PAPER 2

PROBLEM

Prediction

OBJECTIVE

PREDECTION of Alzheimer Disease Using Genome Data

Methods

hybrid machine learning algorithms for classification of late-onset Alzheimer's disease

logistic regression for the association test

multilayer perceptron (MLP) and Convolution Neural network (CNN) for our experiments.

PLINK: Plink v1.9 is used in this study to perform all quality control procedures and preliminary analysis.

Association Test - Logistic Regression: explains and assess the relations between the SNPs and the dependent variable. Association test used in the current work to rank each SNP with a significant value.

Random forest: 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 training dataset

Pooling layer after the conventional layer

Advantages

Reduce the number of features of the GWAS SNPs to allow fast and computational efficiently processing of data.

Improve the classification accuracy of Alzheimer's' vs healthy individuals

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

DATASETS

The GWAS case-control dataset

RESULTS

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

Paragraph

Based on the age of the onset, there are two subtypes to the disease, they are: Early-onset AD (EOAD) and late-onset AD (LOAD). Approximately 5% of AD cases show EOAD the age-onset ranges from the 30's to early 60's. 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%. Genetic and environmental factors are both part of the LOAD which seems showing more complex disorder. They have used 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**. 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.