A Novel Hybrid Machine Learning Approach Using Deep Learning for the Prediction of Alzheimer Disease Using Genome Data

A. Alatrany[1,2(✉)], A. Hussain[1], J. Mustafina[3], and D. Al-Jumeily[1]

[1] School of Computer Science and Mathematics,
Liverpool John Moores University, Liverpool, UK
A.S.Alatrany@2020.ljmu.ac.uk, {A.Hussain,D.Aljumeily}@ljmu.ac.uk
[2] University of Information Technology and Communications, Baghdad, Iraq
[3] Kazan Federal University, Kazan, Russia
DNMustafina@kpfu.ru

Abstract. Genome-wide association studies are aimed at identifying associations between commonly occurring variations in a group of individuals and a phonotype, in which the Deoxyribonucleic acid is genotyped in the form of single nucleotide polymorphisms. Despite the exsistence of various research studies for the prediction of chronic diseases using human genome data, more investigations are still required. Machine learning algorithms are widely used for prediction and genome-wide association studies. In this research, Random Forest was utilised for selecting most significant single nucleotide polymorphisms associated to Alzheimer's Disease. Deep learning model for the prediction of the disease was then developed. Our extesnive similation results indicated that this hybrid model is promising in predicting individuals that suffer from Alzheimer's disease, achieving area under the curve of 0.9 and 0.93 using Convolutional Neural Network and Multilayer perceptron respectively.

Keywords: GWAS · Machine learning · Random forest · CNN · ANN

1 Introduction

Alzheimer disease (AD) is a neurodegenerative disease and a leading cause of dementia. According to World Alzheimer Report (2018) [1] around 50 million patients are diagnosed with dementia. People who suffer from AD account between 60% and 80% of dementia patients. Typically, Alzheimer's symptoms mature after the age of sixty, affecting both the mental and physical condition of the patient. A person diagnosed with Alzheimer's could suffer from various syndromes including memory efficiency decreases, speaking difficulties, lack of attention, decline in the quality of lifestyle. More critically, the disease could develop to cause serious damage, and this could lead patients to start forgetting their family and friends [2].

Based on the age of the onset, there are two subtypes to the disease, they are: Eary-onset AD (EOAD) and late-onset AD (LOAD). Approximately 5% of AD cases [3]

D.-S. Huang et al. (Eds.): ICIC 2021, LNAI 12838, pp. 253–266, 2021.
https://doi.org/10.1007/978-3-030-84532-2_23

show EOAD, the age-onset ranges from the 30's to early 60's. EOAD gene association includes 3 types, amyloid precursor, Presenilin 1 and Presenilin 2 [4]. In comparison with EOAD, LOAD is shown to occur later in life, showing from late 60's onwards. The incidence of LOAD has a rate of 90%–95% [3]. Apolipoprotein is a risk factor gene, associated with LOAD, which has been confirmed as the most common [5]. Genetic and environmental factors are both part of the LOAD which seems showing more complex disorder. Several genetic variants could influence the complex disease AD. Due to the difficulty of accessing pathological information, generally it is a challenge to predict neurodegenerative disorders [6].

The manner of development of AD continues to be difficult to grasp therefore, the course of AD also remains irreversible. Currently, there are no available medication to fight and cure AD, hence the progression of this disease cannot be reversed. Thus, by achieving an early diagnosis of AD, it can provide the patient with medication to slow the disease.

Machine learning algorithms have shown promising results when used for the prediction of diseases using genome wide dataset. Kim and his team [7] used The Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS) datasets to build a deep learning classifier for Type 2 Diabetes (T2D) from genome wide data and clinical factors. The dataset was randomly split into 75% and 25% for training and testing, respectively. Deep neural network was trained as a classifier for T2D using various number of genetic variants and tested on males and females. To evaluate the performance of the classifier, the Area Under the Curve (AUC) is used as a performance measurement. Their simulation results indicated that using high number of Single Nucleotide Polymorphisms (SNPs) during training can improve the performance of their proposed network. Their model outperformed logistic regression. Furthermore, the model accuracy improved when clinical data such as age, weight and hypertension were utilized. In another study [8], the authors tried to build a Deep Learning (DL) classifier to predict T2D from genotype - phenotype data as in the previous study. They used logistic regression to detect the most influencing SNPs associated with the disease, then constructed a multi-layer feedforward Artificial Neural Network (ANN) for the classification task. The authors claim that their proposed model achieved AUC of 0.52, when trained on SNPs with p-value of 5×10^{-8}, and achieved an AUC of 0.96 using larger set of SNPs by setting the p-value to 5×10^{-2}, their results indicated a performance outperformed Random Forest (RF) network.

Ghanem and colleagues [9] used supercomputer to apply parallel deep learning based on Map/Reduce framework to process data by multi-nodes to find SNP-SNP interactions. The proposed method consists of 2 stages: first stage, is the pre-processing of data, and filter SNPs pairs. In the second stage, a super-computer is used to apply the Deep Learning model. The computer clusters is controlled by H2O program accessed by R interface allowing the user to connect and organize the H2O server. Each node is responsible for training a subset of the DL network. The authors used GAMETES tool [10] to generate 12 simulated datasets by presenting different values of minor allele frequency, each dataset contains 1600 samples: 800 cases and 800 controls with 100 SNPs and 2 functional SNPs to simulate the pair-wise interaction. The model showed high accuracy between (83–98) for each of the simulated datasets.

Sun et al. [11] developed a multilayer deep neural network as a prediction tool of age-related macular degeneration (AMD) which is an eye disease using genetic data. Various simulation experiments were conducted to benchmark the performance of their proposed method with other machine learning algorithms. GWAS data of AMD from the age-related eye disease studies (AREDS and AREDSs) contains 7800 samples used for development and evaluation of the model. Deep Neural Network model achieved the best results train on 666 variables of the genetic variant reaching AUC of 81.8% which outperformed all other models.

Ghafouri-Fard et al. [12] constructed an ANN to predict individuals with Autism Spectrum Disorder (ASD) from genetics variants of 487 cases and 455 control of Iranian population. The authors selected only 15 SNPs associated with ASD reported in the literature as a feature set to the model. The model reached an AUC, accuracy, sensitivity, and specificity of 80.59%, 73.67%, 82.75% and 63.95%, respectively.

Different to previous researches, in this work we are proposing a hybrid machine learning model based on random forest and deep learning for the early detection of Alzheimer disease using genetic data.

Genome Wide Association Studies (GWAS) are associated with detecting associations between commonly occurring variations in a group of individuals and a phonotype [13]. Normally, GWAS requires a large number of participants, in which the Deoxyribonucleic Acid (DNA) is genotyped in the form of single nucleotide polymorphisms. Statistical tests are usually conducted to evaluate how each SNP asso-ciated with a human trait being studied [14]. There are two existing approaches of GWAS: family-based studies and population-based studies. Most GWA studies employ a case-control form [15].

GWAS shows a unique way of analysing genetic material, which involve some advantages and disadvantages. Previously, GWAS has been criticised due to the high expense cost. This limitation has been reduced significantly due to technological advancements. GWAS also shows some disadvantages which could prove to be critical in discovering new and existing biomarkers. These include the overlooking of some rare alleles along with an elevated rate of false discoveries [16].

Therefore, the most significant feature of GWA studies are the requirements of using a sample size larger enough to achieve the desired reliability of the results.

Machine Learning (ML) algorithms are computational methods developed to extract information from a raw data and turn it into useful resources [17]. Various studies have looked at individuals who are diagnosed with a specific disease and healthy people, by genotyping their DNA sequence, with the use of machine learnings to predict the susceptibility of individuals to the disease through their sequencing data [8, 9, 18, 19]. Machine learning algorithms, can be trained using supervised and unsupervised learning. In the former, the algorithm is trained on labelled data i.e. each observation includes its output class, a good model should make an accurate prediction when presented with new example. Whereas in the latter, the task is to identify and understand structures of data without prior knowledge of the output [20]. A supervised learning model would be used in the genetics context to aid in the identification of individuals who suffer from a disease vs healthy individual based on their sequence data [8, 12]. Machine learning can be utilised when the output of individuals diagnosis is unknown, therefore,

an unsupervised learning model can be used to from separate groups of these individuals, based on similar characteristics of their blood sequence [21].

Supervised learning algorithms have been extensively used by researchers in the genetics fields. Vivian-Griffiths et al. [22] used support vector machines to build predictive model of schizophrenia from genomic data. Laksshman and his colleagues [23] used decision trees, random forest and neural network to point out genomic mutations for bipolar disorder using exomes data. Yang et al. [24] applied a liner regression model on genetics data after passing quality control procedures in the aim to explain missing heritability of human height. The authors show that considering all 294,831 SNPs can explain 45% of the heritability. Whereas in [25, 26], the authors illustrated the applicability of CNN to DNA sequence data.

For unsupervised learning, Scholz and his team [27] applied non-liner principal component analysis to impute data. While in [28] used class purity maximization clustering to solve the problem of genetic data imbalance.

The reminder of this paper is organised as follows. Section 2 will provide details about our proposed methodology. Results will be presented in Sect. 3 alongside the discussions. Finally, conclusion and future directions are summarized in Sect. 4.

2 Methodology

This section provides the details about the analysis of genetic data and the proposed methodology for the early prediction of Alzheimer Disease.

Our methodology consists from various phases as will be explained in the following subsections.

2.1 Dataset

The GWAS case-control dataset obtained from [29]. The inclusion criteria for participates are those who reported themselves to be from European ethnicity, board-certified neuropathologists confirmed late-onset Alzheimer disease (LOAD) for cases or no neuropathology present for controls according to the National Alzheimer's Coordinating Centre protocols. Death age of participants greater than 65 years. Plaque and tangle assessment (unique structures that effect cells in the brain which could contribute to the pathophysiology of the disease) conducted on all cases and controls. Samples that suffered previously from stroke, Lewy bodies, or any other neurological disease were excluded. The dataset consists of 191 males and 173 females divided into 176 cases and 188 controls with genotype information across 502,627 SNPs. The DNA of participants were genotyped via Affymetrix GeneChip Human Mapping 500K Array Set. Detailed information regarding the dataset can be found in [29].

2.2 Quality Control

Plink [30] is an open-source software utilized to efficiently conduct basic analyses for genotype/phenotype data. Plink v1.9 is used in this study to perform all quality control procedures and pre-liminary analysis. Quality control procedures conducted as follows

on genetic markers filtering. SNPs were removed due to genotype missing rate larger than 5%, Hardly-Weinberg test with p-value less than 0.001, minor allele frequency less than 0.05. In addition, quality control procedures performed on each sample, including missing genotype data rate of 0.05, related individuals, and sex-homozygosity. After completing all the aforementioned steps, a total of 356499 SNPs in the samples were retained for the subsequent analysis.

2.3 Association Test - Logistic Regression

GWAS utilizes highly dimensional data thus making it extremely difficult to process the data directly in which most of the SNPs are irrelevant and uninformative. Therefore, selecting the most important SNPs is essential.

Logistic regression is one of the simplest machine learning algorithms and is the first algorithm to be used for classification problems. Logistic regression explains and assess the relations between the SNPs and the dependent variable.

The association test model is based on a logistic function that is described as:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

Association test used in the current work to rank each SNP with a significant value. Due to computational need, only the first 5000 SNPs with lowest p-value are extracted, to meet the machine learning classifier requirements.

2.4 Feature Selection

Random forest [31] is constructed by combining multiple decision trees to reach a more powerful model than using single model. Random forest algorithm is able to process noisy data. Additionally, random forest can efficiently deal with unbalance data set.

Feature selection techniques is required since many bioinformatics datasets have a high dimensional nature. There are two aims within feature selection, firstly is to gain and understand a more in-depth reasoning and insight to the process of selecting the given data and to enhance the model performance along with trying to avoid any/all overfitting. Machine learning interpretation into how a decision was made through its specific process, is as important as the accuracy of the output results of the model. Specially, in safety critical and medical applications, where some decisions could lead to fatal outcome. Feature selection can be performed using random forest along with the construction of classification rules.

Within the current study, the Gini measure which is one of the most popular and used methods to measure feature importance within RF is utilized as a feature selector based on the traning dataset. The attribute overall importance is calculated by averaging the importance value of all the trees within random forest. Due to the nature of the GWAS data, many SNPs are irrelevant and have an extremely low importance value. Therefore, all SNPs reached a Gini value of 0.0009 or above are selected within the feature set for the classification. The importance threshold of 0.0009 was selected by trial as it can capture the appropriate SNPs which reflect positive results in the classification task.

2.5 Classification

GWA studies use linear models to associate SNPs to a phenotype [32], liner models showed outstanding performance in their ability to make accurate decisions of various problems in different applications. However, GWASs consist of big, complex, and highly interacted data [33], all together making it a flavorful recipe to deep learning models. In recent years DL has been used by scientists as a knowledge discovery tool. Major work presented an understanding to the genetic background of a disease. Previous research [9, 12, 34, 35] showed that by taking advantage of the high computation power of DL, dense relationships from genetic data can be detected. However, the application of DL techniques to GWAS data still requires further investigation.

In this study, we have utilised both multilayer percepton (MLP) and Convolution Neurla network (CNN) for our experriments.

Multilayer perceptron is a feedforward Neural Networks, in which the connections of layers are in one direction. The input layer passes the input signal to the next layer and the process continued until it reaches the output layer, in which an output is produced.

Figure 1A shows basic structure of MLP network consisting of two hidden layers and a single output node. There is no limit or constraints on the number of inputs, outputs, layers, or nodes per layers [36]. The output of such neural networks depends totally on the current input therefore nodes are memoryless. The output of a single layer network is computed by:

$$y = \sigma(wx + b) \tag{2}$$

Where w is the weight of the network, x is the inputs to the network, b is the bias value, and σ represents the activation function. MLP netwrok used in our work consists of 60 inputs, 4 hidden layers with 32,12,6,3 nodes respectivly with using rectified liner activation function at the hidden layers and sigmoid function at the output layer. Learning rate of 0.001 used during the training phase with binary cross entropy loss function.

Convolutional Neural Network reduces the size of the connections from the input layer to the hidden layer by applying a filter over the input matrix. Therefore, neurons in the hidden layer are connected to local regions of the input. For efficient results, it is preferable to add several hidden layers. The layers are not limited to specific filter, but rather each layer can be associated with a different filter. Consequently, different patterns can be extracted from the input data [37].

To reduce the input complexity for the next layer, pooling layer is usually added after the conventional layer. Max-pooling is used in the current work. It divides the input into sub-regions and return the max value of each region. The output of pooling layer then flatten to be fully connected with the next layer.

Figure 1B shows a typical CNN architecture with convolution layers, pooling layers and one fully connected layer. Details of the architecture of CNN used in this paper are presented in Table 1. Learning rate of 0.001 used during the training phase with categorical cross entropy loss function.

The classifier is a constructed model to solve a specific problem. The model requires an input data consist of attributes of each observation, usually known as a feature set. Besides the given attributes, a set of hyper-parameters are also needed to train the model.

To obtain a high accuracy classifier, usually extensive hyper-parameter is required. In the current work a grid search is conducted on user defined hyper-parameters values.

Table 1. CNN architecture

Layer type	Description	Output shape
Input		(60,1)
1D Convolution	16 kernel (1X5) relu	(56,16)
1D Convolution	32 Kernel (1X5) tanh	(54,32)
1D Pooling	Max Pooling (1X2)	(27,32)
Dropout	Dropout_rate = 0.1	(27,32)
1D Convolution	64 Kernel (1 × 4) tanh	(25,64)
1D Pooling	Max Pooling (1 × 2)	(12,64)
Dropout	Dropout_rate = 0.1	(12,64)
Reshape	64 nodes, sigmoid	768
Dese	Flatten	64
Dropout	Dropout_rate = 0.1	64
Output	2 nodes, softmax	2

2.6 The Proposed Model

The proposed system presented in the current work aims to diagnose and early classify individuals with Alzheimer's disease as shown in Fig. 1C. The main intentions of the model are (1) reduce the number of features of the GWAS SNPs to allow fast and computational efficiently processing of data and (2) improve the classification accuracy of Alzheimer's' vs healthy individuals. The hybrid system consists of two stages specifically a feature selection stage and classification stage. Random forest was conducted as feature selector to find the ideal set of SNPs. For the second stage, ANN and CNN models are used as classifier. Algorithm 1 shows the proposed process.

Algorithm 1: Alzahimer Disease Classification using SNPs data
Let A to be the set of Alzehimer GWAS SNPs data
Let B to be the set of pre-prossed data set where B ∈ A
Let C to be the feature seletced set
$\forall\, b \in B, b \in C$ *if b is selected by LR and RF*
Let ML to be a set of ML
ML = { ANN, CNN}
A_z = ML (C) where A_z = { 0, 1}

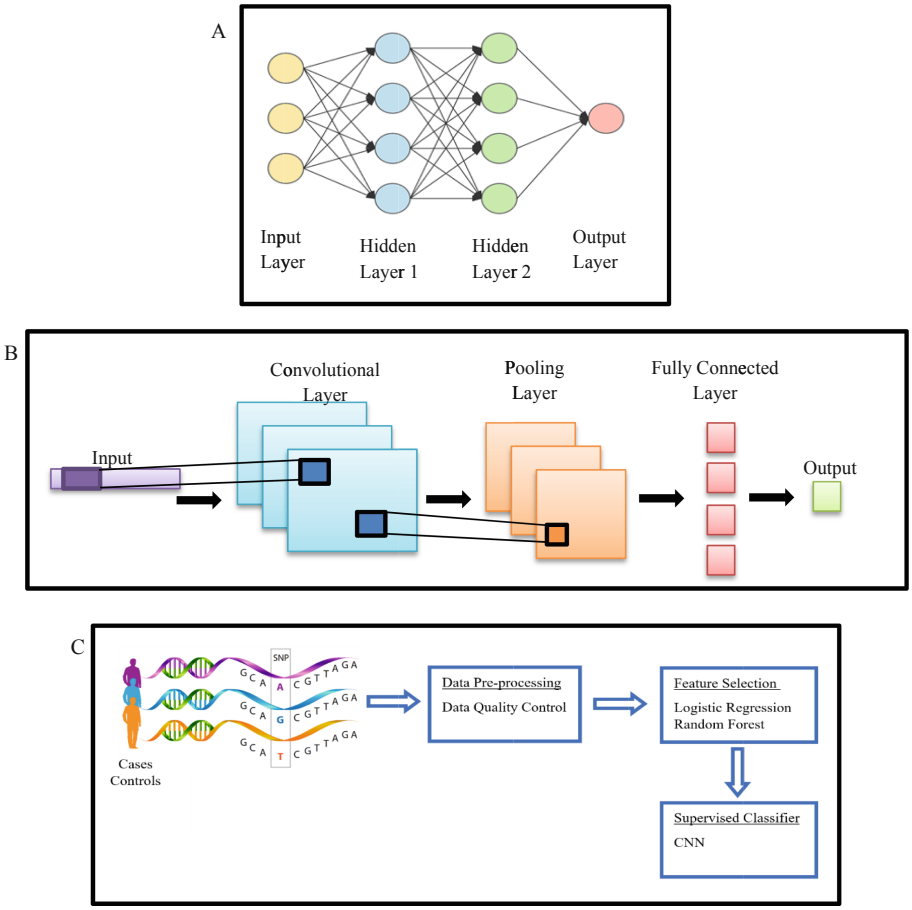

Fig. 1. Proposed model (A) Typical ANN architecture, (B) Typical CNN architecture, (C) Proposed model stages

3 Results and Discussion

This section presents the results of the stages explained in the methods section. After applying logistic regression for the association test, the top 5000 SNPs with the lowest p-value were selected for further analysis (Fig. 2).

Producing the most informative SNPs for the classification task is a challenging problem, in which random forest algorithm was used for SNPs selection in this work. The reason is because the tree-based strategies used by random forests naturally ranks attributes by how well they improve the purity of the node.

Table 2 shows the SNPs associated to AD according to RF Gini metric. The first two SNPs (rs429358 and rs4420638) are proven to be associated with the disease as illustrated in the Manhattan plot which reached the genome-wide significance level ($P < 1.2 \times 10^{-8}$). This suggests that the model is accurate in selecting good features for the

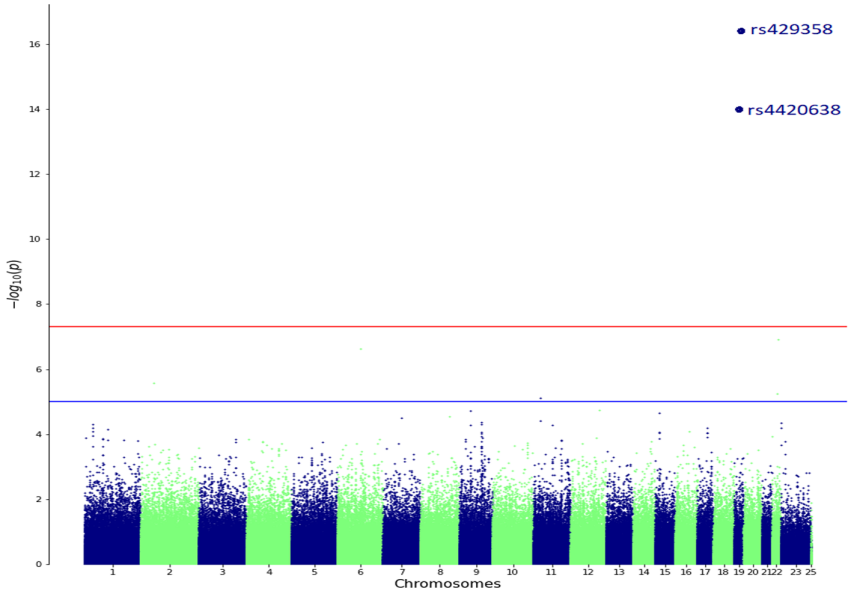

Fig. 2. Manhattan plot of associated SNPs with AD

classification task. However, more investigation is required to examine the associations of the rest of SNPs.

Table 2. Top 10 SNPs

SNP
rs429358
rs4420638
rs867198
rs3011823
rs5939190
rs8030415
rs1522940
rs2383559
rs7677027
rs862245

SNPs selected during the previous stage, are used as a feature set for two deep machine learning models (ANN and CNN) for the classification task of AD and Non-AD. The dataset is split into 80% and 20%, training and testing sets, respectively.

Table 3 shows the performance of the two models on testing set (Fig. 3).

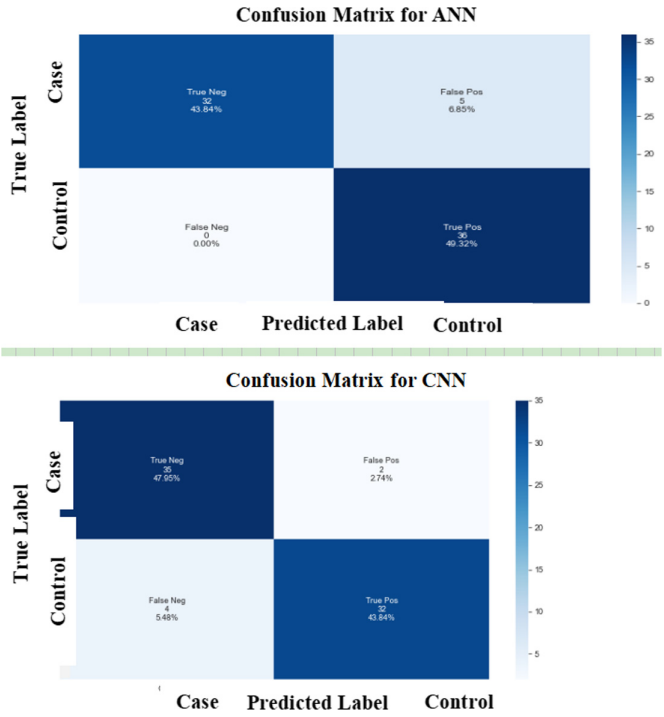

Fig. 3. Confusion matrix for ANN and CN

Table 3. Performance metrics for ANN and CNN testing set

Model	Accuracy	Precision	Recall	F1 score	Cohens kappa
ANN	0.93	0.88	1.00	0.94	0.86
CNN	0.92	0.94	0.89	0.91	0.84

Figure 4 presents the area under the carve for ANN and CNN, the performance of both models is approximately similar. This demonstrates the capability of these models to deal with complex genetic data.

The genetics of complex phenotype such as Alzheimer disease is challenging [38]. Multiple genetic markers contribute into the development of complex human disease. Although GWA studies were successful in revealing SNPs associated to complex traits, this method lacks the identification markers that together have influence on the disease. In our study, we aimed to explore DL models in classification of AD based solely on GWAS. DL offers a well renowned analytical models used in classification applications of diseases.

Few studies reported the use of Deep Learning in diagnostics of complex diseases using SNPs as fea-ture set. Some studies used SNPs that are proven to be associated with

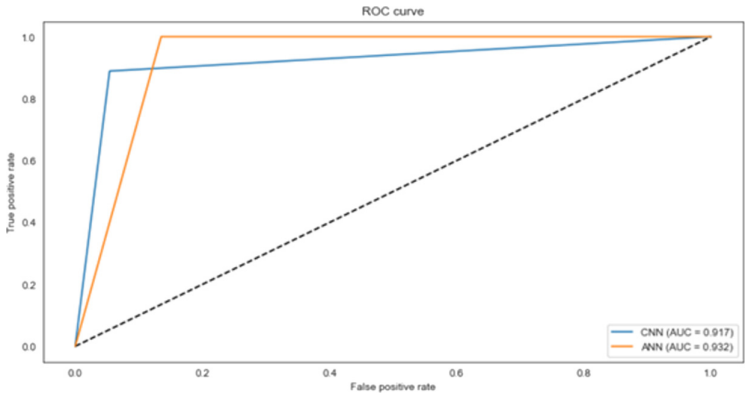

Fig. 4. Performance ROC curve for ANN and CNN

a phonotype. In [12] authors have applied ANN in the diagnosis of Autism Spectrum Disorder using genotyping data of 15 SNPs within four genes. While other studies demonstrated that the use of higher number of SNPs can increase the accuracy of the model [8]. In [39] the authors applied CNN in the diagnosis of amyotrophic lateral scle-rosis using the promoter regions within four genes of the human genome.

This study demonstrates a novel approach for application of genetic markers in the area of Alzheimer disease. Formed of two stages: a feature selection stage and classification stage.

ANN showed a slightly better results than CNN. This could be due to the nature and size of the input data, as CNN are more effective when analysing images data [40]. For data problem solving, ANN is an ideal candidate to do so. Feedforward NN are used to process text data, tabular data and image data with ease. To achieve the desired elevated accuracy, CNN needs a higher number of data inputs. The detection of complex relations, such as nonlinear relations in regard to independent and dependent variables, is achievable by ANN.

The proposed method has few advantages over its competitive. By reducing the feature set size using RF a simpler model can be built. Therefore, it can work on low-performance computing. Feature selection will also reduce the variance of the model, and therefore overfitting and we can reduce the computational cost of training a model.

The results of the proposed model in the current study outperform other models in the area of LOAD GWAS data classification as illustrated in Table 4. However, our study utilized smaller database and future works will involve the use of larger dataset to provide reliable benchmark.

Our work demonstrates the capability of deep learning for classification of AD using correct genetic markers. Romero-Rosales, et al. [41] applied regression analysis method to classify cases from controls by involve the addition of incorrectly classified samples to increase model accuracy.

Table 4. Comparison with previous work

Paper	Year	Dataset size	Classifier	AUC
Romero-Rosales et al. [41]	2020	1,856 AD cases and 2,000 controls	LASSO	0.80
Jansen et al. [42]	2019	71,880 cases and 383,378 controls	PRS	0.827
Proposed model	2021	176 cases and 188 controls	ANN	0.93

4 Conclusion and Future Work

The paper presents the use of hybrid machine learning algorithms for classification of late-onset Alzheimer's disease. To the best of our knowledge, this work is the first of applying these models for Alzheimer's disease. Both models showed promising results, generally ANN showed slightly improved accuracy than CNN. Contrast to various work within the field, our model used two stages for feature selection. First, logistic regression is applied to select the most significant SNPs associated with the disease. Second, Random Forest is applied to further reduces the number of SNPs. Thus, making the classification task computationally efficient.

Future works will invlove the use of larger dataset, helping improve the model, as well as expending our experimnets to other diseases.

References

1. World Alzheimer Report 2018. https://www.alzint.org/u/WorldAlzheimerReport2018.pdf, Accessed 15 Jan 2021
2. Ford, A.: Alzheimer disease. Mol. Chem. Neuropathol. **28**(1–3), 121–124 (1996). https://doi.org/10.1007/BF02815213
3. Isik, A.T.: Late onset alzheimer's disease in older people. Clin. Interv. Aging **5**, 307 (2010)
4. Williamson, J., Goldman, J., Marder, K.S.: Genetic aspects of alzheimer disease. Neurologist **15**(2), 80–86 (2009). https://doi.org/10.1097/NRL.0b013e318187e76b
5. Bekris, L.M., Yu, C.-E., Bird, T.D., Tsuang, D.W.: Review article: genetics of alzheimer disease. J. Geriatr. Psychiatry Neurol. **23**(4), 213–227 (2010). https://doi.org/10.1177/0891988710383571
6. Hofmann-Apitius, M., et al.: Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders. Int. J. Molec. Sci. **16**(12), 29179–29206 (2015). https://www.mdpi.com/1422-0067/16/12/26148
7. Kim, J., Kim, J., Kwak, M.J., Bajaj, M.: Genetic prediction of type 2 diabetes using deep neural network. Clin. Genet. **93**(4), 822–829 (2018). https://doi.org/10.1111/cge.13175
8. Abdulaimma, B., Fergus, P., Chalmers, C., Montanez, C.C.: Deep learning and genome-wide association studies for the classification of type 2 diabetes, pp. 1-8. IEEE (2020)
9. Ghanem, S.I., Ghoneim, A.A., Ghanem, N.M., Ismail, M.A.: High performance computing for detecting complex diseases using deep learning. In: 2019 International Conference on Advances in the Emerging Computing Technologies, AECT 2019 (2020). https://doi.org/10.1109/AECT47998.2020.9194158, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092376858&doi=10.1109%2fAECT47998.2020.9194158&partnerID=40&md5=0252fbd3c9bf9226aaa8482e30f8aaec, https://ieeexplore.ieee.org/document/9194158/

10. Urbanowicz, R., Kiralis, J., Sinnott-Armstrong, N., Heberling, T., Fisher, J., Moore, J.: GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData Mining **5**(1) (2012). https://doi.org/10.1186/1756-0381-5-16
11. Sun, T., Wei, Y., Chen, W., Ding, Y.: Genome-wide association study-based deep learning for survival prediction. Stat. Med. Article (2020). https://doi.org/10.1002/sim.8743
12. Ghafouri-Fard, S., Taheri, M., Omrani, M.D., Daaee, A., Mohammad-Rahimi, H., Kazazi, H.: Application of single-nucleotide polymorphisms in the diagnosis of autism spectrum disorders: a preliminary study with artificial neural networks. J. Mol. Neurosci. **68**(4), 515–521 (2019). https://doi.org/10.1007/s12031-019-01311-1
13. Guo, X., Yu, N., Gu, F., Ding, X., Wang, J., Pan, Y.: Genome-wide interaction-based association of human diseases-a survey. Tsinghua Sci. Technol. **19**(6), 596–616 (2014)
14. Bush, W.S.: Genome-wide association studies. In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (eds.) Encyclopedia of Bioinformatics and Computational Biology, pp. 235-241. Academic Press, Oxford (2019)
15. Clarke, G., Anderson, C., Pettersson, F., Cardon, L., Morris, A., Zondervan, K.: Basic statistical analysis in genetic case-control studies. Nat. Protocols **6**(2), 121–133 (2011). https://doi.org/10.1038/nprot.2010.182
16. Pearson, T.A., Manolio, T.A.: How to interpret a genome-wide association study. JAMA **299**(11), 1335–1344 (2008)
17. Witten, I.H., Frank, E., Hall, M.A.: Chapter 1 - what's it all about? In: Witten, I.H., Frank, E., Hall, M.A. (eds.) Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), pp. 3–38. Morgan Kaufmann, Boston (2011)
18. Lin, E., et al.: A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. Front Psychiatry **9** (2018). https://doi.org/10.3389/fpsyt.2018.00290, (in eng)
19. Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., Aittokallio, T.: Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. **10**(11), e1004754 (2014)
20. Emre Celebi, M., Aydin, K. (eds.): Unsupervised learning algorithms. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-24211-8
21. Lopez, C., Tucker, S., Salameh, T., Tucker, C.: An unsupervised machine learning method for discovering patient clusters based on genetic signatures. J. Biomed. Inf. **85**, 30–39 (2018). https://doi.org/10.1016/j.jbi.2018.07.004
22. Vivian-Griffiths, T., et al.: Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach. Am. J. Med. Genet. B Neuropsychiatr. Genet. **180**(1), 80–85 (2019)
23. Laksshman, S., Bhat, R.R., Viswanath, V., Li, X.: DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning. Hum. Mutat. **38**(9), 1217–1224 (2017)
24. Yang, J., et al.: Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. **42**(7), 565–569 (2010)
25. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. **33**(8), 831–838 (2015)
26. Zhou, J., Troyanskaya, O.G.: Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods **12**(10), 931–934 (2015)
27. Scholz, M., Kaplan, F., Guy, C.L., Kopka, J., Selbig, J.: Non-linear PCA: a missing data approach. Bioinformatics **21**(20), 3887–3895 (2005)
28. Yoon, K., Kwek, S.: An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics. In: Fifth International Conference on Hybrid Intelligent Systems (HIS 2005), p. 6. IEEE (2005)

29. Webster, J.A., et al.: Genetic control of human brain transcript expression in Alzheimer disease (in eng). Am. J. Hum. Genet. **84**(4), 445–458 (2009). https://doi.org/10.1016/j.ajhg.2009.03.011
30. Purcell, S., et al.: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Gen. **81**(3), 559–575 (2007). https://doi.org/10.1086/519795
31. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
32. Cook, J., Mahajan, A., Morris, A.: Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. Eur. J. Hum. Gen. **25**(2), 240–245 (2016). https://doi.org/10.1038/ejhg.2016.150
33. Chang, M., He, L., Cai, L.: An overview of genome-wide association studies. In: Huang, Tao (ed.) Computational Systems Biology. MMB, vol. 1754, pp. 97–108. Springer, New York (2018). https://doi.org/10.1007/978-1-4939-7717-8_6
34. Curbelo, C., et al.: SAERMA: stacked autoencoder rule mining algorithm for the interpretation of epistatic interactions in GWAS for extreme obesity. IEEE Access **8**, 112379–112392 (2020). https://doi.org/10.1109/ACCESS.2020.3002923
35. Fergus, P., Montanez, C.C., Abdulaimma, B., Lisboa, P., Chalmers, C., Pineles, B.: Utilizing deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American women. IEEE/ACM Trans. Comput. Biol. Bioinf. **17**(2), 668–678 (2020). Art no. 8454302, https://doi.org/10.1109/TCBB.2018.2868667
36. Aggarwal, C.C.: Neural networks and deep learning. Springer **10**, 978–983 (2018)
37. Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K.: Convolutional neural networks: an overview and application in radiology. Insights Imag. **9**(4), 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9
38. Bush, W., Moore, J.: Chapter 11: genome-wide association studies. PLoS Comput. Biol. **8**(12), e1002822 (2012). https://doi.org/10.1371/journal.pcbi.1002822
39. Yin, B., et al.: Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype (in eng). Bioinformatics **35**(14), i538–i547 (2019). https://doi.org/10.1093/bioinformatics/btz369
40. Sharma, P., Singh, A.: Era of deep neural networks: a review. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 3–5 July 2017, pp. 1–5 (2017). https://doi.org/10.1109/ICCCNT.2017.8203938.
41. Romero-Rosales, B.-L., Tamez-Pena, J.-G., Nicolini, H., Moreno-Treviño, M.-G., Trevino, V.: Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling. PloS One **15**(4), e0232103 (2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7179850/pdf/pone.0232103.pdf
42. Jansen, I.E., et al.: Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat. Gen. **51**(3), 404–413 (2019). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6836675/pdf/nihms-1031924.pdf