Alzheimer’s disease stages classification

Genetics boosted data science approach

Sijie Sun, sijiesun@g.harvard.edu

Yongcheng Wang, yongchengwang@g.harvard.edu

1.Introduction

Nowadays, the number of people with dementia is over 35 millions worldwide and is expected to grow dramatically as the population ages and competing causes of death in late life continue to recede. The growth rate is even greater in the developing world than in the high-income countries (Fig. 1) [1, 2]. Alzheimer’s Disease (AD) is the most common form of dementia, with both environmental and genetic factors contributing to risk, which affects more than 13% of individuals aged 65 years and older and 30–50% of individuals aged 80 years and older [3,4].

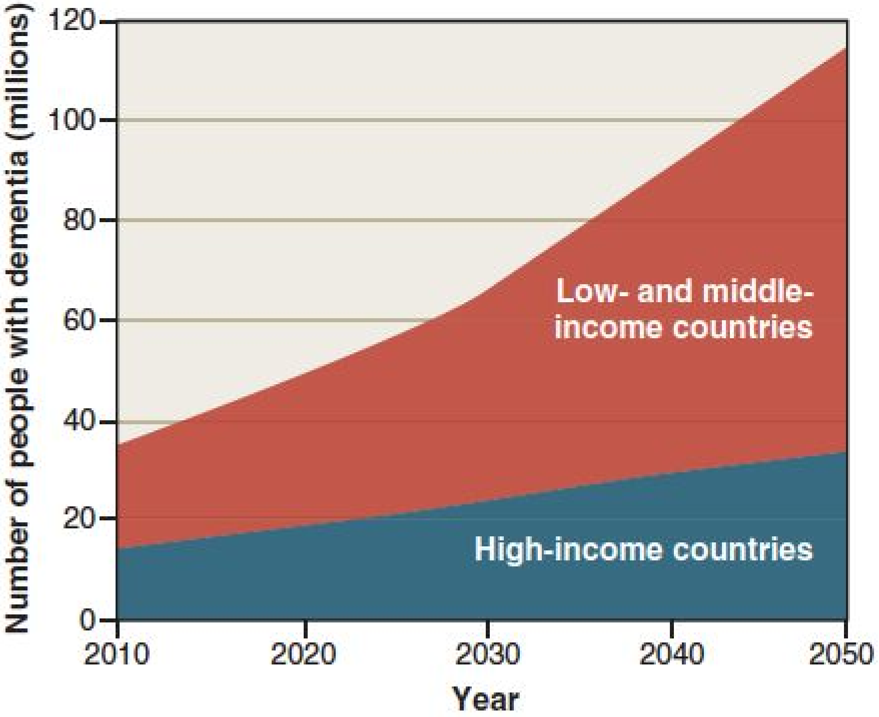

Fig. 1. Projected increases in the numbers of people with dementia in high-income countries and in low- and middle-income countries [1, 2].

Unfortunately, Alzheimer has no current cure, therapeutic interventions directed only at the mild-to-moderate clinical stage is too late to ameliorate symptoms, secondary prevention (diagnosing and treating the disease before overt symptoms) is more likely to slow the pathogenic process [5]. Therefore, it is crucial to develop a method to give long term health prediction before a person head into late age.

2. Goal and Approach

In this project, We target in long-term prediction of AD and MCI (Mild cognitive impairment). To do so, we extract people's clinical background data and genetics data, those data from the database. In practice, compared with imaging, they are easily accessible. so that could be used for widely screening and long-term monitoring. Our clinical data come from AD Challenge Training Data: Clinical (Updated), ADNI. Our genetic data come from ADNI Gene Expression data.
The training data consist of individuals participating in the Alzheimer's Disease Neuroimaging Initiative (ADNI). The study has grown to >1600 participants with mild cognitive impairment (MCI),ranging from Early MCI (EMCI) to Late MCI (LMCI), Alzheimer's Disease (AD) and elderly, cognitively normal (CN) control subjects that have been recruited over three phases ADNI 1, ADNI GO and ADNI 2. The study consists of a multitude of data from assessments, biospecimens, genetic and imaging analysis, in addition to subject characteristics and medical history that are all available through the LONI database. We uses a subset of the ADNI data relevant to our prediction.

Alzheimer's disease is genetically complex and shows heritability of up to 79%. By checking the literatures, we found several genes are associated with Alzheimer. 1. *APOE* (encoding apolipoprotein E) is the most well known gene to increase disease risk for the common form of Alzheimer's disease with late onset. 2. A small fraction (<1%) of all AD cases arises during middle age because of inherited missense mutations in one of three genes: *APP, PSEN1,* or *PSEN2*. 3. The Alzheimer Disease Genetics Consortium (ADGC) performed a genome-wide association study of late-onset Alzheimer disease using a three-stage design consisting of a discovery stage (stage 1) and two replication stages (stages 2 and 3). They identified common variants at *ABCA7*, *MS4A6A/MS4A4E*, *EPHA1*, *CD33* and *CD2AP* are associated with Alzheimer's disease [3], and common variants at *MS4A4/MS4A6E*, *CD2AP*, *CD33* and *EPHA1* are associated with late-onset Alzheimer's disease [4]. Therefore, we picked up the gene expression data of those genes, including *APP, PSEN1, PSEN2, APOE, ABCA7, MS4A6A, MS4A4E, EPHA1, CD33, CD2AP, MS4A4, MS4A6E*, and used those gene expression data as predictors.

Data_Science-wise, we combine EDA and classification algorithms to analyze the data. The potential challenge includes small datasize. Missing data, and so on.

3. Results

Firstly, we conduct elementary EDA towards our data, the basic information includes:

- -Years of Education
- -Age
- -Gender

And there is one genetic information given by the challenge data:

- -APOE4

Which is the most well known mutation towards AD.

Firstly we look into the education distribution of different MCIs (Fig.2). It is known that AD is negatively correlated with Years of Education, this trend is also reflected in this dataset.

We can notice that shorter education time is correlated with AD or LMCI. However, there is one thing remarkable. The years of education among the samples are significantly higher than the averaged Years of Education (YoE) in US, which is 12 years. This implies that the samples are well-educated and perhaps have enough economic support for medical intervention. So this sample is heavily biased in YoE.

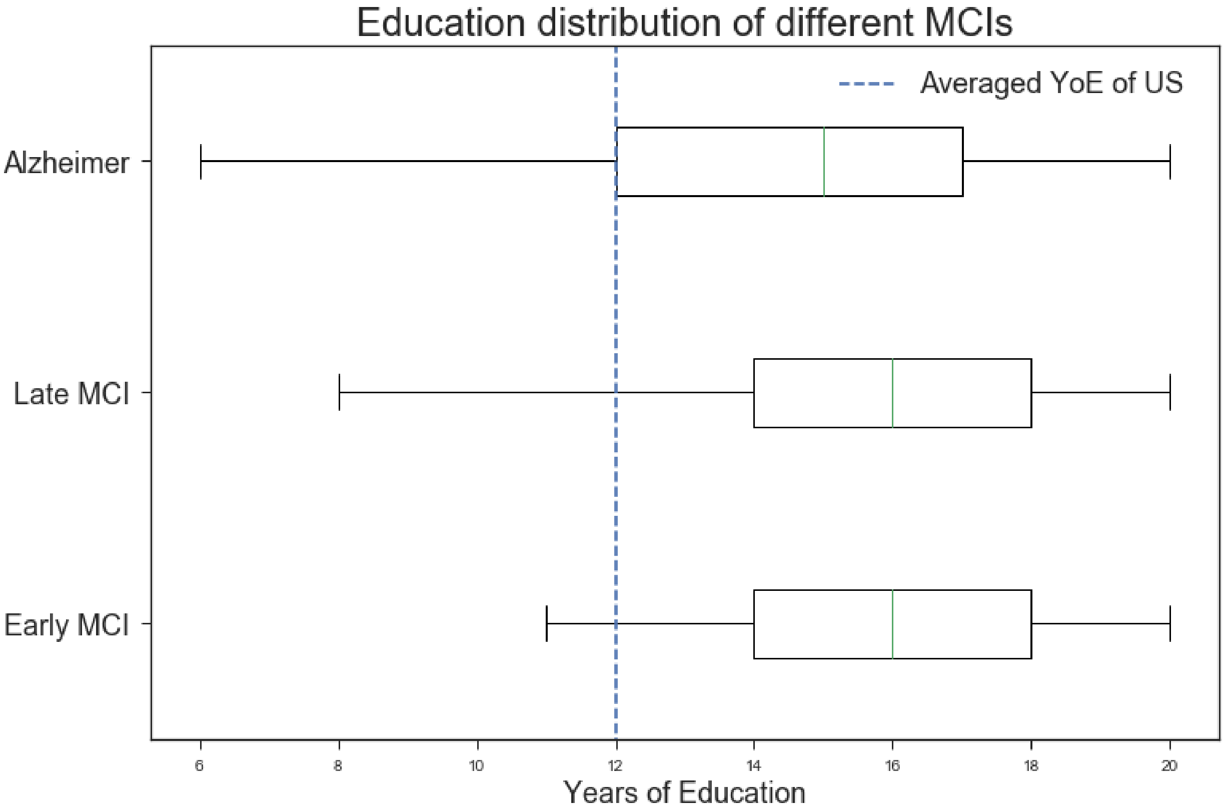

Fig. 2. Education distribution of different MCIs

Moreover, we take a look into the age distribution (Fig. 3). AD and MCI is known to develop in old age. This is reflected in the age distribution in the samples.

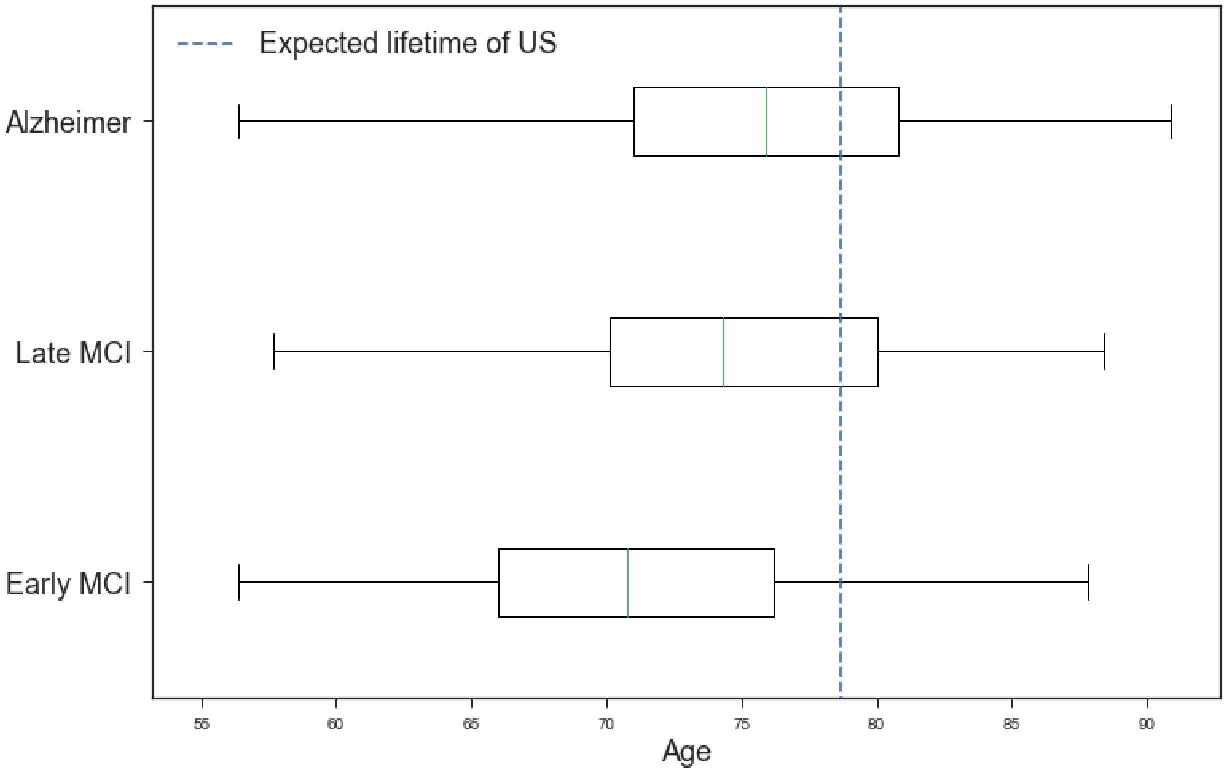

Fig. 3. Age distribution of different MCIs

Notice that expected lifetime in US already bypass the median age of AD patients, which means if patients have genetic defect, or other triggering factors. They are most likely to suffer from AD in their old age.

Regarding the genetic mutation, there is one mutation given by the challenge database: APOE4, this is the most well known mutation related to Alzheimer [3,4]. This is an integer predictor from 0 to 2 (Fig.4). So we use average instead of median. From the dataset, we can see that the predictor is real.

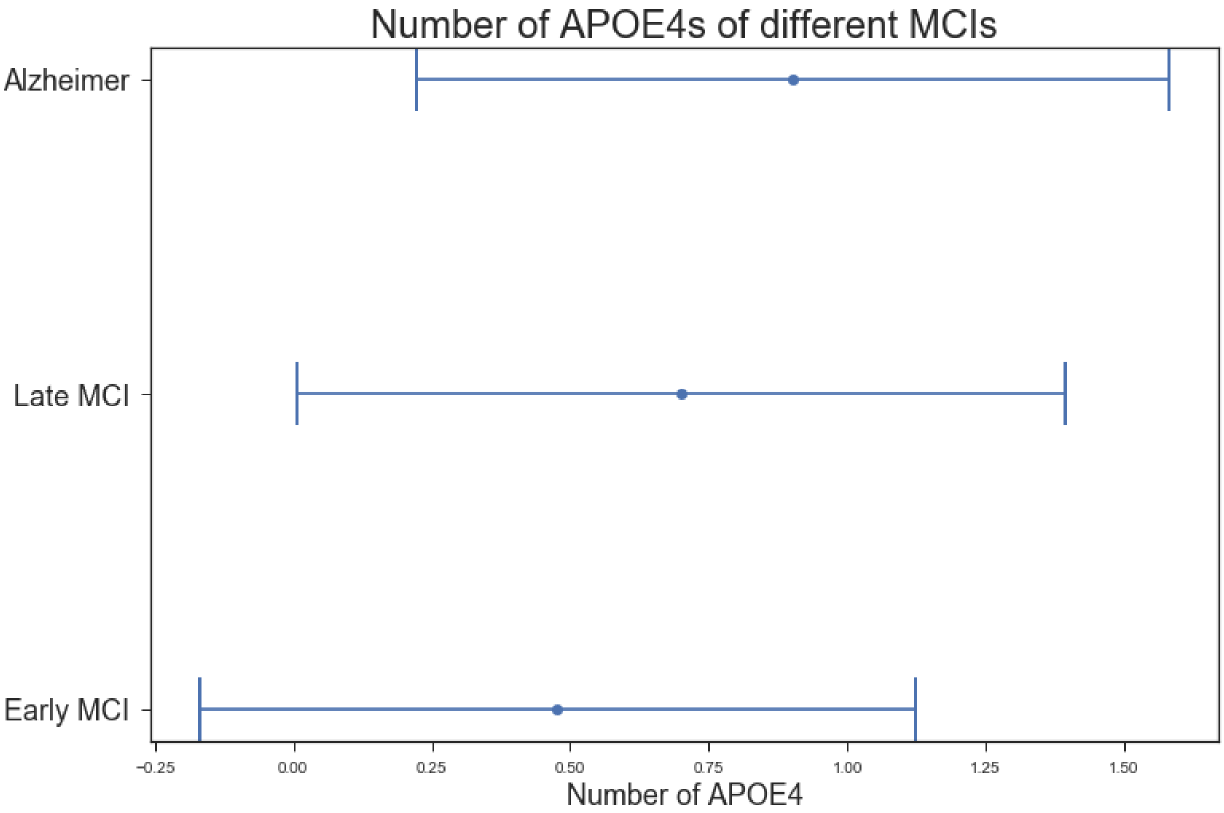

Fig. 4. Number of APOE4s of different MCIs

Finally we take a look at MMSE score (Fig.5). MMSE is a test score reflecting the cognitive impairment. This is not a predictor, as AD leads to cognitive impairment. Here we can see when people are diagnosed as AD, their MMSE still have 20-25, which is still mild cognitive impairment. So it is not too late to interval.

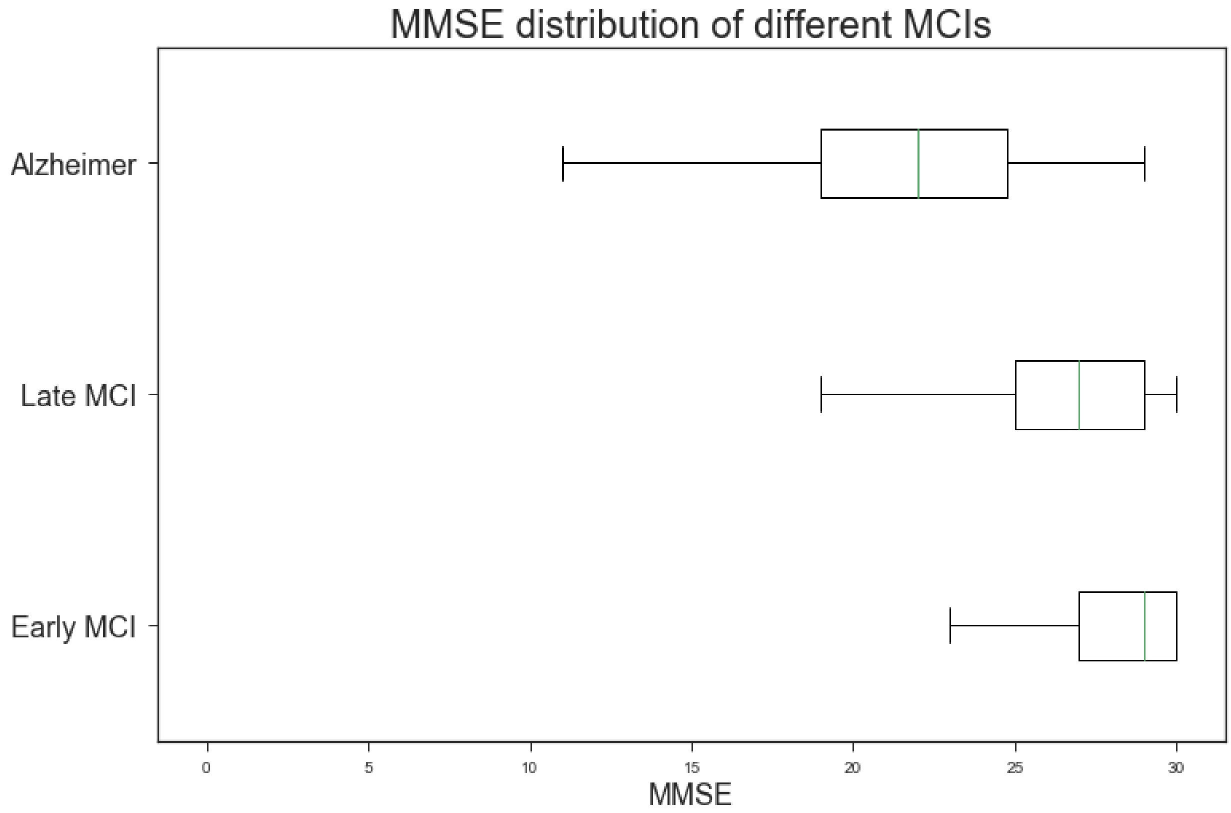

Fig. 5. MMSE distribution of different MCIs

Regarding the benchmark model, we adapt kNN model with cross validation to determine the number of neighbours. In this model we have four predictors: Age, Gender, Years of Education and number of *APOE4* mutation (Fig. 6).
The best result is 39% accuracy, with 14 neighborhoods. Regarding the AD, it has 10% TPR rate. The confusion matrix is given below.

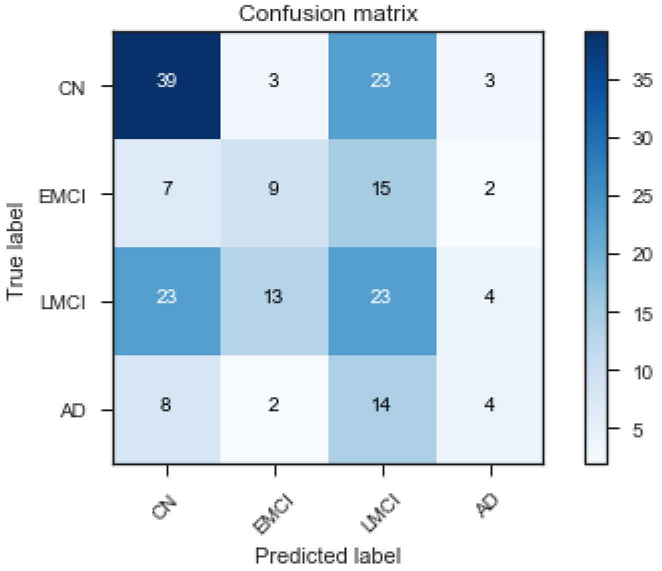

Fig.6. The Confusion matrix for the kNN model.

We firstly try to improve the performance with SVM method (Fig. 7), SVM is known to be a robust classifier when facing complex boundary and can partially overcome overfit problem. The penalty factor is determined by cross validation. The main disadvantage is the computation is really expensive, so we only adapt linear boundary. Since this is a biased dataset, we expect penalty factor to be larger than 1.
This is proven by the calculation. The best accuracy is 45% on test dataset, with penalty factor as 1000. The TPR of AD is also improven to 15%. Here we almost exploited the original dataset.

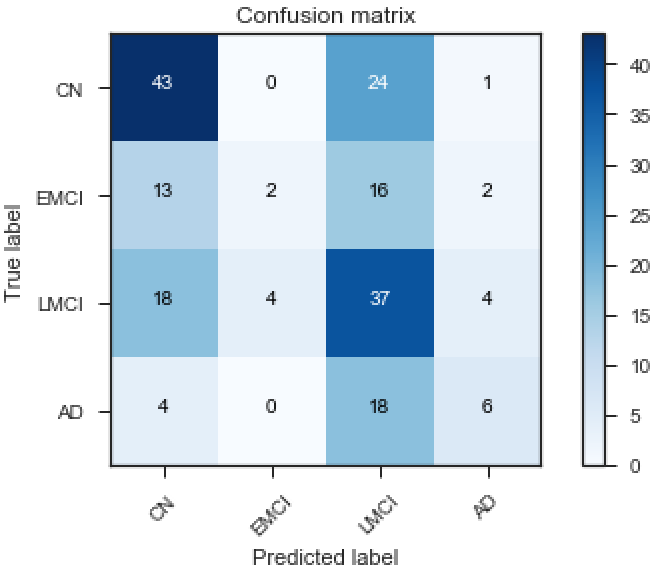

Fig. 7 Confusion matrix of SVM model

Additional predictors

Here we try to introduce more predictors based on their gene expression, the data is aligned with the challenge data we used in the previous question, The first challenge we met is about half of the patients in the challenge data are missing gene information, this is because the data is collected based on different versions of guidelines in span of many years. The early data does not have gene information. And some of the gene data does not pass quality control, The strategy possible including eliminate missing data, median replacement or kNN replacement. Notice half of the data does not have gene information. The only choice is to abort them.

Here we only have 276 data points left. But the number of predictors is greatly improved from 4 to 18. The additional predictors include:
['ABCA7','MS4A6A','MS4A4E','EPHA1','CD33','CD2AP','MS4A4','MS4A6E'])
Some of them have more than one expression data. Then we conduct quality check towards the new predictors. We use random forest method to keep half of the predictors and check the relative importance of the predictors, With 200 trees. The results are given below:

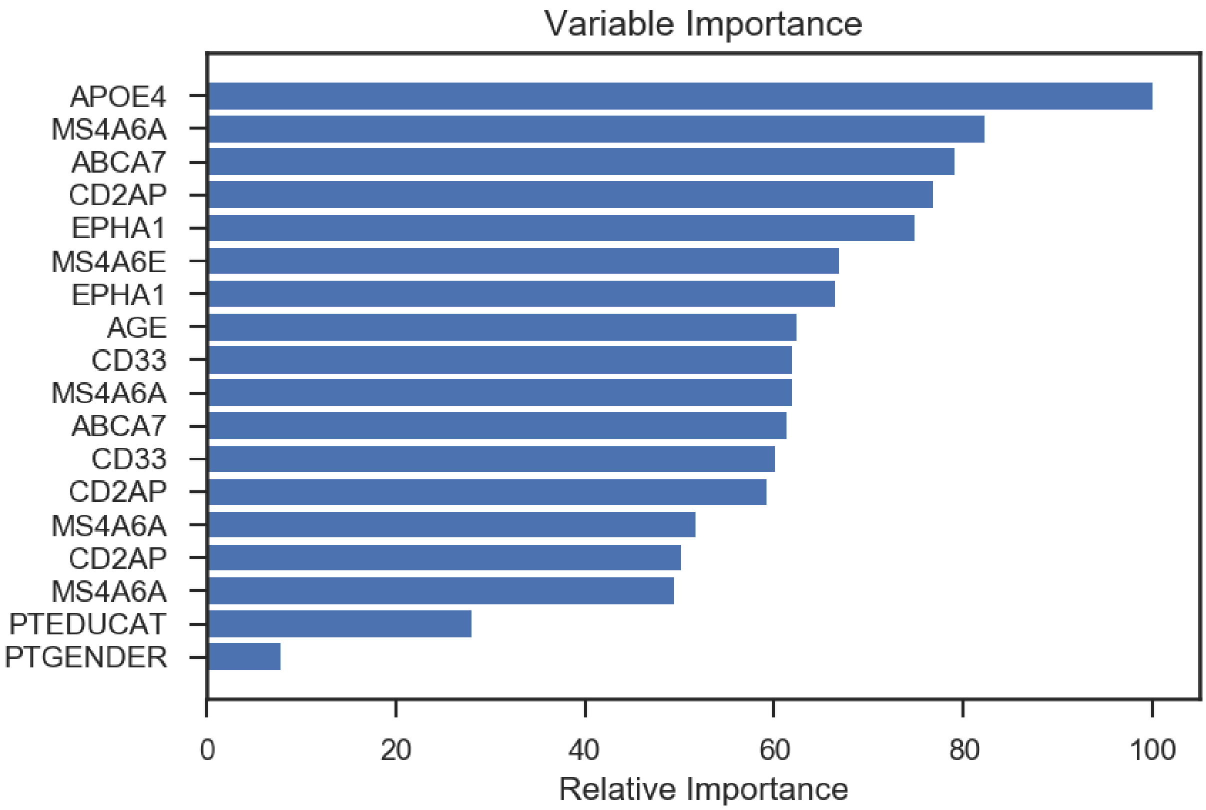

Fig. 8. Relative importance of various predictors

Here we can see, *APOE4* is the most significant predictor. And other gene expressions are also more significant than previous predictors like age, years of education or gender. Also we notice that Gender and Years of education is not very significant in this case.

Then we try to fit model onto the new dataset. We tried LDA, QDA, SVM and random forest four methods. Here we notice the second trouble: overfitting.
The accuracy of SVM is 45%, 40% for RF, 45% for LDA and 40% for QDA. SVM has better performance as it is more robust against overfitting.

The number of predictors is 10% of the training data set. So we notice against the fact we introduced more high quality predictors, the accuracy of the model itself is not improved. To overcome this problem, we introduce PCA into the predictors (Fig. 9). PCA vectors with more than 5% explanation ratios are keeped (first 8). With the new model, the random forest has 58.2% accuracy, which is about 20% higher than our benchmark model.

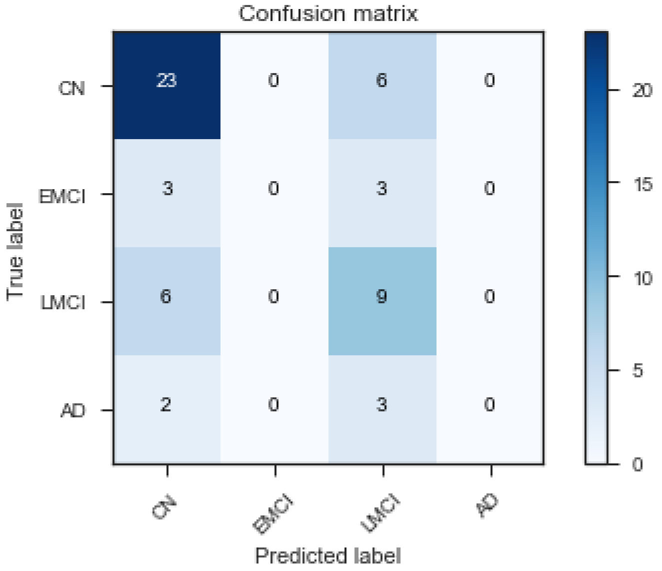

Fig. 9. Confusion matrix of gene expression with PCA

Summary & Shortage & Future work

In this project, we took a look into both the clinical and gene expression data of the patients. The introduction of the gene expression further increase the number of high quality predictors. But the keep increasing number of predictors lead to overfit. This problem can be partially alleviated via SVM or PCA. Especially the PCA method lead to significant boost in the prediction accuracy.

	Clinical Data		Gene expression data included				Gene+PCA
Method	kNN	SVM	SVM	RF	LDA	QDA	RF
Score	39.1%	45.8%	45.5%	40.0%	45.5%	41.8%	58.2%

The main shortage is the TPR is fairly low in all of the cases. Without Gene data, SVM has 15% TPR. And with the gene data added, the number of AD patients down to such a low level that TPR is no longer available. Small dataset training is truly a challenging topic.

In the future work, current model can be further improved by collecting more data from patients. And more significant gene predictors can be included to further boost the performance. Data science wize, if we can qualitatively predict the probability of a person to be AD or MCI, its value will be higher than simple classification. So we may adapt logistic regression for quantitative prediction. In addition to the ADNI gene expression data, we could further incorporate the genome-wide association study (GWAS) results and DNA methylation data into the predictions to give fully coverage of patient’s genetic information.

Reference

1. Selkoe, Dennis J. "Preventing Alzheimer’s disease." *Science* 337.6101 (2012): 1488-1492.
2. A. Wimo, M. Prince, World Alzheimer Report 2010; the Global Economic Impact of Dementia (Alzheimer’s Disease International, London, 2010).
3. Hollingworth, Paul, et al. "Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease." *Nature Genetics* 43.5 (2011): 429-435.
4. Naj, Adam C., et al. "Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease." *Nature Genetics* 43.5 (2011): 436-441.
5. https://www.alz.org/alzheimers_disease_what_is_alzheimers.asp
6. http://www.nationmaster.com/country-info/stats/Education/Average-years-of-schooling-of-adults