Gene Reports 29 (2022) 101673

Contents lists available at ScienceDirect

Gene Reports

journal homepage: www.elsevier.com/locate/genrep

# Discovering epistasis interactions in Alzheimer's disease using deep learning model

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

Edited by Dominic Voon

*Keywords:*
Alzheimer's disease
Epistasis interactions
Personalized medicine
Deep learning model
SHAP

ABSTRACT

Alzheimer's disease (AD) is the most common form of dementia. Single Nucleotide Polymorphisms (SNPs) are single nucleotide alterations that can be used as genomic markers disclosing susceptibility to complex diseases like AD. Epistasis has long been significant for recognizing the function of genetic pathways and the evolutionary dynamics of difficult genetic systems. Discovering epistasis interactions holds a vital key to personalized medicine (PM). PM needs a better understanding of the relationship between human genetic data and complex diseases. In this proposed work, a deep neural network (DNN) is applied using SHapley Additive exPlanations (SHAP) to get top 20, 100, 300, and 500 ranking SNPs responsible for AD risk through epistasis interactions. Multi-locus interaction analysis is performed on these identified SNPs using Multifactor Dimensionality Reduction (MDR). This constructive induction algorithm is integrated with DNN for discovering epistasis interactions in a computationally effective method. The proposed framework is applied to Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The best accuracies are achieved using the top 500 SNPs, and the classification accuracies varied between 0.860 and 0.874 in the five-way interaction model. However, the classification accuracies of 2-way, 3-way, 4-way models varied between 0.663 and 0.670, 0.718 and 0.727, and 0.793 and 0.803, respectively. The results revealed that the reported accuracy scores of the proposed framework outperform the referenced literature work. The proposed framework presents high-ranked risk genes and promising epistasis interactions that may help in explaining the risk of AD.

## 1. Introduction

Alzheimer's disease (AD) is a complex disorder that results in the degeneration of brain cells. This disease is the major cause of dementia, known for troubles in memory, language, problem-solving and thinking abilities. AD has been considered a multifactorial disorder associated with various risk factors like genetic factors, increasing age, injuries in the head, environmental factors, and vascular diseases. The underlying cause of pathological changes in this complex disease is still unknown. This complex disease is one of the top causes of death (Breijyeh and Karaman, 2020).

The genetic mechanisms underlying biological traits are complicated, encompassing the effects of various genetic variants. Epistasis is the interactions between these variants (Schmalohr et al., 2018). Epistasis, the interaction between different genes, is a vital topic of current interest in complex disease genetics. The complex diseases are multiple sclerosis, diabetes, AD, and asthma. Epistatic describes the masking effect in which a variant or allele at one locus stops the variant at another locus from appearing its effect. Discovering gene-gene interactions is vital for understanding the disease mechanism and developing personalized medicine (Ho et al., 2019).

Since we live in the era of big data, converting big biomedical data

*Abbreviations:* AD, Alzheimer's disease; SNPs, Single Nucleotide Polymorphisms; PM, personalized medicine; DNN, deep neural network; SHAP, SHapley Additive exPlanations; MDR, Multifactor Dimensionality Reduction; ADNI, Alzheimer's Disease Neuroimaging Initiative; GWAS, genome-wide association studies; A, adenine; G, guanine; T, thymine; C, cytosine; NB, naïve Bayes; TAN, tree augmented naïve Bayes; QC, quality control; LD, linkage disequilibrium; BA, balanced accuracy.

* 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), yasser.omar@aast.edu (Y.M.K. Omar), mohamed.shaheen@aast.edu (M. Shaheen), msm_eng@yahoo.com (M.S. Mabrouk).

https://doi.org/10.1016/j.genrep.2022.101673
Received 25 March 2022; Received in revised form 31 July 2022; Accepted 24 August 2022
Available online 27 August 2022

into constructive knowledge has been one of the most difficulties in bioinformatics. Simultaneously, deep learning (DL) techniques play an important role in several fields and achieve high performance. Hence, using DL in bioinformatics to get insights from data is essential in recent research (Min et al., 2017). It is noticeable that applying DL to predicting and identifying individuals at risk of developing AD has recently gained considerable attention through its combination with biomarkers. Hence, DL can play an important role in clinical purposes. Deep learning can deal with big and deep biomedical data, which is an essential starting step within a personalized medicine strategy.

DL is a kind of machine learning that is a subset of artificial intelligence. DL uses an artificial neural network of multiple nonlinear layers. The key characteristics of DL are that main features are not determined by human engineers, but learned from the data themselves. DL can learn and explore hierarchical representations of data with a growing level of abstraction as one of the representation learning methods (Berrar and Dubitzky, 2021). Deep learning allows computational models comprising several processing layers for learning representations of data with many levels of abstraction. These methods have improved the state-of-the-art in various domains like visual object recognition, object detection, speech recognition, drug discovery, and genomics (Goodfellow et al., 2016).

The main goal of this proposed framework is to explore significant SNPs responsible for the disease risk through epistasis interactions. Discovering epistasis interactions will give essential insight into complex disease mechanisms and facilitate PM. This paper integrates MDR with DNN to discover epistasis interactions in a computationally effective method. The following paper sections are organized as follows; Section 2 presents the medical background. Section 3 shows the literature review. Section 4 explains the materials and methods. Section 5 describes the results and discussions. Finally, Section 6 presents the conclusions.

## 2. Medical background

Dementia, including AD, has a vital negative effect on individuals' functioning, independence, and the demand for care. Dementia is one of the major health challenges of existing times. It places an intensive load on families and the community, with the expense of care often paid for out-of-pocket. There are massive personal, economic, and social results of dementia (Meyer et al., 2016). Genome-wide association studies (GWAS) played a vital role in detecting the associated genetic variants of complex diseases. A genetic variant can be existed because of an alteration in single nucleotide adenine (A), guanine (G), thymine (T), or cytosine (C) in a certain stretch of DNA (Bailey, 2007). The genetic variants appear throughout a person's DNA and are usually referred to as SNPs. GWAS concentrates on a single-locus approach that aims to investigate each SNP at a time and its association with disease (Niel et al., 2015). However, complex disorders may not have a comprehensible appearance. Hence, the disease could be produced by nonlinear interactions of genetic or environmental factors or both together (Niel et al., 2015).

During the past decade, GWAS have played an important role in discovering genotype-phenotype associations. In GWAS analyses, geneticists focus on DNA polymorphism markers for detecting these associations (Kim et al., 2013). SNP is one of the vital classes of genetic markers that allow the comparison of allelic frequencies between cases and controls. SNPs are examined one by one for statistical association with complex diseases in the standard approach (Patron et al., 2019). Genetic variants have independent effects on the phenotype. Hence, only additive effects are related to this method. This type of analysis has been long-established; however, results are not fascinating as expected. In the strategy of considering one locus at a time, only a small part of the genetic variance interprets the phenotype, the other part apply to missing heritability (Simons et al., 2018). It has been known that missing heritability is partly because of genetic variants presenting effects when they interact with one other variant or more. Epistasis is biologically relevant to complex diseases and refers to the combinatorial effect of the interaction between two or more genes. The study of genetic or epistatic interactions between SNPs is a topic of interest, as these interactions are key to understanding how genes relate functionally (Moore and Williams, 2009).

In explaining the impact of genetic factors on phenotype variation, epistasis interactions (non-additive genetic interactions) must be considered. However, there is a shortage of techniques that can efficiently explore such interactions. Analysis methods that permit or exploit the phenomenon of epistasis are obviously of increasing significance in the genetic dissection of complicated diseases (Cordell, 2002). This paper aims to identify genetic variants that may otherwise have remained undiscovered by permitting epistatic interactions between potential disease loci. Personalized medicine, known as precision medicine, uses the patient's genetic profile to guide decisions to tailor the right therapeutic strategy for the individual or to define the predisposition to the complex disease. Discovering epistasis interactions associated with the complex disease will pave the way for PM and enhance personalized intervention strategies (Dunn et al., 2019). However, PM needs a better understanding of the disease mechanism, and its success depends on the accurate identification of genetic biomarkers. Unfortunately, detecting epistasis interactions is still under research and poorly understood.

## 3. Literature review

Many studies applied several methods for analyzing the individual effect of each SNP and detecting the significant SNPs associated with complex diseases. Different analysis techniques can explore genetic variants associated with complex diseases. A uni-variable analysis approach can examine the association of each SNP independently with the phenotype (Abd El Hamid et al., 2021). A multi-variable analysis approach can capture the interactions between many SNPs and better explain complex diseases' susceptibility (Dorani et al., 2018). Discovering gene-gene interactions is essential to investigate the disease mechanism, and discovering them is still under research. Some diseases like AD are caused by complex interactions among several genes known as epistasis interactions (Dunn et al., 2019).

Romero-Rosales et al. (2020) applied three machine learning techniques that have been proved to build powerful predictive models (genetic algorithms, LASSO, and step-wise). The research method contains procedures for obtaining clinical and genotypic data, filtering data using quality control criteria, reducing dimensionality, analyzing and comparing the models. The authors used the National Institute on Aging—Late-Onset Alzheimer's Disease Family Study. The dataset comprises 5220 subjects and 620,901 SNPs. The results revealed that LASSO models achieved the best accuracy: 0.801, sensitivity: 0.798, and specificity: 0.804, respectively. However, the research gave no attention to exploring genetic interactions.

Sherif et al. (2017) applied different techniques like naïve Bayes, Support vector machine, k-nearest neighbor, logistic regression, random forest, and MDR classifiers. They applied a framework to identify epistasis interactions and improve early AD diagnosis. They used Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. It includes 730,525 SNPs for 125 normal persons and 306 AD patients. MDR achieved the best results compared to the other methods. The achieved classification accuracies of their work varied between 0.7410 and 0.7860. However, the achieved accuracy needed to be improved for better investigation of the disease.

Abd El Hamid et al. (2019) applied some techniques like sequential minimal optimization algorithm with different kernels, naïve Bayes (NB), tree augmented naïve Bayes (TAN), and K2 learning algorithms. The used dataset was whole-genome sequencing dataset that includes 2,379,855 SNPs for 282 normal people, 442 mild cognitive impairment and 48 AD cases. The main target of that research is exploring the best

SNPs associated with the disease. The best results were achieved by NB and K2 learning algorithms with an accuracy of 98 % and 98.40 %, respectively. Even so, the research ignored interaction effects between SNPs.

Chang et al. (2020) present GenEpi, a computational package to uncover epistasis interactions associated with phenotypes. That research aims to discover SNP interactions by building GenEpi package to reveal epistasis interactions associated with the phenotype using machine learning. They applied for their work on an AD cohort used in Alzheimer's disease Dream Challenge. The used cohort includes 767 participants of cases and controls from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The authors defined the interactions between two SNPs only. This study explores the capability of GenEpi in finding disease-related variants and epistasis interactions. However, that paper did not detect higher-order interactions of genes in their analysis.

Several pieces of research concentrated on uni-variable analysis only and inspected the effect of independent SNP loci for identifying genetic variants associated with complex diseases like AD (Romero-Rosales et al., 2020; Abd El Hamid et al., 2019). However, GWAS techniques for detecting epistasis, the interactions between genetic variants associated with phenotypes are still limited. Therefore, discovering multi-locus interactions is essential because they may have more robust associations. The main goal of this paper is to fill the notable gap in the previous research to discover important epistasis interactions up to fifth-order interactions associated with the disease. This can improve the investigation of the biological disease mechanism and identify significant biomarkers that may contribute to the success of personalized medicine (Xie et al., 2018).

## 4. Materials and methods

Fig. 1 demonstrates an overall block diagram of the proposed model. First, the dataset was obtained from ADNI database (Carrillo et al., 2012) and went through consecutive data preprocessing steps to achieve significant SNPs. We used ADNI dataset because it is big and important dataset. ADNI is a global dataset comprises several types of data from study volunteers throughout their participation in the study, using a standard set of protocols and procedures to eliminate inconsistencies and help investigate mechanisms of the disease. The main goal of ADNI study is to support advances in AD intervention, prevention, and treatment through the application of new diagnostic methods.

After completing the data preprocessing phase, DNN was applied using SHAP to detect top-ranking SNPs responsible for AD risk through epistasis interactions. To address the black-box style of the DNN, some techniques like SHAP can create an explanation for this complex model. In this paper, the global level of the model was applied based on aggregations of Shapley values to explain the overall workings. This phase aims to achieve the SHAP feature importance for DNN. The features were sorted by decreasing importance. Then Multi-locus interaction analysis was performed on these identified SNPs using MDR to discover important epistasis interactions. 10-fold cross-validation was used to evaluate the predictive accuracy of all exhaustive 2-, 3-, up to 5-SNP combination models. Finally, the proposed framework presents significant risk genes and epistasis interactions associated with AD.

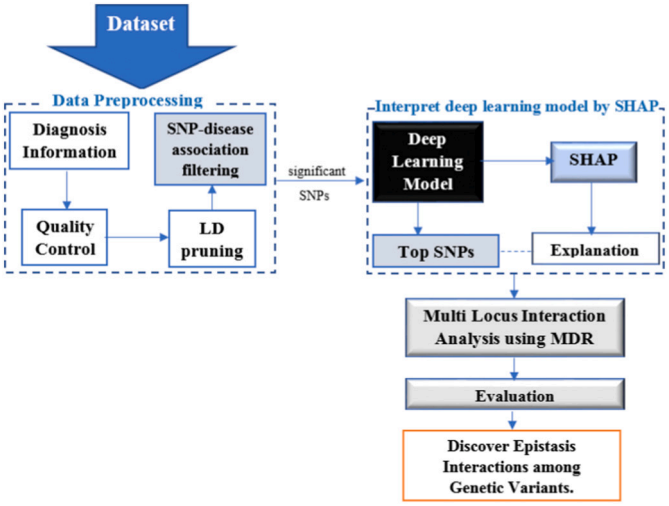


**Fig. 1.** Proposed framework.

The main goal of this paper is to focus on only cases and controls. Hence, we conducted a case-control study to detect high-ranked risk genes and promising epistasis interactions that may help better understand the disease etiology.

### 4.1. Dataset

The used dataset was obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Carrillo et al., 2012). It contained total genotypes for 431 individuals. These individuals were 127 normal individuals and 304 patients. ADNI dataset contained total genotypes of 730,524 SNPs.

### 4.2. Data preprocessing

The data preprocessing phase is an essential phase to achieve meaningful results. In this work, many preprocessing steps were applied as follows: In the first step, the diagnostic information was added to identify the phenotype information for each individual as a case or control. Therefore, the total number of unaffected individuals was 127, while the number of affected individuals was 304. In the second step, quality control (QC) procedures were applied to the dataset to exclude low-quality SNPs and minimize potential false findings. QC steps were applied using PLINK (Purcell, 2012) as follows:

a) People with too much missing genotyping data (10 % missing) were excluded (Purcell et al., 2007).
b) SNPs with a missing genotype rate (10 % missing) were excluded (Purcell, 2012). Only SNPs with 90 % genotyping rate are taken into consideration.
c) SNPs with minor allele frequency $< 10$ % were also excluded (Purcell et al., 2007).

After applying QC steps, the total number of individuals is 431 (304 cases and 127 controls). While, the number of SNPs after QC procedures became 530,750. In the third step, the linkage disequilibrium (LD) pruning phase was applied to enhance the power of complex disease genetic association studies. The LD pruning step was applied on the ADNI dataset to select Informative Markers, leaving 447,538 markers. In the fourth step, SNP-disease association tests (Lehne et al., 2011) were applied to decreasing the huge computational requirements. In this work, three SNP-disease association tests implemented in PLINK were applied. In this paper, the goal of applying independent SNP-disease association tests is to assess the statistical association of each of the 447,538 SNP with the disease.

These SNP-disease association tests are the basic association test, logistic model, and Fisher's exact (allelic association). First, the non-significant SNPs with a $p$-value threshold of over 0.01 were discarded. The $p$-value threshold was used as a significance level to reveal the SNP associations. This paper chose the important SNPs with a $p$-value $< 0.01$ threshold because this threshold has more discriminating power than 0.05 threshold (Wang et al., 2015). The results revealed that the total number of SNPs became 4383, 3863, and 3861 using the basic association test, logistic model, and Fisher's exact test, respectively. Finally, the intersection of the SNPs results from the applied SNP-disease

association tests are achieved using R (Bischl et al., 2016) to get 3502 significant SNPs. The result of the obtained significant SNPs decreases the false-positive association with AD. However, the achieved number of SNP is still huge. Hence, it is important to apply feature selection techniques for detecting the most significant SNPs.

### 4.3. Interpret deep learning model by SHAP

Deep learning plays an important role and makes major progress in fixing problems that have resisted the superior attempts of the artificial intelligence community for several years. However, despite deep neural networks working successfully in several prediction tasks, their black-box style restricts their usage in applications demanding the model explanation, such as GWAS (Cui et al., 2021). With the good progression in interpretable neural networks, interaction effects between attributes can be achieved by evaluating well-principled interaction scores like Shapley interaction scores (Wang et al., 2019). In this paper, the deep learning technique was interpreted using the famous library SHapley Additive exPlanations (SHAP) to assign feature importance to each attribute.

In this work, SHAP was used to evaluate SNPs from a deep learning model trained on the ADNI dataset that may contribute to the risk of AD through epistasis interactions. SHAP is a unified framework for interpreting predictions used to assign each attribute an importance value for a certain prediction (Dickinson and Meyer, 2021). In addition, SHAP was used to assign a numeric measure of credit to each input attribute. This work created a deep learning neural network model in Python using Keras. Keras is a powerful open-source Python library to develop and evaluate deep learning models. Keras wraps the effective numerical computation libraries like Theano and TensorFlow and defines and trains neural network models.

This proposed work used a fully connected network structure with three layers. Fully connected layers are established using the Dense class (Schwing and Urtasun, 2015). Deep learning model was applied with one input layer, two hidden layers (h1 and h2), and one output layer. The model expects rows of data with 3502 features (the input_dim = 3502 argument). The first hidden layer had 16 nodes and used the sigmoid activation function. The second hidden layer had eight nodes and used the sigmoid activation function. Finally, the output layer has one node and used the sigmoid activation function. Fig. 2 shows deep neural network model graph.

DNN was used as the classification method in this work, and its performance is assessed using performance metrics like classification accuracy, precision, recall, and f1-score. After applying the DNN on the ADNI dataset, a robust model characterized was obtained with a classification accuracy of 70.53 %, precision of 66 %, recall of 100 %, and f1-score of 79 %. The main idea of using SHAP feature importance is to detect the important features with large absolute Shapley values. It was applied as a global interpretability method using the function: shap.summary_plot (Molnar, 2020). As we need global importance, we average the absolute Shapley values per attribute across the data. Then, features are sorted in decreasing order by Shapley values. Fig. 3 presents the top 20 features that contribute to helping deep learning algorithms construct decisions.

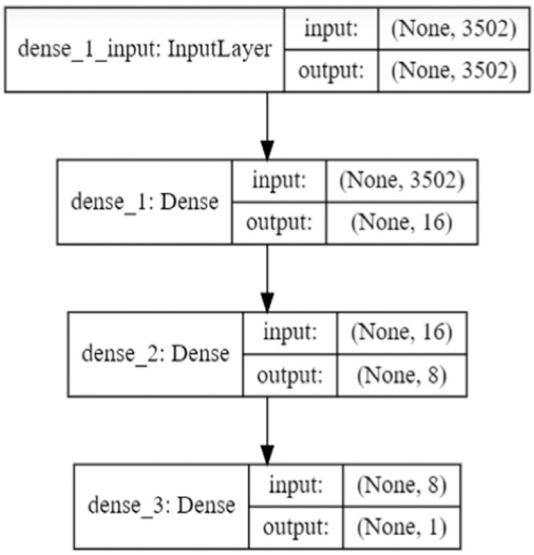


**Fig. 2.** A deep neural network model graph.

### 4.4. Multi-locus interaction analysis using MDR

In this paper, multi-locus interaction analysis was performed on the identified SNPs described in the previous section using MDR for discovering significant epistasis interactions. MDR is designed specifically to detect and interpret gene-gene interactions. The key concept of MDR is a feature construction method that constructs a new attribute by pooling as genotypes from many SNPs (Wu et al., 2011). The procedure of describing a new feature as a function of two or more other features is called attribute construction (constructive induction).

MDR concentrates on the n-dimensional array of genotypes for n variants and their interaction with the complex disease. As the three genotypes per variant, the whole number of genotype arrays is calculated by $3^n$. Each pattern is categorized into low-risk or high-risk groups, relying on a threshold ratio of patient versus normal individuals carrying that pattern (Okazaki et al., 2021). Hence, the analysis problem is reduced effectively from n-dimensions to one dimension. The cross-validation statistical technique is applied to optimize the prediction accuracy of individuals categorized into patients or normal individuals. The outcome models are ranked depending on overall balanced accuracy, which balances between high power and low *p*-value (Velez et al., 2007).

In a GWAS, applying an exhaustive search to detect epistasis interactions is computationally expensive. This computational load is a major problem. Furthermore, the task will be more complex for larger order interactions and larger markers. When the number of markers is huge, the number of multi-locus interactions increases. Hence, we propose a novel approach combining (DNN using SHAP) and MDR methods to decrease some shortcomings of the MDR method by detecting the top-

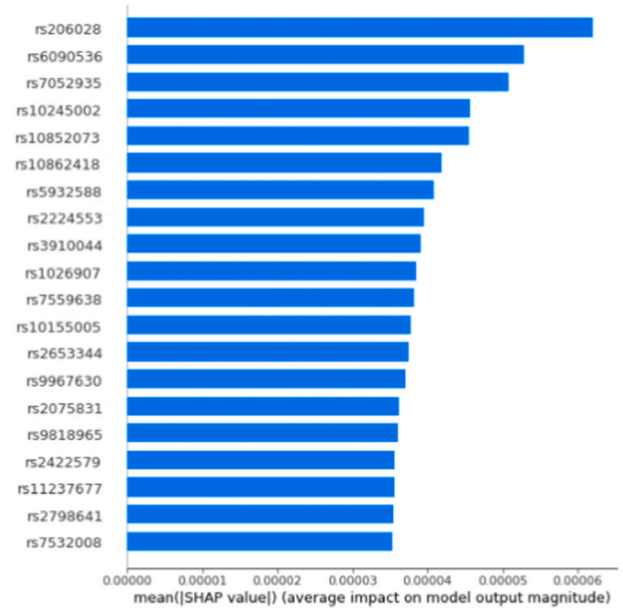


**Fig. 3.** SHAP feature importance.

ranking SNPs.

Hence, this proposed framework presents many approaches for discovering new genes and searching for significant epistasis interactions associated with AD. First, A deep learning algorithm was applied using SHAP for generating top 20, 100, 300, and 500 ranking SNPs. Then, MDR was used for identifying pairwise, 3-way, 4-way, and 5-way interactions models associated with AD. Next, searching for significant interactions models within the dataset was applied to the interactions of the top 20, 100, 300, and 500 rankings. Applying deep learning using SHAP provided rankings on the importance of SNP contribution to AD classification. After selecting the top 20 rankings in the first applied approach, these SNPs' statistical interaction analysis was generated to explore critical gene-gene interactions. The coding SNPs mapped to 11 genes which consists of known AD-related genes (6 identified genes) and genes that have not been discovered previously associated with AD (5 discovered genes), as shown in Table 1. The novel explored genes that can be possible risk association genes with the disease.

The balanced accuracy (BA) metric was used as an assessment metric to deal with imbalanced data. Balanced accuracy (BA) is a metric that can be used when one class appears much more than the other (Wu et al., 2011).

### 4.5. Evaluation criteria

In this proposed work, the classification results of the applied deep learning technique were evaluated by accuracy, precision, recall, and f1-score to detect the best SNPs. These important SNPs were used for contributing to the risk of AD through epistasis interactions and understanding of underlying biological mechanisms of the disease. Cross validation (CV) is accomplished to assess the ability of the model to classify and to predict a disease status. The used metric of model fit was BA averaged for all cross-validation experiments. It is a metric that can be chosen when evaluating how good a binary classifier is. BA is used when the classes are imbalanced, means one of the 2 classes are a lot more than the other class (Velez et al., 2007).

### 4.6. Implementations

The DL algorithms demand large computational operations while training. GPU works properly with the deep learning training. Hence, we used GPU in training deep learning model. Training Deep learning model on GPU accelerates the training process. Hence, running our model on GPU takes about 8 min 15 s to execute. In this proposed work, the used implementation tools are:

- PLINK version 1.07 (Purcell, 2012) is an open-source whole genome association analysis toolset, developed to conduct a range of basic, large-scale analyses in a computationally effective manner.
- R version 3.6.3 (Bischl et al., 2016)
- Python version 3.6.5 (Pedregosa et al., 2011) is an open-source programming language. In this proposed framework, deep learning model was applied.
- Multifactor Dimensionality Reduction (MDR) (Wu et al., 2011), version 3.0.2, is open-source software used to detect and characterize combinations of features.

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

| Gene name | Identified genes | Discovered genes |
|---|---|---|
| *LINC01830* | | Yes |
| *AGBL1* | | Yes |
| *ANXA9* | Yes () | |
| *IZUMO4* | | Yes |
| *SLC13A3* | | Yes |
| *ROBO2* | Yes () | |
| Near genes *NR3C2* and *LOC102724672* | Yes () | |
| *LOC101929507* | | Yes |
| *TENM4* | Yes () | |
| *DPP6* | Yes () | |
| *HCRTR2* | Yes () | |

## 5. Results and discussions

This paper suggests a novel MDR model using a deep learning technique for discovering epistasis interactions related to AD disease. The results revealed that this proposed framework could explore epistasis interactions more powerfully and enhance classification performance effectively. Table 2 shows the top 10 important pairwise models with their overall BA accuracy, BA model training, BA model testing, and *p*-values. The chosen metric of model fit was BA, which characterized the average sensitivity and specificity. The results showed that the training accuracy and the testing accuracy close to each other. This shows the lowering of over-fitting and the rising of generalizability (Moore and Andrews, 2015). These results suggested that these two-way interactions are associated with AD disease. The most significant two-way interaction was found among non-coding SNP rs10862418 and non-coding rs206028 with BA accuracy overall of 0.649, BA model training of 0.650, BA model testing of 0.633, and a significance level of *p*-value 2.16E−06.

Table 3 shows the top ten significant 3-way interaction models with their BA accuracy overall, BA model training, BA model testing, and *p*-values. The achieved results recommended that these three-way synergistic effects among the three SNPs are related to the disease. The most significant three-way interaction was found among (SNP rs7559638 from gene *LINC01830*, non-coding SNP rs10862418, and non-coding SNP rs5932588) with BA accuracy overall of 0.702, BA model training of 0.703, BA model testing 0.680, and a significance level of *p*-value 8.25E−07.

Table 4 presents the top 10 four-way interaction models with their BA accuracy overall, BA model training, BA model testing, and *p*-values. The achieved results showed that these four-way interactions are related to AD disease. The most robust four-way interaction was found among SNP rs7559638 from gene *LINC01830*, non-coding SNP rs10862418, SNP rs2075831 from gene *WWOX*, and non-coding SNP rs5932588. Their BA accuracy overall, BA model training, BA model testing, and *p*-value are 0.767, 0.768, 0.712, and 3.03E−06, respectively.

**Table 2**
The top ten pairwise interaction models (top 20 SNPs approach).

| Models | Overall BA | BA training | BA testing | *p*-Value |
|---|---|---|---|---|
| rs10862418, rs206028 (–, –) | 0.649 | 0.650 | 0.633 | 2.16E−06 |
| rs10862418, rs5932588 (–, –) | 0.645 | 0.647 | 0.620 | 9.97E−06 |
| rs7532008, rs10862418 (*ANXA9*, –) | 0.641 | 0.641 | 0.641 | 4.68E−05 |
| rs206028, rs5932588 (–, –) | 0.638 | 0.638 | 0.638 | 0.00026 |
| rs3910044, rs10862418 (near genes *NR3C2* and *LOC102724672*, –) | 0.637 | 0.638 | 0.621 | 2.05E−05 |
| rs10862418, rs2075831 (–, *WWOX*) | 0.637 | 0.637 | 0.637 | 8.11E−05 |
| rs10155005, rs10862418 (*ROBO2*, –) | 0.636 | 0.638 | 0.614 | 1.70E−05 |
| rs10862418, rs9967630 (–, *IZUMO4*) | 0.633 | 0.634 | 0.608 | 1.46E−05 |
| rs2798641, rs10862418 (*ARMC2*, –) | 0.631 | 0.631 | 0.627 | 0.00011 |
| rs9818965, rs10862418 (–, –) | 0.630 | 0.631 | 0.613 | 7.68E−05 |

**Table 3**
The top ten 3-way interaction models (top 20 SNPs approach).

| Models | Overall BA | BA training | BA testing | *p*-Value |
|---|---|---|---|---|
| rs7559638, rs10862418, rs5932588 (*LINC01830*, –, –) | 0.702 | 0.703 | 0.680 | 8.25E−07 |
| rs10245002, rs11237677, rs10862418 (*DPP6*, *TENM4*, –) | 0.700 | 0.700 | 0.688 | 1.85E–07 |
| rs10245002, rs10862418, rs206028 (*DPP6*, –, –) | 0.688 | 0.690 | 0.661 | 1.05E–06 |
| rs10862418, rs10852073, rs206028 (–, *AGBL1*, –) | 0.687 | 0.688 | 0.658 | 1.05E–05 |
| rs11237677, rs10862418, rs206028 (*TENM4*, –, –) | 0.687 | 0.688 | 0.653 | 3.18E–06 |
| rs9967630, rs6090536, rs5932588 (*IZUMO4*, *SLC13A3*, –) | 0.687 | 0.688 | 0.660 | 7.26E–05 |
| rs7559638, rs10862418, rs206028 (*LINC01830*, –, –) | 0.686 | 0.689 | 0.629 | 4.22E–08 |
| rs10155005, rs10862418, rs2075831 (*ROBO2*, –, *WWOX*) | 0.686 | 0.688 | 0.625 | 9.35E–05 |
| rs10155005, rs10862418, rs206028 (*ROBO2*, –, –) | 0.686 | 0.686 | 0.671 | 2.30E–05 |
| rs10155005, rs10862418, rs5932588 (*ROBO2*, –, –) | 0.684 | 0.687 | 0.643 | 5.37E–05 |

**Table 4**
The top ten 4-way interaction models (top 20 SNPs approach).

| Models | Overall BA | BA training | BA testing | *p*-Value |
|---|---|---|---|---|
| rs7559638, rs10862418, rs2075831, rs5932588 (*LINC01830*, –, *WWOX*, –) | 0.767 | 0.768 | 0.712 | 3.03E–06 |
| rs7559638, rs10862418, rs10852073, rs5932588 (*LINC01830*, –, *AGBL1*, –) | 0.758 | 0.764 | 0.661 | 4.46E–05 |
| rs10155005, rs10862418, rs206028, rs5932588 (*ROBO2*, –, –, –) | 0.753 | 0.756 | 0.669 | 7.01E–05 |
| rs7532008, rs7559638, rs10862418, rs5932588 (*ANXA9*, *LINC01830*, –, –) | 0.753 | 0.757 | 0.665 | 3.67E–05 |
| rs7559638, rs10862418, rs206028, rs5932588 (*LINC01830*, –, –, –) | 0.752 | 0.756 | 0.659 | 3.70E–07 |
| rs7559638, rs10862418, rs9967630, rs5932588 (*LINC01830*, –, *IZUMO4*, –) | 0.750 | 0.755 | 0.663 | 9.50E–06 |
| rs9818965, rs2224553, rs10862418, rs206028 (–, *LOC101929507*, –, –) | 0.750 | 0.755 | 0.641 | 4.93E–05 |
| rs10155005, rs11237677, rs10862418, rs206028 (*ROBO2*, *TENM4*, –, –) | 0.750 | 0.752 | 0.690 | 1.50E–05 |
| rs3910044, rs10862418, rs10852073, rs5932588 (near genes *NR3C2* and *LOC102724672*, –, *AGBL1*, –) | 0.748 | 0.752 | 0.648 | 0.00157 |
| rs9818965, rs3910044, rs10862418, rs5932588 (–, near genes *NR3C2* and *LOC102724672*, –, –) | 0.748 | 0.752 | 0.630 | 6.73E–05 |

**Table 5**
The top ten 5-way interaction models (top 20 SNPs approach).

| Models | Overall BA | BA training | BA testing | *p*-Value |
|---|---|---|---|---|
| rs7559638, rs10862418, rs10852073, rs206028, rs5932588 (*LINC01830*, –, *AGBL1*, –, –) | 0.840 | 0.845 | 0.635 | 3.82E–05 |
| rs10155005, rs10862418, rs2075831, rs206028, rs5932588 (*ROBO2*, –, *WWOX*, –, –) | 0.840 | 0.845 | 0.651 | 0.0009 |
| rs7532008, rs7559638, rs9818965, rs10852073, rs7052935 (*ANXA9*, *LINC01830*, –, *AGBL1*, –) | 0.838 | 0.844 | 0.611 | 0.0079 |
| rs9818965, rs10862418, rs10852073, rs7052935, rs206028 (–, –, *AGBL1*, –, –) | 0.836 | 0.840 | 0.631 | 0.0014 |
| rs7559638, rs9818965, rs10862418, rs10852073, rs206028 (*LINC01830*, –, –, *AGBL1*, –) | 0.834 | 0.840 | 0.591 | 0.0012 |
| rs7532008, rs7559638, rs10862418, rs2075831, rs5932588 (*ANXA9*, *LINC01830*, –, *WWOX*, –) | 0.833 | 0.838 | 0.658 | 0.0002 |
| rs7559638, rs3910044, rs2075831, rs6090536, rs5932588 (*LINC01830*, near genes *NR3C2* and *LOC102724672*, *WWOX*, *SLC13A3*, –) | 0.833 | 0.839 | 0.601 | 0.0110 |
| rs7532008, rs7559638, rs10862418, rs9967630, rs5932588 (*ANXA9*, *LINC01830*, –, *IZUMO4*, –) | 0.832 | 0.836 | 0.692 | 5.23E–05 |
| rs7559638, rs9818965, rs3910044, rs10852073, rs2075831 (LINC01830, –, near genes *NR3C2* and *LOC102724672*, *AGBL1*, *WWOX*) | 0.832 | 0.840 | 0.565 | 0.0047 |
| rs7559638, rs10862418, rs2075831, rs6090536, rs5932588 (*LINC01830*, –, *WWOX*, *SLC13A3*, –) | 0.832 | 0.840 | 0.633 | 0.0007 |

Table 5 presents the top 10 five-way interaction models with their BA accuracy overall, BA model training, BA model testing, and *p*-values. The achieved results showed that these five-way interactions are associated with AD disease. The most robust five-way interaction was found among SNP rs7559638 from gene *LINC01830*, non-coding SNP rs10862418, SNP rs10852073 from gene *AGBL1*, non-coding SNP rs206028, and non-coding rs5932588. Their BA accuracy overall, BA model training, BA model testing, and *p*-value are 0.840, 0.845, 0.635, and 3.82E–05, respectively.

Pathway analysis gives meaning to high-throughput biological data. SNPs are contextualized in biological processes through the gene(s) to which they were mapped. Hence, after applying SNPs rankings approaches, SNPs are mapped to genes within each pathway using the mapping in NCBI's dbSNP database (Sherry et al., 2001).

In this paper, 20 SNPs were ranked as the most robust SNPs using deep learning. First, the genes were mapped from these SNPs containing previously detected AD association genes and unknown but potential AD association genes. After that, epistasis interaction analysis was applied on the 20 SNPs using MDR to explore important pairwise, 3-way, 4-way, and 5-way interactions. The generated analysis of the 11 genes that are

mapped from the top-ranked SNPs also recommends robust epistasis interactions that may help interpret the risk of the disease.

After selecting the top 500 rankings in the last approach, these SNPs' statistical interaction analysis was generated to explore significant gene-gene interactions associated with AD. Table 6 shows the top 10 important pairwise models with their BA accuracy overall, BA model training, BA model testing, and *p*-values based on the last approach.

Table 7 shows the top ten significant 3-way interaction models using MDR based on top 500 ranking SNPs.

Table 8 shows the top ten significant 4-way interaction models using MDR based on top 500 ranking SNPs.

Table 9 shows the top ten significant 5-way interaction models based on top 500 ranking SNPs. The best accuracies were achieved using the top 500 SNPs rankings approach. The most robust five-way interaction was found among SNP rs11691402 from gene *LINC01317*, SNP rs10758578 from gene *GLIS3*, SNP rs17557796 near genes *ELAVL2* and *LOC105375992*, SNP rs2764808 near genes *CTNNA3* and *LRRTM3*, and non-coding rs1883105. Their BA accuracy overall, BA model training, BA model testing, and p-value are 0.874, 0.881, 0.589, and 0.00025, respectively. It was shown that the *ZEB2* gene is repeated with genes *ELAVL2* and *LOC105375992*, as observed in Tables 8 and 9.

The achieved results of the proposed framework outperformed the results reported in (Romero-Rosales et al., 2020; Sherif et al., 2017; Abd El Hamid et al., 2019; Chang et al., 2020). This research work was not limited to examining the association of each SNP independently with the phenotype as reported in (Romero-Rosales et al., 2020; Abd El Hamid et al., 2019) but also concentrated on the interaction between multiple SNP loci up to fifth-order interactions. In this paper, the same dataset (ADNI) used in (Sherif et al., 2017) was used with the proposed framework, and the achieved results outperformed the results reported in (Sherif et al., 2017).

Chang et al. (2020) focused only on pairwise epistasis interactions and did not explore higher-order interactions of genes in their analysis. Hence, the proposed work presents considerable improvement using an integration of DNN and MDR over previous methods.

Several genes mapped from the applied SNPs rankings approaches, including *GRID2*, *ELAVL2*, *ANXA9*, *ROBO2*, *NR3C2*, *TENM4*, *DPP6*, and *HCRTR2*, have been known previously related to AD disease.

After applying SNPs rankings approaches, the results suggested novel genes associated with AD disease like *LINC01830*, *AGBL1*, *IZUMO4*, *SLC13A3*, *LOC101929507*, *LOC105374292*, *NSUN7*, and *LINC01482*. It was not shown that these genes are explored before, and they can be possible risk association genes to AD disease. Furthermore, this paper suggests that these novel genes may be associated with this complex disease since they have strong interactions with other genes, as shown in the previous tables.

**Table 6**
The top ten pairwise interaction models (top 500 SNPs approach).

| Models | Overall BA | BA training | BA testing | *p*-Value |
|---|---|---|---|---|
| 1-rs17021105, rs9927963 (*GRID2*, –) | 0.670 | 0.671 | 0.667 | 1.83E–07 |
| 2-rs10862406, rs959144 (–, *PXMP4*) | 0.670 | 0.670 | 0.667 | 7.40E–08 |
| 3-rs17021105, rs246718 (*GRID2*, –) | 0.670 | 0.670 | 0.662 | 1.04E–07 |
| 4-rs17021105, rs9456815 (*GRID2*, *PACRG*) | 0.669 | 0.669 | 0.666 | 3.02E–08 |
| 5-rs17021105, rs13169441 (*GRID2*, *SAP30L-AS1*) | 0.667 | 0.668 | 0.661 | 1.10E–07 |
| 6-rs10862418, rs207036 (–, –) | 0.667 | 0.667 | 0.659 | 1.96E–07 |
| 7-rs1925616, rs11002688 (near genes *CTNNA3*, *LRRTM3*, and *LOC101928961*, –) | 0.666 | 0.666 | 0.659 | 5.62E–07 |
| 8-rs7557276, rs10862418 (*CLASP1*, –) | 0.665 | 0.665 | 0.658 | 3.10E–07 |
| 9-rs177138, rs17021105 (*FHIT*, *GRID2*) | 0.663 | 0.663 | 0.657 | 7.38E–08 |
| 10-rs17021105, rs11963648 (*GRID2*, *PACRG*) | 0.663 | 0.663 | 0.659 | 2.48E–07 |

**Table 7**
The top ten 3-way interaction models (top 500 SNPs approach).

| Models | Overall BA | BA training | BA testing | *p*-Value |
|---|---|---|---|---|
| 1-rs7557276, rs1796518, rs10862418 (*CLASP1*, *BTN2A2*, –) | 0.727 | 0.729 | 0.660 | 1.85E–07 |
| 2-rs6760326, rs17021105, rs17813753 (*LINC01873*, *GRID2*, –) | 0.724 | 0.724 | 0.703 | 1.15E–09 |
| 3-rs8056021, rs6025639, rs1883105 (–, –, –) | 0.721 | 0.723 | 0.670 | 1.48E–08 |
| 4-rs6760326, rs1410421, rs17813753 (*LINC01873*, –, –) | 0.720 | 0.721 | 0.685 | 9.06E–08 |
| 5-rs7557276, rs2529489, rs10862418 (*CLASP1*, *IMMP2L*, –) | 0.720 | 0.721 | 0.693 | 7.84E-10 |
| 6-rs242263, rs17557796, rs9409912 (*GJB7*, near genes *ELAVL2* and *LOC105375992*, *LOC100506532*) | 0.720 | 0.721 | 0.676 | 1.92E–08 |
| 7-rs6760326, rs17021105, rs10942387 (*LINC01873*, *GRID2*, –) | 0.718 | 0.719 | 0.710 | 4.38E–08 |
| 8-rs10862418, rs933561, rs7230479 (–, *ZNF423*, *LINC01478*) | 0.718 | 0.719 | 0.703 | 2.37E–07 |
| 9-rs17021105, rs8056021, rs2205637 (*GRID2*, –, –) | 0.718 | 0.719 | 0.706 | 1.72E–08 |
| 10-rs17557796, rs9301365, rs959144 (near genes *ELAVL2* and *LOC105375992*, –, *PXMP4*) | 0.718 | 0.721 | 0.669 | 1.34E–06 |

## 6. Conclusions

In this paper, the MDR constructive induction algorithm was integrated with the deep learning algorithm to explore epistasis interactions in a computationally efficient method. It was shown that deep learning algorithms have been worked successfully in various biomedical domain tasks using large amounts of data. This paper shows a proposed framework using deep learning to search for epistasis interactions associated with the disease. The result of the higher-order SNP interactions is presented in the previous tables. The results showed that the presented framework could assign a numeric measure of credit to each input attribute using SHAP feature importance and improve the classification performance. The deep learning technique was interpreted using SHAP to assign feature importance to each feature and get top 20, 100, 300, and 500 ranking SNPs. The primary goal is to discover relevant SNPs responsible for AD risk through epistasis interactions. Discovering SNPs for predicting disease risks is vital to contribute to personalized medicine.

The best remarkable interaction models associated with AD were detected in this work. In the top 20 rankings approach, the classification accuracies of five-way interaction models varied between 0.832 and 0.840. However, the classification accuracies of two-way, three-way, four-way models varied between 0.630 and 0.649, 0.684 and 0.702, and 0.747 and 0.758, respectively. In the top 500 rankings approach, the results have been reached to (0.860–0.874) in the five-way interaction model. While, the classification accuracies of 2-way, 3-way, 4-way models varied between 0.663 and 0.670, 0.718 and 0.727, and 0.793

**Table 8**
The top ten 4-way interaction models (top 500 SNPs approach).

| Models | Overall BA | BA training | BA testing | *p*-Value |
|---|---|---|---|---|
| rs878698, rs17557796, rs4537681, rs206028 (*TBC1D7-LOC100130357*, near genes *ELAVL2* and *LOC105375992*, –, –) | 0.803 | 0.807 | 0.701 | 4.37E–07 |
| rs9883073, rs242263, rs17557796, rs9409912 (–, *GJB7*, near genes *ELAVL2* and *LOC105375992*, *LOC100506532*) | 0.800 | 0.803 | 0.708 | 4.24E–08 |
| rs17557796, rs9301365, rs959144, rs206028 (near genes *ELAVL2* and *LOC105375992*, –, *PXMP4*) | 0.798 | 0.802 | 0.689 | 5.99E–07 |
| rs878698, rs1796518, rs10961303, rs17557796 (*TBC1D7-LOC100130357*, *BTN2A2*, near genes *LINC00583* and *LOC101929507*, near genes *ELAVL2* and *LOC105375992*) | 0.798 | 0.802 | 0.650 | 5.10E–06 |
| rs878698, rs1796518, rs17557796, rs4537681 (*TBC1D7-LOC100130357*, *BTN2A2*, near genes *ELAVL2* and *LOC105375992*, –) | 0.796 | 0.800 | 0.677 | 2.14E–06 |
| rs11934708, rs17557796, rs1925616, rs1548906 (*CXXC4-AS1*, near genes *ELAVL2* and *LOC105375992*, near genes *CTNNA3*, *LRRTM3* and *LOC101928961*, *ADAMTS14*) | 0.796 | 0.800 | 0.653 | 5.64E–07 |
| rs10758578, rs17557796, rs207036, rs2205637 (*GLIS3*, near genes *ELAVL2* and *LOC105375992*, –, –) | 0.795 | 0.802 | 0.672 | 9.98E–07 |
| rs17557796, rs9301365, rs959144, rs207036 (near genes *ELAVL2* and *LOC105375992*, –, *PXMP4*) | 0.795 | 0.799 | 0.673 | 3.08E–06 |
| rs7604762, rs7597006, rs17557796, rs930016 (–, *ZEB2*, near genes *ELAVL2* and *LOC105375992*, *LOC105370805*) | 0.794 | 0.798 | 0.655 | 1.43E–07 |
| rs11934708, rs17557796, rs2147886, rs1548906 (*CXXC4-AS1*, near genes *ELAVL2* and *LOC105375992*, near genes *CTNNA3* and *LRRTM3*, *ADAMTS14*) | 0.793 | 0.799 | 0.644 | 9.85E–07 |

**Table 9**
The top ten 5-way interaction models using MDR (top 500 SNPs approach).

| Models | Overall BA | BA training | BA testing | *p*-Value |
|---|---|---|---|---|
| rs11691402, rs10758578, rs17557796, rs2764808, rs1883105 (*LINC01317*, *GLIS3*, near genes *ELAVL2* and *LOC105375992*, near genes *CTNNA3* and *LRRTM3*, –) | 0.874 | 0.881 | 0.589 | 0.00025 |
| rs6711065, rs6452399, rs2596501, rs1755779, rs1880769 (*GMCL1*, *ACOT12*, *HLA-B*, near genes *ELAVL2* and *LOC105375992*, *CA10*) | 0.866 | 0.872 | 0.640 | 0.00020 |
| rs6798458, rs2309949, rs10053765, rs179651, rs17557796 (*OSBPL10*, *STOX2*, –, *BTN2A2*, near genes *ELAVL2* and *LOC105375992*) | 0.866 | 0.871 | 0.653 | 0.00088 |
| rs2529489, rs1755779, rs10862418, rs959144, rs207036 (*IMMP2L*, near genes *ELAVL2* and *LOC105375992*, –, *PXMP4*, –) | 0.865 | 0.871 | 0.621 | 5.63E–05 |
| rs1050316, rs2309949, rs246718, rs1796518, rs17557796 (*MEF2D*, *STOX2*, –, *BTN2A2*, near genes *ELAVL2* and *LOC105375992*) | 0.865 | 0.870 | 0.626 | 0.00373 |
| rs7597006, rs6851877, rs17557796, rs276480, rs5920524 (*ZEB2*, *ANK2*, near genes *ELAVL2* and *LOC105375992*, near genes *CTNNA3* and *LRRTM3*, –) | 0.864 | 0.871 | 0.614 | 0.00033 |
| rs2596501, rs6466401, rs17557796, rs592637, rs892596 (*HLA-B*, *DOCK4*, near genes *ELAVL2* and *LOC105375992*, near genes *MS4A10* and *LOC105369322*, *ZNF180*) | 0.863 | 0.869 | 0.627 | 6.76E–05 |
| rs11691402, rs759700, rs17557796, rs2764808, rs1880769 (*LINC01317*, *ZEB2*, near genes *ELAVL2* and *LOC105375992*, near genes *CTNNA3* and *LRRTM3*, *CA10*) | 0.860 | 0.868 | 0.530 | 0.00086 |
| rs7604762, rs7597006, rs9812746, rs1755779, rs9301365 (–, *ZEB2*, *CNTN4*, near genes *ELAVL2* and *LOC105375992*, –) | 0.860 | 0.866 | 0.613 | 0.00010 |
| rs826009, rs9818965, rs1796518, rs4935126, rs4238639 (–, –, *BTN2A2*, *PCDH15*, –) | 0.860 | 0.865 | 0.647 | 0.00017 |

and 0.803, respectively.

In this proposed work, an efficient framework was proposed to identify epistatic interactions between all pairs of nucleotides in DNA sequence input. The primary target of this paper is to integrate constructive induction algorithm MDR with deep learning techniques to make the proposed framework more robust to build models for other complex diseases.

In the top 20 rankings approach, there are 11 genes mapped from the 20 SNPs containing previously detected genes and novel genes associated with AD. Both types of genes (previously identified and newly discovered) may help investigate the risk of AD disease. There are several genes mapped from the applied SNPs rankings approaches (top 20, 100, 300, 500), including previously detected AD association genes and unknown but potential AD association genes. This paper proposes *LINC01830*, *AGBL1*, *IZUMO4*, *SLC13A3*, and *LOC101929507* genes are potentially associated with AD since they have strong interactions with other genes as appeared in the preceding tables. One of the most repeated interactions is between *LINC01830/AGBL1* in models 1, 2, 4, 6, and 10, as observed in Table 5 (5-way). It was shown that the *LINC01830* gene is repeated with *ANXA9* (*LINC01830/ANXA9*) in models 2, 5, and 7, as shown in Table 5. Also, one of the most repeated interactions is between *GRID2* and *PACRG* in models 4 and 10, as shown in Table 6. It was shown that the *GRID2* gene is repeated with *LINC01873* in models 2 and 7, as observed in Table 7.

## CRediT authorship contribution statement

**Marwa M. Abd El Hamid:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft.

**Yasser M.K. Omar:** Supervision, Conceptualization, Data curation, Methodology, Resources, Validation, Writing – review & editing.

**Mohamed Shaheen:** Supervision, Conceptualization, Data curation, Methodology, Resources, Validation, Writing – review & editing.

**Mai S. Mabrouk:** Conceptualization, Data curation, Methodology, Resources, Validation, Writing – review & editing.

## Declaration of competing interest

All 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.

## Data availability

The dataset was obtained from ADNI database

## Acknowledgements

The ADNI database was released in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. In this paper, the used dataset was obtained from the ADNI database (). The investigators within the ADNI study participated in the design and implementation of ADNI and/or given data. However, they did not contribute to the analysis/writing of this paper.

## References

Abd El Hamid, M.M., Mabrouk, M.S., Omar, Y.M., 2019. Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer's disease using machine learning techniques. Biomed. Eng. 31, 1950040.
Abd El Hamid, M.M., Shaheen, M., Mabrouk, M.S., Omar, Y.M., 2021. Machine learning for detecting epistasis interactions and its relevance to personalized medicine in Alzheimer's disease: systematic review. Biomed. Eng. 33, 1016–2372, 2150047.
Bailey, P., 2007. Biological markers in Alzheimer's disease. Can. J. Neurol. Sci. 34, S72–S76.
Berrar, D., Dubitzky, W., 2021. Deep Learning in Bioinformatics and Biomedicine. Oxford University Press.
Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., et al., 2016. Mlr: machine learning in R. J. Mach. Learn. Res. 17, 5938–5942.
Breijyeh, Z., Karaman, R., 2020. Comprehensive review on Alzheimer's disease: causes and treatment. Molecules 25, 5789.
Carrillo, M.C., Bain, L.J., Frisoni, G.B., Weiner, M.W., 2012. Worldwide Alzheimer's disease neuroimaging initiative. Alzheimers Dement. 8, 337–342.
Chang, Y.-C., Wu, J.-T., Hong, M.-Y., Tung, Y.-A., Hsieh, P.-H., Yee, S.W., et al., 2020. GenEpi: gene-based epistasis discovery using machine learning. BMC Bioinformatics 21, 1–13.
Cordell, H.J., 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet. 11, 2463–2468.
Cui, T., Mekkaoui, K.El, Reinvall, J., Havulinna, A.S., Marttinen, P., Kaski, S., 2021. Gene-Gene Interaction Detection with Deep Learning. bioRxiv.
Dickinson, Q., Meyer, J.G., 2021. Positional SHAP for Interpretation of Deep Learning Models Trained from Biological Sequences. bioRxiv.
Dorani, F., Hu, T., Woods, M.O., Zhai, G., 2018. Ensemble learning for detecting gene-gene interactions in colorectal cancer. PeerJ 6, e5854.
Dunn, A.R., O'Connell, K.M., Kaczorowski, C.C., 2019. Gene-by-environment interactions in Alzheimer's disease and Parkinson's disease. Neurosci. Biobehav. Rev. 103, 73–80.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT press.
Ho, D.S.W., Schierding, W., Wake, M., Saffery, R., O'Sullivan, J., 2019. Machine learning SNP based prediction for precision medicine. Front. Genet. 10, 267.
Kim, J., Sohn, I., Kim, D.D.H., Jung, S.-H., 2013. SNP selection in genome-wide association studies via penalized support vector machine with MAX test. Comput. Math. Methods Med. vol. 2013.
Lehne, B., Lewis, C.M., Schlitt, T., 2011. From SNPs to genes: disease association at the gene level. PloS one 6, e20133.
Meyer, J.C., Harirari, P., Schellack, N., 2016. Overview of Alzheimer's disease and its management. SA Pharm. J. 83, 48–56.
Min, S., Lee, B., Yoon, S., 2017. Deep learning in bioinformatics. Brief. Bioinform. 18, 851–869.
Molnar, C., 2020. Interpretable Machine Learning: Lulu. com.
Moore, J.H., Andrews, P.C., 2015. Epistasis analysis using multifactor dimensionality reduction. In: Epistasis. Springer, pp. 301–314.
Moore, J.H., Williams, S.M., 2009. Epistasis and its implications for personal genetics. Am. J. Hum. Genet. 85, 309–320.
Niel, C., Sinoquet, C., Dina, C., Rocheleau, G., 2015. A survey about methods dedicated to epistasis detection. Front. Genet. 6, 285.
Okazaki, A., Horpaopan, S., Zhang, Q., Randesi, M., Ott, J., 2021. Genotype pattern Mining for Pairs of interacting variants underlying digenic traits. Genes 12, 1160.
Patron, J., Serra-Cayuela, A., Han, B., Li, C., Wishart, D.S., 2019. Assessing the performance of genome-wide association studies for predicting disease risk. PLoS One 14, e0220215.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Purcell, S., 2012. PLINK (1.07). Documentation.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.
Romero-Rosales, B.-L., Tamez-Pena, J.-G., Nicolini, H., Moreno-Treviño, M.-G., Trevino, V., 2020. Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling. PloS one 15, e0232103.
Schmalohr, C.L., Grossbach, J., Clément-Ziza, M., Beyer, A., 2018. In: Detection of Epistatic Interactions With Random Forest. bioRxiv, p. 353193.
Schwing, A.G., Urtasun, R., 2015. Fully Connected Deep Structured Networks. arXiv preprint arXiv:1503.02351.
Sherif, F.F., Zayed, N., Fakhr, M., Wahed, M.A., Kadah, Y.M., 2017. Integrated higher-order evidence-based framework for prediction of higher-order epistasis interactions in Alzheimer's disease. Int. J. Biol. Biomed. Eng. 11, 16.
Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., et al., 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311.
Simons, Y.B., Bullaughey, K., Hudson, R.R., Sella, G., 2018. A population genetic interpretation of GWAS findings for human quantitative traits. PLoS Biol. 16, e2002985.
Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., et al., 2007. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet. Epidemiol. 31, 306–315.
Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., et al., 2019. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nat. Commun. 10, 1–14.
Wang, Y.-T., Sung, P.-Y., Lin, P.-L., Yu, Y.-W., Chung, R.-H., 2015. A multi-SNP association test for complex diseases incorporating an optimal P-value threshold algorithm in nuclear families. BMC Genomics 16, 1–10.
Wu, Y., Zhang, L., Liu, L., Zhang, Y., Zhao, Z., Liu, X., et al., 2011. A multifactor dimensionality reduction-logistic regression model of gene polymorphisms and an environmental interaction analysis in cancer research. Asian Pac. J. Cancer Prev. 12, 2887–2892.
Xie, F., Chan, J.C., Ma, R.C., 2018. Precision medicine in diabetes prevention, classification and management. J. Diabetes Investig. 9, 998–1015.

## Further Reading

Bertram, L., Tanzi, R., 2019. Alzheimer disease risk genes: 29 and counting. Nat. Rev. Neurosci. 15, 191–192.
Biffi, A., Anderson, C.D., Desikan, R.S., et al., 2010. Genetic variation and neuroimaging measures in Alzheimer disease. Arch. Neurol. 67, 677–685.
Mostafa, M., MK, Y., Mabrouk, M., 2016. Identifying genetic biomarkers associated to Alzheimer's disease using Support Vector Machine. In: 2016 8th Cairo International Biomedical Engineering Conference. IEEE, pp. 5–9.
Seshadri, S., Fitzpatrick, A.L., Ikram, M.A., et al., 2010. Genome-wide analysis of genetic loci associated with Alzheimer disease. J. Am. Med. Assoc. 303, 1832–1840.