Ain Shams Engineering Journal 14 (2023) 101986

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Discovering epistasis interactions in Alzheimer's disease using integrated framework of ensemble learning and multifactor dimensionality reduction (MDR)

Marwa M. Abd El Hamid [a,b,*], Mohamed Shaheen [b], Yasser M.K. Omar [b], Mai S. Mabrouk [c]

[a] Computer Science Department, The Higher Institute of Computers and Information Technology, El Shorouk Academy, El Shorouk City, Cairo, Egypt
[b] College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport, Egypt
[c] Biomedical Engineering Department, Misr University for Science and Technology, 6th of October City, Egypt

ARTICLE INFO

Article history:
Received 6 March 2022
Revised 29 June 2022
Accepted 17 September 2022
Available online 30 September 2022

Keywords:
Epistasis Interactions
Alzheimer's disease
Personalized Medicine
Ensemble learning techniques

ABSTRACT

Alzheimer's disease (AD) is a complex disorder with strong genetic factors. The proposed framework is applied to Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. We present a novel framework integrating ensemble learning and MDR constructive induction algorithm to discover epistasis interactions associated with AD in a computationally efficient method. Discovering epistasis interactions is a big challenge and significantly impacts personalized medicine (PM). The applied ensemble learning algorithms are random forests (RF) with Gini index and permutation importance, Extreme Gradient Boosting (XGBoost), and classification and regression trees (CART). The classification accuracy of 5-way models varied between (0.8674–0.8758), whereas the accuracy of 2-way, 3-way, and 4-way models varied between (0.6515–0.6649), (0.7071–0.7170), and (0.7811–0.7878) respectively. The promising results of this proposed framework show high-ranked risk genes and up to 5-way epistasis models that contribute to the disease risk efficiently and at higher accuracy.

1. Introduction

Personalized medicine, also named precision medicine, is a significant model that tailors the patient's treatment to his characteristics. Investigating epistasis interactions holds a substantial key to PM. Networks of gene interactions have clues to the treatment response and disease susceptibility. Precision medicine needs an enhanced understanding of the relationship between genetic data and complex diseases [1]. The main target of genome-wide association studies (GWAS) is to investigate the genetic variants across the human genome for detecting the SNPs most associated with complex diseases such as cancer, autoimmune, and AD [2]. Different computational and statistical approaches can identify genetic variants related to complex diseases. These methods can be categorized into single-locus and multi-locus analyses [3].

The single-locus approach investigates each SNP independently and its relation with the phenotype. Unfortunately, the single-locus analysis failed to demonstrate disease heritability and discover genetic risk factors. The multi-locus approach investigates the SNPs combinations and captures the interactions between several SNPs in the genetic data [4]. Most risky diseases are caused due to complex interactions among multiple genes. However, some computational and statistical challenges exist for modeling nonlinear SNP interactions, discovering the most significant genetic variables, and investigating the explored epistasis interaction models. Hence, the epistasis interactions better explain the susceptibility of risky diseases than individual genetic variants [5].

Feature selection techniques are considered the core concept in machine learning, which improves the model's performance. The data features that are used for training the machine learning algorithms have a significant influence on the achieved performance.

* Corresponding author at: Computer Science Department, The Higher Institute of Computers and Information Technology, El Shorouk Academy, El Shorouk City, Cairo, Egypt; College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport, Egypt.
E-mail addresses: marwa.mustafa@sha.edu.eg, marwa.ramadan5@student.aast.edu (M.M. Abd El Hamid), mohamed.shaheen@aast.edu (M. Shaheen), yasser.omar@aast.edu (Y.M.K. Omar), msm_eng@yahoo.com (M.S. Mabrouk).

Peer review under responsibility of Ain Shams University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.asej.2022.101986
2090-4479/

Furthermore, the relevant features can positively impact model performance. Therefore, feature selection is essential in designing the model to eliminate irrelevant or partially relevant features [6]. Common feature selection techniques are odds ratio, chi-square, logistic regression, and Relief-based algorithms. It was shown that Relief-based methods, specifically Tuned ReliefF (TuRF) method achieved remarkable results in filtering data features and ranking the most significant data attributes [7]. In this proposed work, TuRF method is applied for filtering the most significant SNPs associated with the disease that contribute to the phenotypic result through epistasis interactions.

The ensemble learning method is a machine learning model where many algorithms are trained to solve the same problem. In contrast, traditional machine learning techniques try to learn one hypothesis from training data. The main advantage of ensemble learning methods is trying to contrast a group of hypotheses and integrate them. Ensemble learning aims to incorporate the decisions from several models for enhancing predictive or classification performance [8]. The importance weight of the features is obtained using ensemble methods. These methods are still an important research area that may contribute to the success of PM. Hence, TuRF feature selection technique was applied in this study to reduce the ADNI data features to be manageable.

Feature importance techniques assign a score to input attributes depending on their significance in predicting the goal variable. They play a significant role in providing insight into the data and the model for enhancing the performance and the efficiency of the predictive model [9]. The scores are important in the predictive model on the problem for understanding the data and the model and input attribute reduction. Feature importance scores can be used to select the features with the highest scores and delete the features with the lowest scores to improve the model performance.

Although the notable success of these methods lately, there is a lack of studies that provide insights about how these techniques shall be utilized in discovering the risk susceptibility SNPs associated with AD through epistasis interactions. The main target of this paper is to fill the notable gap in the literature and comprehensively apply ensemble learning methods to explore the most significant SNPs through gene-gene interactions. In this proposed work, various methods in ensemble learning were used, such as RF with Gini index and permutation importance [10], XGBoost [11], and CART [12].

The main contributions of this paper are as the following:

- Integrating constructive induction algorithms like MDR with ensemble learning algorithms in the proposed framework makes it robust, applicable, and reliable to create models for other risky diseases.
- The proposed framework suggests novel genes associated with AD
- Our results show high-ranked risk genes and up to 5-way epistasis models that may help better understand the disease etiology to diagnose and improve personalized therapeutic strategies.
- The results revealed that the reported accuracy scores of the proposed framework outperform the referenced literature work.

The following sections of the paper are organized as follows; Section 2 shows the background. Section 3 introduces the literature review. Section 4 describes the materials and methods. Section 5 presents the results. Section 6 presents the discussion. Finally, Section 7 explains the conclusions.

2. Background

Alzheimer's disease (AD) is a complex neurodegenerative disease that leads to memory problems, confusion, and impairments in semantic memory [14]. This disease is considered the cause of 60:70 % of dementia cases. Difficulty in remembering recent events is regarded as one of the AD symptoms. The advancement of AD leads to more problems like motivation loss, planning, behavioral issues, and self-care management problems. AD is a complex, irreversible, progressive brain disease that causes memory destruction and difficulty carrying out the simplest tasks. AD disease is one of the currently leading causes of death all over the world. This disease is the most common cause of dementia among older people [15].

Dementia is the loss of thinking, behavioral abilities, reasoning, and remembering. Scientists try to explore the brain changes involved in the onset and advancement of the disease. The complex brain changes lead to memory and cognitive problems. These changes are considered toxic changes in the brain [15]. Apolipoprotein E4 (apoE4) is one of the most widespread genetic risk factors for this complex disease. It appeared in several AD cases at loci rs429358 and.

rs7412 [16]. Hence, this risky gene is a vital possible therapeutic target for this disease. Detecting the most effective and important biomarkers in the early stages of the disease can improve searching for treatment and slow its progression.

Epistasis interactions are considered a genetic phenomenon in which the function of a given gene depends on the activation or inactivation of one modifier gene or more. This genetic phenomenon is a deviation from the expected phenotype of combining two alleles. It has been shown that epistasis interactions effectively predict phenotype from genotype for a subject. They are examined in population studies to identify genetic risk factors in risky diseases. Identifying the most significant epistasis interactions has become an essential topic in complicated trait genetics [17].

It was shown that genes are interacted with each other and do not function alone. These interactions are essential for gene regulation and different developmental pathways. Some genes can alter other genes' functions, resulting in risky diseases. The researchers concentrating on candidate genes cannot thoroughly investigate complex diseases [18]. Therefore, many genes can interact with each other to increase or decrease the susceptibility of a disease. If the effect of the disease-bearing gene is altered by the effects of another gene, then identifying the first gene can be difficult. Identifying the most significant genes will be more difficult if there are several epistasis interactions related to a disease. Ensemble learning algorithms are machine learning techniques that hybrid many base models to generate one powerful predictive model. It was shown that they could be utilized in training highly accurate classifiers and discovering new genetic biomarkers by ranking the features depending on their significance in classification [8].

3. Literature review

Several studies applied many techniques to analyze the individual effect of each SNP and identify the crucial SNPs associated with common diseases. There are different related AD diagnoses and multi-modal methods, such as 1. fusion of multimodality data using a stage-wise deep neural network to use the complementary information from both the neuroimaging and genetic data for diagnosing AD and its related early statuses [19]. 2. hybrid-fusion network for multi-modal MR image synthesis to explore the modality-specific properties within each modality and simultaneously exploits the correlations across multiple modalities [20]. 3. Latent representation learning method for multi-modality-based AD diag-

nosis to learn an accurate AD prediction model from the incomplete multimodality dataset [21].

Exploring the SNPs of an affected person to predict the disease risks is a crucial stage in contributing to PM [22]. In addition, exploring gene-gene interactions is significant for investigating the mechanism of the disease and developing PM. However, discovering them is still under research [23]. Two different analysis methods can be used to detect genetic variants associated with diseases. The first approach is a uni-variable analysis that examines the association of each SNP independently with the phenotype. The second approach is a multi-variable analysis that can capture the interactions between multiple SNPs and better explain some diseases' susceptibility [24]. In [25], the authors applied machine learning techniques for predicting AD-affected individuals from genetic variation. They downloaded the used dataset from ADNI. The dataset consists of 230 patients and 241 normal people. In that paper, Least Absolute Shrinkage and Selection Operator (LASSO) KNN, naive Bayes, RF, and SVM were applied to identify new genetic markers associated with AD disease. The results revealed that LASSO scored the best result (0.719). Even so, that research gave no attention to detecting genetic interactions.

In [26], the authors applied sequential minimal optimization algorithms with different kernels, Naïve Bayes (NB), Tree Augmented Naive Bayes and K2 learning algorithms. The authors used a Whole-genome sequencing dataset including 2,379,855 SNPs for 282 controls, 442 mild cognitive impairment, and 48 AD. That research focused on detecting the most significant SNPs associated with AD with high classification accuracy. The results revealed that NB and K2 learning algorithms achieved 98 % and 98.40 % accuracy, respectively. However, that paper ignored interaction effects between SNPs. The authors [27] implemented a framework for detecting epistasis interactions and enhancing early disease diagnosis. The authors obtained the dataset from ADNI database. This dataset consists of 730,525 SNPs for 125 unaffected individuals and 306 affected individuals. Some algorithms were applied like naïve Bayes, Support vector machine, k-nearest neighbor, logistic regression, RF, and MDR classifiers. The best results were achieved by MDR. The achieved classification accuracy of their proposed framework varied between 0.7410 and 0.7860. However, the achieved accuracy needed to be enhanced to investigate this complex disease better.

In [28], the authors conducted a differential network analysis to reveal potential networks involved in the neuropathogenesis of AD and detect genes to predict AD. The dataset was obtained from the Religious Orders Study and the Rush Memory and Aging Project, including 193 CE patients and 172 controls. That research developed machine learning techniques like joint density-based non-parametric differential interaction network analysis (JDINAC), RF, and Logistic Regression. The JDINAC technique achieved the best results. It achieved accuracy: 0.791, sensitivity: 0.776, and specificity: 0.808. However, that study did not identify higher-order interactions of genes in their analysis.

In [29], the authors present GenEpi, a computational package to uncover epistasis interactions associated with phenotypes. That study aims to discover SNP interactions by building GenEpi package to explore epistasis interactions related to the phenotype using machine learning. The authors applied their work to an AD dataset for Alzheimer's disease Dream Challenge. The used dataset includes 767 individuals of cases and controls from ADNI database. They defined the interactions between two SNPs only. Unfortunately, that paper did not detect higher-order interactions of genes in their analysis. In [30], the authors applied RF method to detect and model epistasis interactions. They performed a comparative analysis of feature importance metrics for improving the interpretability of RF with complex interactions. The authors obtained the dataset from ADNI database. They established that the permutation feature importance metric provides a more accurate feature importance rank estimation in the presence of epistasis interactions. The applied model was tuned with grid search and had a classification accuracy of 62.6 % for the AD dataset. The achieved accuracy needed to be increased for better investigation of AD disease.

In [31], the authors applied elastic net machine learning methodology to identify the strongest predictors for the risk of AD, combining all genotyping data (direct effects and epistatic interactions). The used dataset consists of 1078 individuals (602 controls and 476 cases). After applying the EN method to the data, the model achieved an accuracy of 72 %, including epistatic interactions between the assessed variants as predictors of AD risk. Unfortunately, there is a lack of identifying higher-order interactions of genes in their analysis for more understanding of the biological mechanism of the disease. Many studies focused on univariable analysis and investigated the effect of independent SNP loci to detect genetic variants associated with the disease [25,26]. However, the studies that focus on discovering multi-locus interactions are still limited and may have more robust associations. This paper aims to fill the notable gap in the literature for finding the most significant epistasis interactions up to fifth-order interactions associated with AD.

4. Materials and methods

The detected genetic causes of the disease mainly concentrated on individual genes, but it was shown that risky diseases might be affected by the interaction of genes. Though deep-learning-based methods can solve many problems, such as feature selection, model interpretability is one of the biggest challenges with deep learning. Ensemble learning is a machine learning model where many algorithms are trained to solve the same problem. In contrast, traditional machine learning techniques try to learn one hypothesis from training data.

The reasons for applying ensemble learning techniques over a single model are as the following:

- Ensemble learning techniques can make better predictions and perform better than any traditional machine learning model.
- They are used to improve robustness or reliability in the average performance of a model by reducing the spread or dispersion of the predictions and model performance

The proposed system workflow is described in Fig. 1 to show the most critical steps. First, the ADNI dataset was loaded and went through preprocessing steps. The preprocessing steps aim to add affectional information of individuals identified as normal or case. Subsequently, APOE genotyping was added by estimating alleles of the SNPs: rs7412 and rs429358. Then QC was applied to exclude SNPs with insufficient genotyping quality. After that, highly correlated SNPs were excluded using linkage disequilibrium. Then SNP-disease association tests were applied to assess the statistical association of each SNP with the disease using a p-value < 0.01. Finally, their results were intersected to extract the significant SNPs.

After the preprocessing phase, dimensionality reduction was applied using TuRF algorithm to remove the irrelevant markers. Then, different ensemble learning algorithms were applied to detect SNPs that contribute to AD risk through epistasis interactions. The top twenty rankings provided by these algorithms were integrated to detect 76 possible susceptibility SNPs. Finally, Multi-locus interaction analysis was performed on these identified SNPs using MDR to discover and validate the most significant gene-gene interactions. The critical aspect of this proposed work is to develop

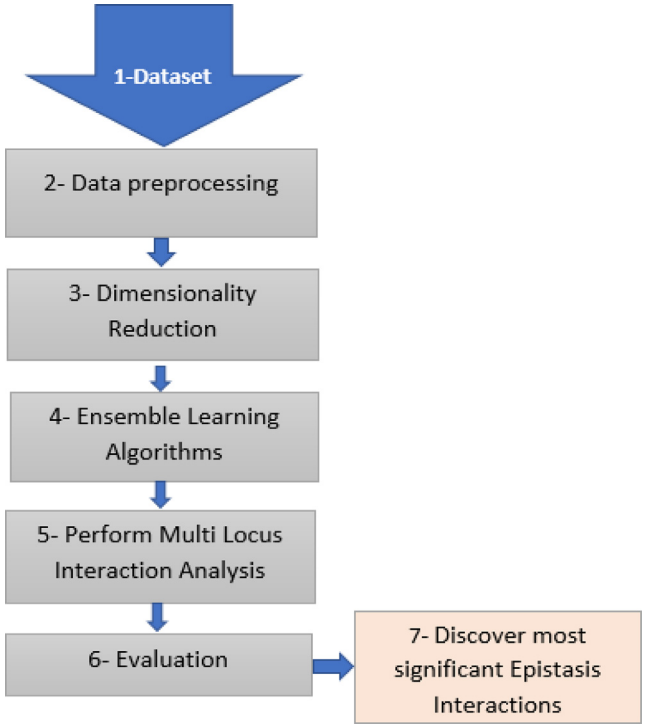


Fig. 1. Proposed framework.

a high-dimensional model for discovering epistasis interactions among genetic variants.

4.1. Dataset

The used dataset was downloaded from ADNI database [15]. ADNI GWAS data contained total genotypes for 431 subjects (127 controls and 304 cases). The total number of SNPs is 730,524 for both the unaffected individuals and the affected AD patients.

4.2. Data preprocessing

In this proposed work, the dataset was downloaded from ADNI database, and preprocessing steps were applied. Data preprocessing is a vital stage for achieving significant results. Several phases were used as follows:

(A) The phenotype that described the affection status column was added as a case or normal person. The number of normal subjects is 127, while the total number of cases is 304

(B) The SNPs (rs429358 and rs7412) of APOE were not presented in the ADNI dataset. However, the genotyping of APOE was performed at the time of participant enrollment and inserted into the ADNI website. Hence, APOE genotyping was added to the dataset. Therefore, the total number of SNPs became 730,526 after merging the APOE genotyping.

(C) Apply Quality Control (QC) using PLINK [32] to filter SNPs and minimize potential false findings. QC procedures were applied to the dataset as the following:

- Exclude the persons with a lot of missing genotyping data. The missing genotyping, which is more than 10 %, was removed.
- Exclude the SNPs with missing genotype rates. The only SNPs with a 90 % genotyping rate were included.
- Subsequent analyses can be set to include only SNPs with MAF >=0.1.

After applying these steps, this yielded 431 subjects, including 304 patients and 127 controls. The total number of SNPs after QC became 530,750.

(D) Applying the LD pruning step is vital for improving the power of complex disease genetic association studies. The SNPs of high correlation were removed from the ADNI dataset, leaving 447,538 markers.

(E) Apply SNP–disease association tests [33] to reduce the enormous computational requirements. Three SNP-disease association tests were applied using PLINK. Every 447,538 SNPs were independently tested for association with AD disease in the basic association test, logistic model, and Fisher's exact (allelic association) test. The p-value threshold was used as a significance level for detecting the SNP associations. The irrelevant SNPs with a p-value of more than 0.01 were excluded. The significant SNPs with a p-value less than the threshold of 0.01 can lead to a higher power than using the threshold of 0.05 [34]. Hence, the total number of SNPs decreased to 4383, 3863, and 3861 using the Basic association test, logistic model, and Fisher's Exact test, respectively. A total of 3,502 significant SNPs were gained using R [35] by applying the intersection of the SNP results from the three tests.

(F) The obtained significant SNP result can reduce the false-positive association with AD. However, the remaining total number of SNP is still large. Hence, it is an important step to apply feature selection techniques.

4.3. Dimensionality reduction

Feature reduction techniques were used to reduce the high dimensionality and focus on discarding redundant and non-significant features from the used dataset. TuRF feature selection method usually works best for big volume data problems and always fits the complex nature of biology [7]. Furthermore, TuRF feature selection algorithm was used to improve the performance of a well-known algorithm called ReliefF algorithm.

This algorithm is a typical representative of the filter method. Filter algorithm is often used to rank or order features in the data set. TuRF algorithm is vital in filtering SNPs by adding an iterative component. This technique can recursively remove the law-ranked (irrelevant) SNPs in each iteration. For example, if the number of iterations of this technique is R and the total number of SNPs is N. TuRF algorithm will discard the N/R least discriminative SNPs in each iteration. TuRF algorithm was applied using a percentage of significant SNPs at 30 %. After applying TuRF algorithm, the total number of SNPs was reduced from 3,502 to 1,050.

4.4. Ensemble learning algorithms

Imbalanced data occurs in real life, like in medical diagnoses in which patients' records outnumber normal individuals. Ensemble learning techniques are considered one of the best solutions for imbalanced data classification problems. Hence, this paper demonstrates a novel framework using ensemble learning methods to solve the imbalanced data problem by enhancing the performance indicators for the classes that have been the majority of the samples.

Ensemble learning algorithms are powerful in decreasing variance and overfitting by using a group of diverse base learners [36]. Cross-validation is an effective preventative measure against overfitting. This proposed work used 10-fold cross-validation to get a better insight into the models, which eventually helped avoid overfitting.

We applied ensemble learning methods, including RF with Gini index and permutation importance, XGBoost, and CART, to detect each method's top 20 ranking SNPs. Then these identified significant SNPs were integrated to discover their power in exploring non-linear epistasis interactions for GWAS.

4.4.1. Random forests (RF) feature importance

RF classifier is a powerful tool to classify a new sample based on a collection of decision trees. A majority of voting was used for deciding on the class label. This algorithm is an extension of the bagging of several decision trees [10]. RF is typically treated as a black box. Hence, we computed feature importance by getting an insight into the RF model. RF feature importance is computed using different methods, including Gini importance and permutation methods. We used Gini importance and permutation methods to explain which SNPs are relevant.

In this proposed work, RF algorithm was used to apply feature importance using RandomForestClassifier class implemented in scikit-learn. However, applying this algorithm with default settings of hyper-parameters would not yield the best results for large GWAS datasets. There is no specific method to calculate how the change in the hyperparameter values will optimize the model architecture. Hence, the models' hyperparameter values were set by experimentation and specifying a range of possible values for all the hyperparameters. Thus, we applied parameter tuning for RF using RandomizedSearchCV method to find optimal parameters.

The tuned hyperparameters are: criterion= 'gini', n_estimators = 1800, bootstrap: false, min_samples_leaf = 1, max_features='auto', min_samples_split = 5, max_depth = 90. Gini index was used as a hyper-parameter to choose which feature would be used for splitting the data. We used Gini Index [10] to quantify SNPs' importance to determine the top-ranking SNPs.

The other method of interpreting RF is to compute feature importance using permutation feature importance. Permutation feature importance method randomly shuffles each attribute and evaluates the change in the model's performance [13]. The attributes which impact the performance the most are the most significant ones. In this paper, the permutation_importance function was used to calculate the feature importance of estimators for the ADNI dataset. We measured the importance of a SNP by measuring the increase in the model's prediction error after permuting the SNPs. In this work, python's ELI5 library was used to explain the black-box of RF model by measuring how the score decreases when a SNP is unavailable. This library provides a convenient way to calculate permutation importance. We used Gini importance and permutation methods to determine the top-ranking SNPs by each and discover their power in finding non-linear epistasis interactions.

4.4.2. XGBoost

Extreme Gradient Boosting (XGBoost) is a scalable end-to-end tree boosting system used by data scientists to achieve promising results on several machine learning algorithms challenges. It is a widely used algorithm for supervised learning in machine learning [11].

In this paper, the model was built using trees as base learners depending on XGBoost's scikit-learn compatible API. The library XGBoost and other libraries were imported using python. Finally, we applied parameter tuning for XGBoost using RandomizedSearchCV method to find optimal parameters.

The tuned hyperparameters are: base_score = 0.5, booster='gbtree', n_estimators = 100, colsample_bylevel = 1, n_jobs = 1, nthread = None, objective='binary:logistic', and so on for other parameters. In this work, XGBoost was used as it provides estimates of feature importance from a trained predictive model to determine the top-ranking SNPs. The top rankings provided by this algorithm are essential to show high-ranked risk genes and epistasis interactions that may explain the risk of the disease.

4.4.3. Classification and regression trees (CART)

A Decision Tree (DT) is considered an effective algorithm for predictive modeling machine learning. The classical DT techniques have been around for decades. However, modern algorithms like CART are one of the most powerful techniques available. CART is used for classification/regression predictive modeling problems. CART is a non-parametric DT learning algorithm that generates either regression or classification trees depending on whether the dependent variable is numerical or categorical. The implemented algorithm in sklearn is called CART and works with categorical and numerical targets [12].

In this work, we applied parameter tuning for CART using RandomizedSearchCV method to find the best parameters. The tuned hyperparameters are: criterion: 'entropy', max_depth: 10, max_features: 18, min_samples_leaf: 7. CART algorithm offers importance scores based on the reduction in the criterion parameter used to select split points, like Gini or entropy. The optimum split was chosen by the SNP with less entropy. We used the feature_importances property to retrieve the relative importance scores for each input SNP. In this paper, the top-ranking SNPs by CART algorithm were detected to discover their power in finding high-ranked epistasis interactions.

4.5. Multi-locus interaction analysis

MDR is a powerful machine learning method that detects interacting combinations of genetic variations related to complex diseases like AD [37]. The top 20 ranking SNPs provided by the applied ensemble learning techniques were integrated to detect 76 possible susceptibility SNPs. In this proposed work, multi-locus interaction analysis was performed on the identified SNPs described in the previous section using MDR to discover significant epistasis interactions. In this work, MDR was used to collect the genotypes from 2 SNPs or more into an attribute with high/low-risk groups. Hence, MDR reduced the high dimensionality of the features from N dimensions to only one dimension. This process is called constructive induction, wherein the new feature is defined as a function of 2 or more other features. The newly generated features were assessed for their ability to predict and classify AD status. The binary feature is defined as a high-risk attribute if the ratio of patients to normal in that group is greater than the original ratio of affected patients to normal people in ADNI dataset. Other than that, the binary attribute is low risk.

In a GWAS, using an exhaustive search to explore epistasis interactions is computationally expensive. Therefore, this task requires a computational load for larger order interactions and markers. Moreover, when the number of markers is massive, the number of multi-locus interactions increases. Hence, we present a novel approach combining ensemble learning and MDR methods to decrease some shortcomings of the MDR method by determining the top-ranking SNPs.

In this paper, MDR was used to evaluate pairwise, 3-way, 4-way, and 5-way interaction predictive accuracy. Since searching these SNP interaction models within the dataset is computationally complex. Therefore, searching for the significant interactions was limited to the interactions of the top 20 rankings provided by the applied ensemble algorithms. These ensemble methods generated rankings on the significance of SNP contribution to AD classification. By integrating the top 20 rankings provided by the ensemble methods, the total number of features became 76 SNPs.

Then the statistical interaction analysis of the 76 identified SNPs was performed, and they could discover critical gene-gene interactions. As a result, the coding SNPs mapped to 38 genes, con-

sisting of known AD-related genes (27 identified genes) and genes that have not been explored previously related to AD (11 discovered genes), as shown in Table 1.

The newly discovered genes can be possible risk association genes to AD.

A novel MDR method was proposed using ensemble learning methodologies to discover new epistasis interactions associated with AD disease. The results demonstrate that this proposed framework can discover new risk genes and epistasis interactions that may help better understand the disease etiology.

4.6. Evaluation criteria

In this proposed work, the predictive performance of the applied ensemble learning algorithms was evaluated using classification accuracy, precision, recall, and f1-score for exploring the best SNPs. These significant SNPs were utilized in the proposed framework to contribute to the risk of AD through epistasis interactions. The performance of the models was estimated from 10-fold cross-validation along with the training and testing datasets. The chosen metric of model fit was balanced accuracy (BA) averaged for all cross-validation experiments.

BA metric is the average of the sensitivity and specificity. It was shown that it outperforms the traditional measure of classification accuracy [38]. The main target of this paper is to discover new epistasis interactions from large-scale genotyping. This can enhance the understanding of the biological mechanism of the disease and detect effective biomarkers that can contribute to the success of PM.

Table 1
Known AD association genes, and unknown but potential AD association genes.

Gene Name	Identified Genes	Discovered Genes
HPCAL1	Yes [40]	
CTNNA2	Yes [41]	
GRID2	Yes [42]	
CRYL1	Yes [43]	
ANK2	Yes [44]	
FHOD3	Yes [44]	
CPNE4	Yes [44]	
LINC02880		Yes
PDE1C		Yes
VAV3	Yes [44]	
F5	Yes [45]	
KIAA1217	Yes [45]	
TAFA2	Yes [45]	
IGF1R	Yes [45]	
ZFP90	Yes [45]	
LDLRAD4	Yes [45]	
CYP24A1	Yes [45]	
ARHGAP15	Yes [46]	
LOC105374122		Yes
LOC105374292		Yes
SLC2A9	Yes [47]	
NSUN7		Yes
TRPC3	Yes [48]	
PDE4D	Yes [45]	
LOC105379107		Yes
EBF2	Yes [49]	
EXTL3	Yes [50]	
ADCY8	Yes [51]	
LOC101929507		Yes
ELAVL2	Yes [52]	
SCUBE2	Yes [47]	
IGSF9B	Yes [53]	
GALNT8	Yes [54]	
LOC112268136		Yes
LINC01482		Yes
FAM20A	Yes [47]	
LINC01837		Yes
KRTAP27-1		Yes

4.7. Implementations

In this proposed work, I7 PC was used with the following software:

- PLINK version 1.07 [32] was used for filtering the SNPs and decreasing potential false findings. Hence, the SNPs with insufficient genotyping quality were eliminated.
- R version 3.6.3 [35]
- Python version 3.6.5 [39] is an open-source programming language that applies ensemble learning methods.
- MDR [37], version 3.0.2, is open-source software used to identify gene interactions in genetic association studies.

5. Results

This section presents TuRF feature selection method results, application results of different ensemble methods, and the discovered epistasis interactions. First, a statistical hypothesis test was applied using a p-value for reducing the massive number of SNPs to significant SNPs only. ADNI dataset was applied by detecting SNPs with a p-value < 0.01. Then, TuRF feature selection method was used on all 3,502 SNPs. Each was assigned a score depending on the SNP contribution to AD status. The number of SNPs after applying TuRF is 1,050. The achieved SNP subset became a suitable size for identifying significant gene-gene interactions. Then, different ensemble learning algorithms (RF with Gini index and permutation importance, XGBoost, and CART) were applied to discover interacting genetic attributes associated with AD. Finally, the predictive accuracy of 2, 3, 4, and 5-SNP interaction models were detected and evaluated using MDR [37]. The performance metrics of the ensemble learning algorithms are shown in Fig. 2.

Table 2 presents the top ten significant pairwise models with their BA model training, BA model testing, and p-values. The used metric of model fit was BA which described the average of the sensitivity and specificity. The results revealed that the training and testing accuracies are close. This explains the decrease in overfitting and the increase in generalizability [55]. The most robust two-way interaction was found between SNP rs17021105 from gene GRID2 and SNP rs17774281 near genes LINC01837 and LOC105372364) with BA model training of 0.6649, BA model testing of 0.6505, and a significance level of p-value 9.48E-08. These

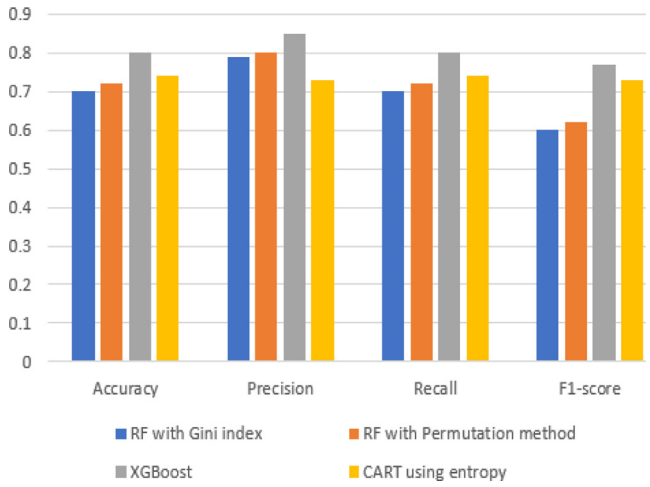


Fig. 2. The performance metrics of the ensemble learning algorithms.

Table 2
The top ten pairwise interaction models using MDR.

Model SNP(Gene)	BA model training	BA model testing	P-value
rs17021105, rs17774281 (GRID2, Near genes LINC01837 and LOC105372364)	0.6649	0.6505	9.48E−08
rs10961291, rs17565918 (LOC101929507, FHOD3)	0.6636	0.6398	9.69E−07
rs17557796, rs1405904 (near genes ELAVL2 and LOC105375992, –)	0.6594	0.6560	3.32E−06
rs528785, rs10862418 (–, –)	0.6573	0.6525	7.17E−07
rs17021105, rs1428896 (GRID2, –)	0.6568	0.6568	3.01E−07
rs17021105, rs1008975 (GRID2, EBF2)	0.6570	0.6518	5.40E−07
rs12621622, rs10862418 (HPCAL1, –)	0.6561	0.6561	6.14E−07
rs17021105, rs17557796 (GRID2, Near genes ELAVL2 and LOC105375992)	0.6556	0.6416	1.05E−06
rs528785, rs2505389 (–,–)	0.6545	0.6257	2.23E−06
rs17021105, rs1405904 (GRID2, –)	0.6515	0.6490	2.14E−06

results suggested that these two-way interactions are associated with AD disease.

Table 3 presents the ten most significant (p-value < 0.01) SNP trios with BA model training, BA model testing, and p-values. The most robust three-way interaction was found among (non-coding SNP rs13013095, SNP rs17021105 from gene GRID2, and non-coding SNP rs1428896) with BA model training of 0.7170, BA model testing of 0.6849, and a significance level of p-value 9.52E-09. These results suggested that these pure three-way synergistic effects among the three SNPs are associated with AD disease.

Table 3
The top ten 3-way interaction models using MDR.

Model	BA model training	BA model testing	P-value
rs13013095, rs17021105, rs1428896 (–, GRID2, –)	0.7170	0.6849	9.52E−09
rs7604762, rs17021105, rs1428896 (–, GRID2, –)	0.7144	0.7020	3.81E−08
rs11682196, rs17021105, rs17774281 (CTNNA2, GRID2, near genes LINC01837 and LOC105372364)	0.7143	0.6816	9.18E−09
rs17021105, rs17557796, rs17774281 (GRID2, near genes ELAVL2 and LOC105375992, near genes LINC01837 and LOC105372364)	0.7135	0.6931	1.41E−08
rs1545077, rs17021105, rs1428896 (NSUN7, GRID2, –)	0.7117	0.6957	1.55E−07
rs17021105, rs6486112, rs10784277 (GRID2, near genes SCUBE2 and LOC105376541, TAFA2)	0.7101	0.6934	8.97E−09
rs17021105, rs8071496, rs17774281 (GRID2, LINC01482, near genes LINC01837 and LOC105372364)	0.7093	0.6679	2.78E−09
rs528785, rs352098, rs17774281 (–, LINC02880, near genes LINC01837 and LOC105372364)	0.7084	0.6708	1.98E−08
rs7604762, rs17557796, rs907808 (–, near genes ELAVL2 and LOC105375992, IGF1R)	0.7088	0.6445	2.39E−07
rs17557796, rs1405904, rs907808 (near genes ELAVL2 and LOC105375992, –, IGF1R)	0.7071	0.6648	2.00E−07

Table 4 presents the top ten four-way interaction models with their BA model training, BA model testing, and p-values. The most significant four-way interaction was found among SNP rs17557796 near genes ELAVL2 and LOC105375992, non-coding SNP rs1405904, SNP rs907808 from gene IGF1R, and SNP rs4799866 from gene FHOD3 with BA model training of 0.7878, BA model testing of 0.6977, and a significance level of p-value 3.46E-06. These results suggested that these four-way interactions are associated with the disease.

Table 5 shows the top ten five-way interaction models with their BA model training, BA model testing, and p-values. The best five-way interaction was found among non-coding SNP rs787994, SNP rs7630827 from gene CPNE4, SNP rs17557796 near genes ELAVL2, and LOC105375992, SNP rs6486112 near genes SCUBE2 and LOC105376541, and SNP rs8097433 from gene FHOD3. In the 5-way interaction models, the best-achieved BA model training, BA model testing, and significance level (p-value) are 0.8758, 0.6126, and 7.07E-05, respectively.

6. Discussion

Exploring genetic markers associated with complex human diseases like AD help in better understanding the disease etiology to diagnose, treat, improve personalized therapeutic strategies, and

Table 4
The top ten 4-way interaction models using MDR.

Model	BA model training	BA model testing	P-value
1-rs17557796, rs1405904, rs907808, rs4799866 (Near genes ELAVL2 and LOC105375992, –, IGF1R, FHOD3)	0.7878	0.6977	3.46E−06
2-rs7604762, rs17557796, rs10784277, rs1405904 (–, Near genes ELAVL2 and LOC105375992, TAFA2, –)	0.7860	0.6492	3.04E−06
3-rs13013095, rs17557796, rs907808, rs4799866 (–, near genes ELAVL2 and LOC105375992, IGF1R, FHOD3)	0.7850	0.6587	4.35E−07
4-rs7604762, rs17557796, rs907808, rs4799866 (–, Near genes ELAVL2 and LOC105375992, IGF1R, FHOD3)	0.7844	0.6547	3.11E−06
5-rs7604762, rs17557796, rs1405904, rs907808 (–, Near genes ELAVL2 and LOC105375992, –, IGF1R)	0.7819	0.6715	2.20E−06
6-rs10084340, rs17557796, rs1405904, rs11857366 (ARHGAP15, near genes ELAVL2 and LOC105375992, –, IGF1R)	0.7824	0.6502	1.05E−06
7-rs10084340, rs17557796, rs1405904, rs907808 (ARHGAP15, near genes ELAVL2 and LOC105375992, –, IGF1R)	0.7817	0.6675	4.89E−07
8-rs13013095, rs17557796, rs907808, rs8097433 (–, near genes ELAVL2 and LOC105375992, IGF1R, FHOD3)	0.7816	0.662	9.35E−07
9-rs7604762, rs7630827, rs17557796, rs1405904 (–, CPNE4, near genes ELAVL2 and LOC105375992, –)	0.7815	0.6591	1.09E−05
10-rs7604762, rs17557796, rs17305480, rs907808 (–, near genes ELAVL2 and LOC105375992, CRYL1, IGF1R)	0.7811	0.6094	9.66E−08

Table 5
The top ten 5-way interaction models using MDR.

Model	BA model training	BA model testing	P-value
1-rs787994, rs7630827, rs17557796, rs6486112, rs8097433 (–, CPNE4, near genes ELAVL2 and LOC105375992, near genes SCUBE2 and LOC105376541, FHOD3)	0.8758	0.6126	7.07E−05
2-rs787994, rs352098, rs17557796, rs6486112, rs4799866 (–, LINC02880, near genes ELAVL2 and LOC105375992, near genes SCUBE2 and LOC105376541, FHOD3)	0.8771	0.6011	2.60E−05
3-rs787994, rs7630827, rs17557796, rs6486112, rs4799866 (–, CPNE4, near genes ELAVL2 and LOC105375992, near genes SCUBE2 and LOC105376541, FHOD3)	0.8717	0.6165	0.0001
4-rs7630827, rs17557796, rs2505389, rs1405904, rs8071496 (CPNE4, near genes ELAVL2 and LOC105375992, –, –, LINC01482)	0.8715	0.6450	3.15E−05
5-rs787994, rs352098, rs17557796, rs6486112, rs8097433 (–, LINC02880, Near genes ELAVL2 and LOC105375992, Near genes SCUBE2 and LOC105376541, FHOD3)	0.8740	0.6073	3.18E−05
6-rs7604762, rs7630827, rs17557796, rs1405904, rs4799866 (–, CPNE4, near genes ELAVL2 and LOC105375992, –, FHOD3)	0.8723	0.6509	2.75E−05
7-rs7604762, rs352098, rs17557796, rs6486112, rs4799866 (–, LINC02880, near genes ELAVL2 and LOC105375992, Near genes SCUBE2 and LOC105376541, FHOD3)	0.8709	0.5925	0.0001
8-rs787994, rs17557796, rs6486112, rs8097433, rs3013042 (–, near genes ELAVL2 and LOC105375992, Near genes SCUBE2 and LOC105376541, FHOD3, –)	0.8723	0.5830	9.08E−05
9-rs7630827, rs17557796, rs4880575, rs2505389, rs12587274 (CPNE4, near genes ELAVL2 and LOC105375992, –, –, –)	0.8673	0.6378	0.0007
10-rs17557796, rs2505389, rs10862418, rs1405904, rs8071496 (near genes ELAVL2 and LOC105375992, –, –, –, LINC01482)	0.8674	0.6815	0.0001

even prevent the disease. Given the complexity of AD disease, the causing factors are interactions among multiple genetic attributes instead of individual genetic variants. However, searching for combinations of features requires a significant challenge for genome-wide association research. In this work, we applied different ensemble learning algorithms to identify genetic attributes associated with AD. We performed data preprocessing and filtering on the ADNI dataset using TuRF feature selection method. We optimized the parameters for the applied ensemble learning methods for GWAS data analyses by parameter tuning.

The advantages and disadvantages of the applied ensemble learning techniques can be described as follows. We applied RF because it can handle several input variables and identify the most significant SNPs. This algorithm outputs the importance of the variable, which can be a significant SNP. RF can rank the SNPs using Gini importance or permutation methods. Further, it can deal with imbalanced dataset problems. The main disadvantage of RF is considered a black box technique as there is little control over what the model does. Hence, trying different parameters and random seeds is a suitable solution.

The advantage of using XGBoost is a highly flexible algorithm does not need normalized features and can handle missing data in the used dataset with its in-build features. Furthermore, it can output the importance of each SNP. The main disadvantage of XGBoost algorithm is the difficulty of interpretation. CART algorithm was applied as it needs minimal supervision and performs easy-to-understand models. We used this algorithm to find significant SNPs automatically. The disadvantage of using CART algorithm is having a limited number of positions suitable for accessible predictors.

Hence, we used ensemble learning algorithms to produce rankings on the importance of SNP contribution to the disease classification. By integrating the rankings provided by the applied algorithms, we identified a set of 76 top-ranked SNPs from both coding and non-coding regions of DNA. The coding SNPs mapped to 38 genes, including known AD association genes and unknown but potential AD association genes. Our results suggested novel genes associated with AD, as shown in Table 1. We further applied epistasis interaction analysis on the 76 SNPs using MDR to discover important pairwise, 3-way, 4-way, and 5-way interactions, as shown in the previous section. The performed analysis of the identified genes also suggests significant epistasis interactions that may explain the risk of AD. Gene information of these SNPs was detected using the NCBI database [56]. The achieved results of our work outperformed the results reported in [25–31]. This proposed work was not limited to examining the association of each SNP independently with the phenotype as reported in [25,26] but also focused on the interaction between multiple SNP loci up to fifth-order interactions. In this work, the same dataset (ADNI) used in [27] was used with the proposed paper, and the achieved results outperformed the results reported in [27].

We highlight some significant findings in the following paragraphs. One of the most repeated interactions is between IGF1R and FHOD3 in models 1,3,4, and 8, as observed in Table 4. Also, it was shown that IGF1R gene is repeated with ARHGAP15 in models 6 and 7, as observed in Table 4.

It was shown that gene TAFA2 has strong three-way interactions with gene GRID2 and a SNP rs6486112 near genes SCUBE2 and LOC105376541. TAFA2 has significant 4-way interactions with non-coding SNP rs7604762, SNP rs17557796 Near genes ELAVL2 and LOC105375992, gene LOC105375992, and non-coding SNP rs1405904. This paper observed that gene LINC01482 has strong 5-way interactions with gene CPNE4, SNP rs17557796 near genes ELAVL2 and LOC105375992, non-coding SNP rs2505389, and non-coding rs1405904). Also, there are significant interactions between LINC02880 gene and FHOD3 gene in models 2,5 and 7, as shown in Table 5.

In this work, statistical analysis was performed using statistical significance, which is the main criterion for interpreting our results. A p-value less than 0.01 was used as a significance level. The proposed framework presents the most significant epistasis interactions (p-value < 0.01). Hence, the achieved results are statistically significant.

Discovering novel risk genes and epistasis interactions are beneficial for clinical and public health practices to assist disease diagnosis, predict disease risk, guide patient care, and facilitate the development of suitable drug discoveries. Furthermore, we suggest future directions by linking clinical, genetic, and other biomarkers data such as PET, MRI, and CSF assays to define disease staging and possibly help detect etiology. This combination may shape the future therapeutic approach toward pre-symptomatic PM for AD prevention.

7. conclusions

In this paper, MDR constructive induction algorithm was integrated with ensemble learning algorithms to discover epistasis

interactions in a computationally efficient method. The proposed framework was implemented and applied to ADNI dataset, and the results of the most significant two-, three-, four-, and five-way interactions are shown. Discovering new interactions from large-scale genotype data is an important research goal. This paper aims to reduce high dimensionality and improve the performance by selecting a subset of powerful SNP features from a dataset with many features using feature selection methods. The results demonstrate that the proposed framework can detect feature subsets more efficiently and improve classification performance.

It was shown that machine learning techniques have great abilities in modeling difficult relationships between many attributes. In GWAS, suitable algorithms are needed for discovering the non-linear, non-additive interactions between several genetic factors that may contribute to the complex disease outcome. This paper demonstrates a novel framework using different ensemble learning methods to search for epistasis interactions associated with AD. We hope our work will help better understand AD disease etiology and facilitate the development of PM approaches and suitable drug discoveries. In this paper, the most significant interaction models associated with the disease were identified. The five-way interaction models were identified with classification accuracy varied between (0.8674–0.8758). While the accuracy of the two-way, three-way, and four-way models varied between (0.6515–0.6649), (0.7071–0.7170), and (0.7811–0.7878), respectively.

Integrating constructive induction algorithms like MDR with ensemble learning algorithms in the proposed framework makes it robust, applicable, and reliable to create models for other risky diseases. In This paper, there are 38 genes mapped from the 76 SNPs, including genes that have been already identified previously for AD disease and the newly discovered genes. These determined and discovered genes can help to explore significant interactions among them.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Data used in the preparation of this paper were obtained from the ADNI database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this paper.

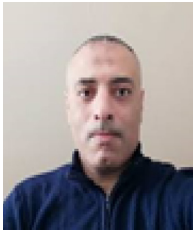
References

[1] Xie F, Chan JCN, Ma RCW. Precision medicine in diabetes prevention, classification and management. J Diabetes Investig 2018;9(5):998–1015.
[2] Hamid MM, Ali N, Saad M, Mabrouk M, Shaker O. Multiple sclerosis: an associated single-nucleotide polymorphism study on Egyptian population. Network Modeling Analysis in Health Informatics and Bioinformatics 2020;9(1):1–7.
[3] Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. Front Genet 2015;6:285.
[4] Weigelt B, Reis-Filho JS. Epistatic interactions and drug response. J Pathol 2014;232(2):255–63.
[5] Moore JH, Williams SM. Epistasis and Its Implications for Personal Genetics. Am J Hum Genet 2009;85(3):309–20.
[6] Bron EE, Smits M, Niessen WJ, Klein S. Feature selection based on the SVM weight vector for classification of dementia. IEEE J Biomed Health Inform 2015;19(5):1617–26.
[7] Urbanowicz RJ et al. Benchmarking relief-based feature selection methods for bioinformatics data mining. J Biomed Inform 2018;85:168–88.
[8] Brunese L, Mercaldo F, Reginelli A, Santone A. An ensemble learning approach for brain cancer detection exploiting radiomic features. Comput Methods Programs Biomed 2020;185:105134.
[9] Huynh-Thu VA et al. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics 2012;28(13):1766–74.
[10] Menze BH et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinf 2009;10(1):1–16.
[11] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining 2016; 785-794.
[12] Yang L, Liu S, Tsoka S, Papageorgiou LG. A regression tree approach using mathematical programming. Expert Syst Appl 2017;78:347–57.
[13] Altmann A et al. Permutation importance: a corrected feature importance measure. Bioinformatics 2010;26(10):1340–7.
[14] Mostafa M, Mabrouk MS, Omar YMK. Identifying genetic biomarkers associated to Alzheimer's Disease Using Support Vector Machine. 2016 8th CIBEC IEEE 2016:5–9.
[15] Mostafa MM et al. Machine learning for detecting epistasis interactions and its relevance to personalized medicine in Alzheimer's disease: systematic review. Biomed Eng 2021;33(6):2150047.
[16] Karch CM, Goate AM. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. Biol Psychiatry 2015;77(1):43–51.
[17] Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 2002;11(20):2463–8.
[18] Uppu S, Krishna A, Gopalan RP. Towards Deep Learning in genome-Wide Association Interaction studies. PACIS; 2016.
[19] Zhou T, Thung KH, Zhu X, Shen D. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum Brain Mapp 2019;40(3):1001–16.
[20] Zhou T et al. Hi-net: hybrid-fusion network for multi-modal MR image synthesis. IEEE Trans Med Imaging 2020;39(9):2772–81.
[21] Zhou T et al. Latent representation learning for Alzheimer's disease diagnosis with incomplete multimodality neuroimaging and genetic data. IEEE Trans Med Imaging 2019;38(10):2411–22.
[22] Mostafa M et al. Discovering epistasis interactions in Alzheimer's disease using deep learning model. Gene Reports 2022;29:101673.
[23] Dunn AR, O'Connell KMS, Kaczorowski CC. Gene-by-environment interactions in Alzheimer's disease and Parkinson's disease. Neurosci Biobehav Rev 2019;103:73–80.
[24] Dorani F, Hu T, Woods MO, Zhai G. Ensemble learning for detecting gene-gene interactions in colorectal cancer. PeerJ 2018:e5854.
[25] De Velasco OJ et al. Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. BMC Bioinf 2019;20(1):1–17.
[26] Mostafa M, Mabrouk M. Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer's disease using machine learning techniques. Biomed Eng: Appl Basis Commun 2019;31(5).
[27] Sherif FF, Zayed N, Fakhr M, Wahed MA, Kadah YM. Integrated higher-order evidence-based framework for prediction of higher-order epistasis interactions in Alzheimer's disease. Int J Biol Biomed Eng 2017;11:16–24.
[28] Chen H, He Y, Ji J, Shi Y. A Machine Learning Method for Identifying Critical Interactions Between Gene Pairs in Alzheimer's Disease Prediction. Front Neurol 2019;10:1162.
[29] Chang Y-C, Wu J-T, Hong M-Y, Tung Y-A, Hsieh P-H, Yee SW, et al. GenEpi: gene-based epistasis discovery using machine learning. BMC Bioinf 2020;21(1):68.
[30] Orlenko A, Moore JH. A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. BioData mining 2021;14(1):1–17.
[31] Petrelis AM, Stathopoulou MG, Kafyra M, Murray H, Masson C, Lamont J, et al. VEGF-A-related genetic variants protect against Alzheimer's disease. Aging 2022;14(6):2524–36.
[32] Purcell S. PLINK (1.07). Documentation; 2010, pp. 1–293.
[33] Lehne B, Lewis CM, Schlitt T, Khanin R. From SNPs to genes: disease association at the gene level. PLoS ONE 2011;6(6):e20133.
[34] Wang Y-T, Sung P-Y, Lin P-L, Yu Y-W, Chung R-H. A multi-SNP association test for complex diseases incorporating an optimal P-value threshold algorithm in nuclear families. BMC Genom 2015;16(1):1–10.
[35] Lantz Brett. Machine learning with R. Packt publishing ltd; 2013.
[36] Dietterich TG. Ensemble methods in machine learning. International workshop on multiple classifier systems. Berlin, Heidelberg: Springer; 2000. p. 1–15.
[37] Wu Y, Zhang L, Liu L, Zhang Y, Zhao Z, Liu X, et al. A multifactor dimensionality reduction-logistic regression model of gene polymorphisms and an environmental interaction analysis in cancer research. Asian Pac J Cancer Prev 2011;12(11):2887–92.
[38] Velez DR et al. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet Epidemiol 2007;31(4):306–15.
[39] Pedregosa F et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[40] Braunewell KH. The visinin-like proteins VILIP-1 and VILIP-3 in Alzheimer's disease—old wine in new bottles. Front Mol Neurosci 2012;5(20).
[41] Kawalia SB et al. Analytical strategy to prioritize Alzheimer's disease candidate genes in gene regulatory networks using public expression data. J Alzheimers Dis 2017;59(4):1237–54.

[42] Pérez-Palma E, Bustos BI, Villamán CF, Alarcón MA, Avila ME, Ugarte GD, et al. Overrepresentation of glutamate signaling in Alzheimer's disease: network-based pathway enrichment using meta-analysis of genome-wide association studies. PLoS ONE 2014;9(4):e95413.
[43] Tindale LC et al. Lipid and Alzheimer's disease genes associated with healthy aging and longevity in healthy oldest-old. Oncotarget 2017;8(13):20612.
[44] Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res 2012;40(D1):D1047–54.
[45] https://www.targetvalidation.org/ (last seen 2020).
[46] Wang H, Bennett DA, De Jager PL, Zhang Q-Y, Zhang H-Y. Genome-wide epistasis analysis for Alzheimer's disease and implications for genetic risk prediction. Alzheimer's Res Therapy 2021;13(1):1–13.
[47] http://twas-hub.org/genes/ (last seen 2020).
[48] Selvaraj S, Sun Y, Singh BB. TRPC channels and their implications for neurological diseases. CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders) 2010;9(1):94–104.
[49] Altuna M, Urdánoz-Casado A, Sánchez-Ruiz de Gordoa J, Zelaya MV, Labarga A, Lepesant JMJ, et al. DNA methylation signature of human hippocampus in Alzheimer's disease is linked to neurogenesis. Clinical 2019;11(1):1–16.
[50] Yamada S. Specific functions of Exostosin-like 3 (EXTL3) gene products. Cell Mol Biol Lett 2020;25(1):1–12.
[51] Silver M, Janousova E, Hua X, Thompson PM, Montana G. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. Neuroimage 2012;63(3):1681–94.
[52] Wang L, Cukier HN, Rajabli F, Hofmann NK, Adams LD, Rodriguez VC, et al. Functional analysis of candidate genes identified through whole genome sequencing in Caribbean Hispanic families for late-onset Alzheimer disease. Alzheimer's & Dementia 2020;16(S3).
[53] Zhang Qi, Ma C, Chin L-S, Li L. Integrative glycoproteomics reveals protein N-glycosylation aberrations and glycoproteomic network alterations in Alzheimer's disease. Sci Adv 2020;6(40).
[54] Kim Bo-Hyun et al. Identification of Novel Genes Associated with Cortical Thickness in Alzheimer's Disease: Systems Biology Approach to Neuroimaging Endophenotype. J Alzheimer's Disease Preprint 2020;75(2):531–45.
[55] Moore JH, Andrews PC. Epistasis analysis using multifactor dimensionality reduction. In: Epistasis. Springer; 2015. p. 301–14.
[56] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29(1):308–11.

Marwa M. Abd El Hamid received her B.SC degree from Computer Science at El-Shorouk Academy in 2012. She received her M.Sc. from the College of Computing and Information Technology at Arab Academy for Science, Technology and Maritime Transport (AASTMT) in 2017, where she is currently pursuing a Ph.D. degree in bioinformatics. She is currently a teaching assistant in the Department of Computer Science at El-Shorouk Academy. She has several publications in bioinformatics, machine learning, and text mining fields. She is also a reviewer for several international journals and conferences throughout her career.

Mohamed Shaheen Elgamel received the B.Sc. degree in computer science from Alexandria University, Egypt, 1991, the M.Sc. degree in computer engineering from Arab Academy for Science, Technology and Maritime Transport(AASTMT), Alexandria, Egypt, 1998, the M.Sc. and the Ph.D. degrees in computer engineering from University of Louisiana at Lafayette, Louisiana, USA, in 2000 and 2003 respectively. He is currently a Professor at the College of Computing and Information Technology, AASTMT since 2010. From 2003-2010, he has been a Graduate Faculty at the University of Louisiana at Lafayette, USA. From 1999 to 2003, he served as an Adjunct Faculty at Computer Science Department, University of Louisiana at Lafayette, USA. His research interests include Software Engineering, Wireless Sensor Networks and Artificial Intelligence. His research is supported by funds from federal agents like DOE, NSF, the Louisiana Governor ITI, the Qatar National Research Fund ("QNRF"), and ITIDA, Egypt.



Yasser M.K. Omar is currently a Professor at AASTMT's department of Computer Science and Engineering's Computer Science College of Computing and Information Technology. His research interests are in artificial intelligence, namely agent technology, multi–agent systems, and machine learning. He has published several articles in various areas.

Mai S. Mabrouk received her B.Sc., M.Sc. and PhD degrees from the Biomedical Engineering Department at Cairo University in 2000, 2004 and 2008 respectively. She is currently an Associate professor of and department head of Biomedical Engineering at Misr University for Science and Technology. Her biography was selected to appear in Marquis Who's Who in the World in 2012. Along her career, she was a technical reviewer and editorial board member for several international journals and conferences. She published over 60 peer-reviewed journal and conference articles in the areas of medical imaging processing, Bioinformatics and human computer interface.