

Detection of Alzheimer disease using mRNA gene expression data.

Canberk Arıcı

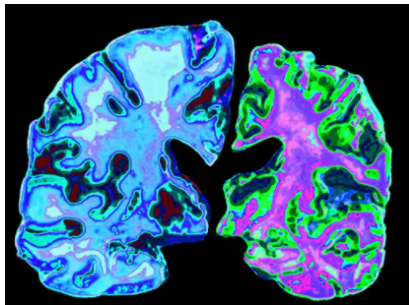
Advisor: Assoc. Prof. Habil Kalkan

15 Jun 2022



1. Project Definition
2. Project Design
3. Project Results
4. Success Criteria
5. References





In this project, Alzheimer's Disease will be detected by using mRNA gen expression data. Three different datasets are provided as combined. 5 different machine learning models are performed to classify these samples. These models are CNN, MLP, SVM, Random Forest and KNN.



Dataset Information:

MCI(Mild Cognitive Impairment) is a transitional stage between normal aging and Alzheimer's Disease.

Total "Gene Symbol" being 11618; There are 482 AD(Alzheimer Disease), 313 MCI, 467 Control samples.

- ▶ GSE63060 Normalized Dataset; 29958 Gene Symbol, 329 Samples: 145 AD, 80 MCI, 104 Control
- ▶ GSE63061 Normalized Dataset; 24900 Gene Symbol, 388 Samples: 139 AD, 109 MCI, 134 Control
- ▶ GSE140829 Normalized Dataset; 15987 Gene Symbol, 551 Samples: 198 AD, 124 MCI, 229 Control

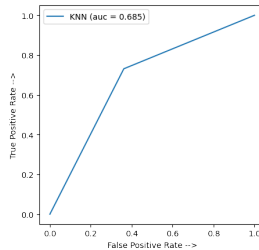
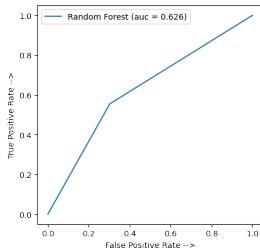
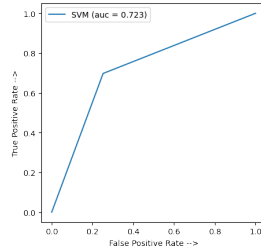
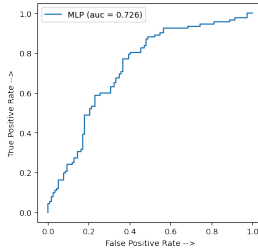
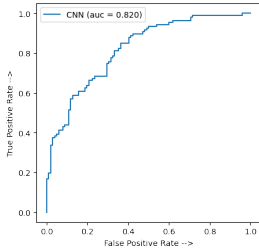








- ▶ PCA is applied to dataset and optimal number of components is decided.
- ▶ One Hot Encoding is applied to the class feature.
- ▶ Different well known model architectures such as AlexNet are used with CNN and other well known model architectures are used with MLP but results were not satisfactory therefore model architectures are created from scratch.
- ▶ Model performances are tried to be improved both manually and by using some techniques such as GridSearchCV.
- ▶ GridSearchCV is used to find optimal parameters for models and is used to evaluate models' performances.



- ▶ EarlyStopping of Keras is used to prevent overfitting.
- ▶ ROC Curve is plotted and AUC Score is calculated for all models then these plots and scores are observed.
- ▶ An additional dataset to the existing datasets is tried to be found but couldn't be found.
- ▶ 2 of the 3 classes are used in order to improve models' performances. These classes are AD and CTL.






 MODEL / AUC SCORE	AUC SCORE
 CNN	0.82
 MLP	0.726
 SVM	0.723
 KNN	0.685
 Random Forest	0.626



Success Criteria

- ▶ Having a minimum 70% AUC score with RNA classifier on the testing data set.
- ▶ Performances of 5 different classification models will be compared. At least one of the classification algorithms will be deep learning model.
- ▶ At least one additional dataset will be combined with provided datasets.



-  *Integrated analysis of differential gene expression profiles in hippocampi to identify candidate genes involved in alzheimer's disease.*
-  *Prediction of alzheimer's disease using blood gene expression data.*
-  *A novel multi-tissue rna diagnostic of healthy ageing relates to cognitive health status.*
-  Vo Van Giau Eva Bagyinszky and SeongSoo A.An, *Transcriptomics in alzheimer's disease: Aspects and challenges.*
-  Justin Miron, Cynthia Picard, Nathalie Nilsson, Josée Frappier, Doris Dea, Louise Thérourx, and Judes Poirier, *Cdk5rap2 gene and tau pathophysiology in late-onset sporadic alzheimer's disease, Alzheimer's Dementia* **14** (2018).

