi, "A multifactor dimensionality reduction-logistic regression model of gene polymorphisms and an environmental interaction analysis in cancer research," *Asian Pac J Cancer Prev*, vol. 12, pp. 2887-2892, 2011.
[39] J. H. Moore and P. C. Andrews, "Epistasis analysis using multifactor dimensionality reduction," in *Epistasis*: Springer, 2015.
[40] J. Hoh, A. iWlle, R. Zee, S. Cheng, R. Reynolds, K. Lindpaintner, and J. Ott, "Selecting SNPs in two-stage analysis of disease association data: a model-free approach," *Annals of Human Genetics*, vol. 64, pp. 413 - 417, 2000.
[41] D. Brassat, A. A. Motsinger, S. J. Caillier, H. A. Erlich, K. Walker, L. L. Steiner, B. A. C. Cree, L. F. Barcellos, M. A. Pericak-Vance, S. Schmidt, S. Gregory, S. L. Hauser, J. L. Haines, J. R. Oksenberg, and M. D. Ritchie, "Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans," *Genes and Immunity*, vol. 7, pp. 310-315, 2006.
[42] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, J. H. Moore, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genet Epidemiol*, vol. 31, no. 4, pp. 306-315, 2007.