Early detection of Alzheimer’s disease using blood gene expression data

Progress Report – 5-12-2022

Hariharan J Dr.Jothi R

**Abstract:** Alzheimer’s a type of neurological disorder is seeing a surge in the numbers of cases from the past decade. Especially in India the cases are expected to grow to 11,422,692 by 2050 from 3,848,118 measured in 2019 according to Lancet report as of July 2022. Early detection of the genes responsible for Alzheimer's is an important pre-requisite for its diagnosis. This study focuses to explore the various feature extraction and classifier algorithms that make use of gene expression data and find and design the best among them extract the specific genes (feature selection) and their characteristics that translate to Alzheimer’s disease (classification). Gene expression datasets used for the study are curated from multiple public repositories such as ADNI, AddNeuroMed1, AddNeuroMed2. The classifiers algorithms are put for internal validation (training and testing dataset are from same set) and external validation (training and testing dataset are from different sets). Study also aims to explore explainable artificial intelligence methods (XAI) of classification for a simple human interpretation and measure its trustworthiness.

**Work Done:**

**Dataset Collection:** Datasets are curated from three sources viz.

* <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63060>
* <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63061>
* <https://adni.loni.usc.edu/data-samples/access-data/>

**Dataset Preprocessing:**

* Explored different gene expression processing packages in python and found GEOparse to be suitable.
* Extracted only the columns necessary for classification such as RNA values, Sample ID, Status (AD/CTL/MCI), Age, Gender, Ethnicity.
* Renaming the column names having Probe Id with their respective Gene Id.
* Grouping the columns with same gene identifier and taking a median RNA value for the resultant column.
* Removing row with Status values as MCI as study is to focus of classifying AD vs CTL.

**Feature Extraction:**

* **Chi Square:** 
  + Normalized the RNA expression values of Gene Identifier using Min-Max normalization.
  + X = values of Gene Identifier columns and Y = Status column.
  + With K = 200, X, Y chi square fit is applied
  + X\_Significant = top 200 gene from chi square fit output.

**Classification:**

* **SVM:**
  + 70:30 split is done on the dataset to get training and testing set.
  + With X\_Signification and Y, SVM liner kernel classification carried out.
  + 93.67% and 77.33% accuracy is found on training and testing set respectively.