Early detection of Alzheimer's disease using Brain MRI scans

Akshay Iyer
Department Of Robotics
Worcester Polytechnic Institute
Worcester, Massachusetts
abiyer@wpi.edu

Yuchen Shen
Department Of Data Science
Worcester Polytechnic Institute
Worcester, Massachusetts
yshen4@wpi.edu

Omkar Kulkarni
Department Of Data Science
Worcester Polytechnic Institute
Worcester, Massachusetts
okulkarni@wpi.edu

Shengmei Liu
Department Of Computer Science
Worcester Polytechnic Institute
Worcester, Massachusetts
sliu7@wpi.edu

***Abstract*—Alzheimer's disease(AD) is a neurodegenerative disease and is known to be the primary cause of dementia in old individuals[1]. While it is very difficult to cure, treatment is most effective when started early in the disease process[2]. In this study, we present and compare models which will help doctors classify a patient as being diagnosed with Alzheimer's/ normal condition/ having a mild cognitive impairment (early stage). We implemented several machine learning models individually, then created ensemble methods to improve the accuracy and finally ran a deep learning model on the raw images. The performance was measured using accuracy and f1 score metrics and we performed a comparative analysis of all the various models based on the performance metrics. Our study shows that Gradient Boosting (among machine learning algorithms) and Convolutional Neural Networks give the highest accuracy in detecting Alzheimer's correctly.**

Keywords—Alzheimer's, Machine Learning, deep learning, accuracy, ensemble

I. Introduction

Alzheimer's disease (AD) is a progressive, neurodegenerative brain disorder that attacks neurotransmitters, brain cells, and nerves, affecting brain functions, memory, and behaviors. Worldwide, nearly 44 million people have Alzheimer's or a related dementia [3]. It is the 6th leading cause of death in the US and it kills more than breast cancer and prostate cancer combined. The global cost of Alzheimer's and dementia is estimated to be $605 billion, which is equivalent to 1% of the entire world's gross domestic product. Currently, only 1-in-4 people with Alzheimer's disease get diagnosed[17]. Treatment of Alzheimer's and other dementia-causing diseases is typically most effective when started early in the disease process[17] and early and accurate diagnosis could save up to 7.9 trillion USD in medical and care costs[17]. Therefore we designed a series of individual classification models, create different ensembles of them and also applied deep learning, in an endeavor to find an optimal classifier to detect Alzheimer's. Since our classifier would also be able to predict if patients have a Mild Cognitive Impairment, we would be able to detect Alzheimer's in the early stages itself. The main challenges faced were a lack of a good dataset which contains features extracted from scans and enough data to avoid overfitting when there are 2000+ features. We used multiple methods to tackle the same which are mentioned in the paper. The machine learning models could augment the capabilities of a doctor since then, doctors can feed the MRI scan/ data of a patient whom they suspect of developing Alzheimer's to the algorithm. If diagnosed positively and correctly, their chances of recovery increases.

II. Background

A. State of the art methods:

1. Kloeppel S et al (2008) Automatic classification of MR scans in Alzheimer's disease[4]
    a. Used support vector machines with Voxel Based Morphometry to extract features to achieve an accuracy of 94.62%
    b. However, it is a very small and balanced dataset consisting of only 34 positive and negative examples each
2. Gaussian process regression (GPR) model by Fan Zhu, Yuanfang Guan of University of Michigan[5] :
    a. Achieved the highest accuracy score in Alzheimer's Big Data Dream Challenge

   b. However, solution considered only 2 clinical feature and 20 image features of all 2170 features and discarded others
   c. Method did not consider boosting/ bagging/ applying deep learning
3. Nearest shrunken centroid method by The University of Texas Southwestern Medical Center[6]
   a. The advantage is its a simple model and achieved an accuracy score of 44%
   b. However, used just one model and could try ensemble methods to improve accuracy
4. Ensemble of SVM and Gaussian Process Regression[7]
   a. Used ensemble methods and achieved an accuracy of 50%
   b. Could try more models for the ensemble
5. Early diagnosis of Alzheimer's disease with deep learning - Siqi Liu, Sidong Liu, Weidong Cai, Sonia Pujol, Ron Kikinis, Dagan Feng [18]
   a. Used stacked autoencoders and a softmax output layer to diagnose AD/MCI/CN
   b. Requires less labeled training samples and minimal domain prior knowledge
6. Ensembles of Deep Learning Architectures for the Early Diagnosis of the Alzheimer's Disease - Andrés Ortiz, Jorge Munilla, Juan M. Górriz and Javier Ramírez [19]
   a. Two deep learning based structures and four different voting schemes were implemented and compared
   b. Achieved an accuracy of 90%

B. Our computational approach

We are using the same AD challenge Training data dataset (as used by models 2,3,4 above) compiled by ADNI - Alzheimer's Disease Neuroimaging Initiative. It consists of data about 628 individuals that are part of the ADNI1.

Steps followed:

1. Cleanse the dataset consisting of the morphometric and clinical features
2. Perform manual feature selection to remove features like Patient ID, Directory Number, Roster ID
3. Run the state of the art machine learning algorithms using sklearn libraries[9] - Random Forests, Support Vector Machines, Gaussian Process Classification, Artificial Neural Networks and Logistic Regression individually on the dataset and perform
4. Tabulate the performance measures - accuracy score and f-measure of each of the algorithms
5. Implement bagging for the individual algorithms and tabulate results
6. Implement boosting - Adaboost for the appropriate individual algorithms and Gradient Boosting Classifier and tabulate results
7. Implement several combinations of weak classifiers to form a strong classifier using majority voting scheme
8. Repeat Steps 3-7 after performing dimensionality reduction with PCA
9. Train and test a Convolutional Neural Network (CNN) on the corresponding raw images and tabulate results
10. Compare the performance of all models

C. Advantages and disadvantages of our approach:

Advantages:

1. Multiple classical algorithms :
   a. Not limiting solution to just one or two algorithms as the current methods above but instead trying 5 individual algorithms
2. Ensemble learning :
   a. Using various ensemble learning methods and trying several combinations of the same
   b. Performing bagging on individual algorithms, boosting (Adaboost, Gradient Boosting), ensemble using majority voting
3. Applying deep learning (CNN) on the corresponding brain MRI scans

Disadvantages:

1. Less training examples - there are only 628 training examples in the dataset
2. No direct extraction of features from the raw images directly using methods like VBM
3. Models require data processed by the pipeline of software mentioned above which is not always possible
4. Did not use raw image data in the entirety due to computational limitations

D. Paper organization

The paper is organized as follows: Section IIIA - 'Dataset Description' introduces the dataset used for the project, Section III B - 'Basic Models' explains the various machine learning/ deep learning models used to classify the patients and the assessment protocol used for comparing the various models. Section IV - Results presents the results of the comparative analysis and Section V - Conclusion presents the conclusions and the way forward.

III. Methodology

Overview: The problem is to classify a patient as having Alzheimer's, being cognitively normal or having a Mild cognitive impairment using Brain MRI scans. The approach we follow is to implement classical machine learning models and then ensemble models on the features extracted from the

to extract from the scans. Individual models like Logistic Regression, SVM, Random Forest, Gaussian Process Classifier and ANN are implemented. Based on these individual classifier, ensemble learning models, such as bagging, boosting are implemented. Besides those basic machine learning algorithms, we also present the deep learning model, CNN, to deal with the original image data, and comparing the result with result from processed data using machine learning. We perform hyperparameter tuning and principal component analysis in order to improve accuracy. Finally, to evaluate the performance, both accuracy and f-measure are used to show a combined result.

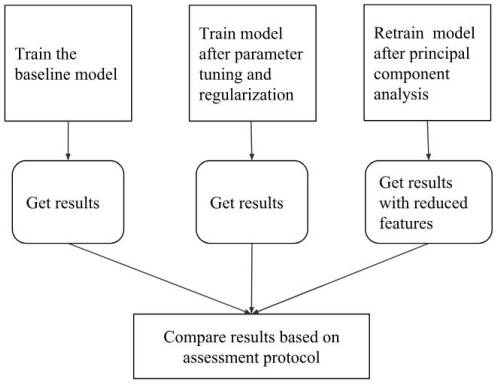

Figure 1. Training and assessment strategy employed in the study

A. Data description :

The dataset consists of Complete 1Yr 1.5T collection[8] and their assessments at baseline. It contains raw images which are the brain MRI scans and features derived from them. Features are morphometric (shape) data such as surface area, travel depth, geodesic depth, etc of the various regions of the brain and other features provided include clinical data like age, years of education, gender, APOE4 and APOE allele 1 and allele2 genotypes. There are 2160 features in total. All images were processed using three neuroimaging software pipelines: FreeSurfer, Advanced Normalization Tools (ANTs), and Mindboggle.

1. Processed image data[9]

We are using the same AD challenge Training data dataset (as used by models above) compiled by ADNI - Alzheimer's Disease Neuroimaging Initiative. It consists of data about 628 individuals that are part of the ADNI1 : Complete 1Yr 1.5T collection and their assessments at baseline.

There are 639 sample patients, and each sample has 2159 features; 2150 features extracted and processed from the original image data, including the following per cortical surface label

- Surface area
- Travel depth
- Geodesic depth
- Mean curvature
- Convexity (FreeSurfer)
- Thickness (FreeSurfer)

The remaining are clinical data about the patient like

- Age
- Gender
- Education
- Ethnicity
- Race
- Mini Mental State Examination (MMSE)
- Gene Type

and the brain status which could be

- Cognitively Normal (CN),
- Alzheimer Disease (AD),
- Mild Cognitive Impairment (MCI)

and these three are our target class names.

2. Original image[9]

From the processed data, we download the corresponding image from ADNI collections. The data include NIFTI image and XML with metadata; Every NIFTI image file contains different numbers of brain slicing images with size 256*256 for a single patient, which can form a 3-D image; The XML file contains date, labels and etc.

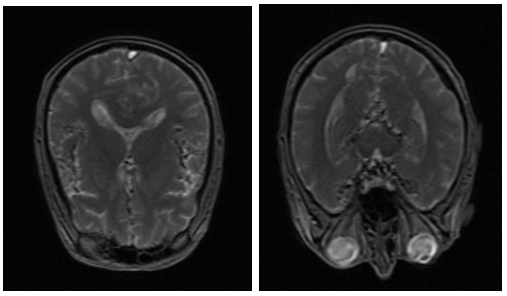

Figure 2. **Example image of patient brain scanning slice in NIFTI image file with size 256*256.** The image is created out of the .nii files in the ADNI dataset

B. Basic Models :

We employed different models to the dataset, and in order to handle the challenge of small data set, we apply bootstrap which is a method of resampling, that will help us to generate more data for training. The final result will be the max-voting result from each bootstrap model.

1. Logistic Regression:

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.[10]

We also did parameter tuning in Logistic Regression. In logistic regression new use C as tuning/regularization parameter. $C = 1/\lambda$. Lambda (λ) controls the trade-off between allowing the model to increase its complexity as much as it wants with trying to keep it simple. We used c = 2.5 and random state = 0 and retrained the model.

2. Support Vector Machines :

A Support Vector Machine (SVM) is a discriminative classifier which, given labeled training data, outputs an optimal hyperplane which categorizes new examples. In 2D space this hyperplane is a line dividing a plane into two parts where in each class lay in either side.[11]

We used baseline SVM with Linear and RBF kernels and default parameters which gave very low accuracy.

Hyper-Parameter Tuning in SVM :

Hyper-parameters are parameters that are not directly learnt within estimators. In SVM they are passed as arguments to the constructor of the estimator classes. Typical examples include changing values of C, types of kernel and values of gamma for Support Vector Classifier.[12]

After doing hyper-parameter tuning with the baseline SVM on our dataset, we got C=10 and Kernel = Linear.

we used those tuned parameters to retrain the model and got better accuracy and finally used PCA with tuned parameters to retrain the model.

Regularization in SVM :

Regularization parameter lambda is used in SVM for two main reasons. To overfitting and underfitting.

SVM categorizes multidimensional data with the goal of fitting training data well but sometimes the testing data fails to generalize the accuracy of the training data that's when algorithm overfills. To avoid overfitting and underfitting we used lambda to penalize the equation. With penalty 'l2' as a regularization parameter,we again retrained the model and got better results.

3. Random Forest

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks.[13]

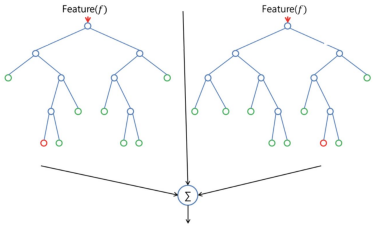

Figure 3. Random Forest General Structure (https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

In our case, we used Random Forest for classification. Even though Random forest doesn't require much parameter tuning, we applied some extent of tuning as our base Random Classifier wasn't giving us the desired accuracy.

4. Gaussian Processing Classification:

Different from many other methods, it is a nonparametric classification method, and the final classification is determined as the one that provides a good fit for the observed data, while at the same time guaranteeing smoothness.

It focuses on modeling the posterior probabilities, by defining certain latent variables f_i : is the latent variable for pattern i.[14]

The posterior probability for a class is:

$$P(C|x_i) = P(y_i = 1|f_i)$$
$$= \int_{-\infty}^{f_i} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

Equation 1. The posterior probability for a class y = 1in our 3 classes problem. fi is the latent variable.

The Gaussian Preprocessing Classification is using the prior probabilities to estimate a good estimation for posterior probability.

5. Artificial Neural Network

Since we have 2158 features for prediction, a deep neural network can have great advantages on dealing this high dimensional data.

We use a 4-layer fully-connected neural network, where the inputs are our 639 records and 2158 features. The output is the classification result of the 3 classes. The figure shown below is our 4 layer ANN.

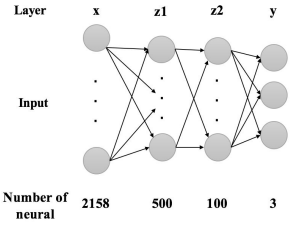

Figure 4. Example structure of artificial neural network. The input is our 2158 features samples, number of neural will decrease as the process goes. Final get 3 neural output as class probability.(Image taken from WPI Slides)

the network we are using implement the function f : $R^{2158} \rightarrow R^3$. For every layer, $z_i = w_i x + b_i$, where x is the output of the previous layer and b is the bias. We choose the ReLU function as the activate function before the last output, and use softmax function to get the final classification probability result. The cost function is cross entropy:

$$J(W,b) = -\frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{3} y_k^{(j)} log\hat{y}_k^{(j)}$$

Equation 2. The cross entropy for our 3 class classification problem. Sum of *m* samples and 3 classes.

Hyperparameter tuning is an important part in the neural network, we tried different numbers of units in the hidden layer, learning rate, mini-batch size and regulation strength.

C. Ensemble Learning :

1.Bagging Model:

Bagging is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. It is a special case of the model averaging approach. Bagging is to generate several new sets by sampling samples or features or both uniformly and with replacement, then train the model with these different datasets, then using max-voting to generate the final result.[15]

We are using bagging in both samples and features, which will resample 600 records and 500 features to generate new data set, to feed in the model. Separately, we will use bagging on all basic model we mentioned above.

2.Ensemble Model:

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Feeding the data into several different models, and using max-voting to generate the final result.

We are using ensemble model which is several combinations of basic model we mentioned above, including model ensembled with Logistic regression+Decision Tree+SVM, Logistic regression+Decision Tree+SVM+GPC and Logistic regression+Decision Tree+GPC.

3.Boosting Model:

Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones.

We are using AdaBoost which selects only those features known to improve the predictive power of the model, reducing dimensionality and potentially improving execution time as irrelevant features need not be computed.[16]

D. Convolutional Neural Network :

For the original image data, we choose CNN as our method. Similar to the deep neural network, it is a network combined by 2 convolutional layers and following by 2 fully connected layers. Different from the processed data, the original data is 3-D image with many slices for each patient, which mean the input of our network has many channels with 256 pixels on width and height.

Limited by the computational power, we resize each image to 40*40 and only use 20 slices in the middle of 3-D image, which makes the input of our network to 40*40*20.

We also choose ReLU and softmax as our activation function, and we additionally add the max pooling after each convolutional layer and dropout after the fully-connected layer. Max pooling reduces the number of the parameters and leaves the most important information; Dropout will randomly choose a part of neural network for training, which can reduce overfitting. Finally, we add the fully connected layer at the end with softmax function to get the classification probability result.

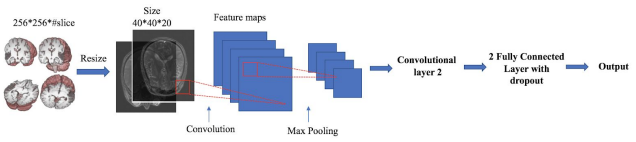

Figure 5. The structure of image preprocessing and convolutional neural network. the process include resize the image to small scale, 2 convolutional layers and 2 fully connected layer, and the output is the possibility of classified into each class. (http://nipy.org/nibabel/neuro_radio_conventions.html)

In the training process, we are using mini-batch gradient descent approach of size 32 to reduce the computational pressure. Mini-batch gradient descent will update the weight matrix only using a part of the whole dataset, and will eventually converge similar to the result of normal gradient descent.

E. Description Of Assessment Protocol :

We evaluate our advanced models by comparing the result with the basic model. Since correctly detecting Alzheimer is the most important target for our models, we evaluate the performance using accuracy, high accuracy means the better model.

The data we are using is not very balanced with 20% AD, 30%, CN and 50% MCI labels, the accuracy metrics sometimes can mislead the performance, so we also measure our model with f1-score, which will measure the performance both on the accuracy and the balance to avoid predicting all samples to the dominating class.

F-1 Score is function of Precision and Recall.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Equation 3. F-1 Score function

IV. Results

The following tables present a comparative analysis of the various machine learning classifiers :

TABLE 1: Performance of the classifiers without PCA.

Accuracy score and f1 score for each classifier was measured on the same dataset without PCA

Classifier	Performance metric	
	Accuracy Score	**F1 score**
Logistic Regression (LR)	0.4921	0.4829
Support Vector Machines (SVM)	0.5556	0.5174
Random Forests (RF)	0.6587	0.5916
Gaussian Process Classification (GPC)	0.5397	0.2337
Artificial Neural Networks	0.4841	0.4559
Bagging - Logistic Regression	0.4762	0.4682
Bagging - SVM	0.5159	0.3016
Ensemble (LR + RF + SVM)	0.5159	0.4670
Ensemble (LR + RF + SVM + GPC)	0.5625	0.5137
Ensemble (LR + RF + GPC)	0.4932	0.4873
Ada boosting - Decision trees	0.5793	0.5909
Ada boosting - LR	0.5079	0.5104
Gradient Boosting	0.6904	0.6897

TABLE 2: Performance of the ensemble machine learning algorithms with PCA.

Accuracy score and f1 score for classifier was measured on the same dataset after PCA

Classifier	Performance metric	
	Accuracy Score	**F1 score**
Logistic Regression	0.3917	0.4206
Support Vector Machines	0.4683	0.4297
Random Forests	0.4444	0.3223
Gaussian Process Classification	0.4286	0.1999
Bagging - Logistic Regression	0.3968	0.3348
Bagging - SVM	0.5397	0.2337
Ensemble (LR + RF + SVM)	0.4891	0.4732
Ensemble (LR + RF + SVM + GPC)	0.5213	0.4975
Ensemble (LR + RF + GPC)	0.4610	0.4533
Ada boosting Decision Trees	0.4365	0.4268
Ada boosting - LR	0.4841	0.4855
Gradient Boosting	0.4523	0.4141

Additionally:

1. Mini Mental State Examination (MMSE) was found to be the most important which agrees with the current diagnostic standards
2. However, education was not to be found to be in the top 40 features and had negligible importance compared to other more important features which is a new finding contradicting what is commonly assumed in the industry
3. PCA found out 390 features among 2160 features to have 98% of the variance.But using PCA has dropped the performance of most methods.
4. Important results of Convolutional Neural Networks:

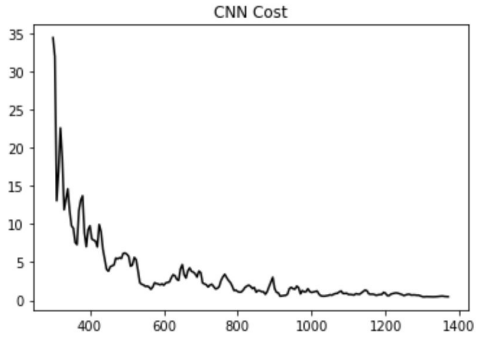

Figure 6. Convolutional Neural Network Training Cost vs the number of iterations. The cost is decreasing with iterations.

We implemented mini-batch Convolutional Neural Network on the original brain image data. The figure 6 shows the cost is continuously decreasing from 25000 to lower than 1. The figure shows on the right that the cost is nearly converged.

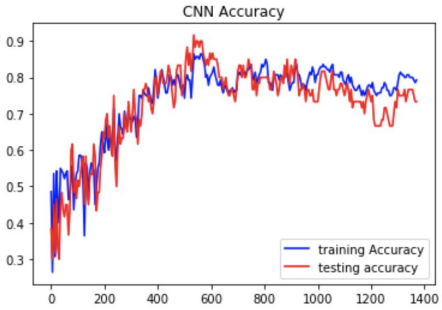

Figure 7. Convolutional Neural Network Training and Testing Accuracy vs number of iterations. the Accuracy is stable at 75%.

The classification accuracy is generally increasing on both training set and test set, with the increasing of iteration. The accuracy for training is quite stable around 75% and f1-score is around 72%; due to the limited number of iteration, the testing accuracy is fluctuating from 70% to 83%. More training iteration will increase stability and accuracy.

In Summary the following six important results were obtained. First being that the highest accuracy for classical machine learning obtained by the Gradient Boosting. Secondly, Performing dimensionality reduction using PCA deteriorates the performance of all the models. Third being that Ada boosting is found to not give any substantial improvements in performance but Random Forest is found to work well on this dataset individually. Although all ensemble models include random forest which gives the highest individual accuracy, the result of ensemble models is much lower than RF. Also, using original brain image data, we generate the convolutional neural network classification model with the accuracy around 75%.

V. Conclusion

We have proposed an approach to detect Alzheimer's using Brain MRI scans and features extracted from Brain MRI scans. We implemented several individual machine learning algorithms, ensemble machine learning algorithms and a convolutional neural network and found out that Gradient Boosting gives the highest accuracy in classifying the patients in Machine Learning.

PCA found features which have the most variance and helped in dimensionality reduction however since it consistently deteriorated the performance of all models, we conclude that it was not able to capture the important discriminative features.

Ensemble models with different types of classifiers it reduced the overall performance. This indicates that bad models can potentially deteriorate the performance of a high model in the ensemble model and combining different types of classifiers is not always a good approach.

CNN improved over the accuracy by dealing with raw images which indicates the strength of CNN and probably shows the disadvantages of image processing method that can lose useful features compared to using original image.

We also observed that MMSE is found out to be the most important feature agreeing with current industry practices. However our models found that education is not as important as the other morphometric data features, which is unlike what is currently perceived by the industry.

In the future, we will try to make ensemble model which would use weighted voting instead of max-voting. Also, we would build models which would extract features from raw images and are not bound to the type of processing done on the images and build a deep learning model which would use the images in its entirety.

Acknowledgment

We would like to extend our gratitude to Prof. Korkin and Prof. Nephew from WPI for their support and guidance

References

[1]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5494120/

[2]https://www.alzwisc.org/Importance%20of%20an%20early%20diagnosis.htm

[3] https://www.alzheimers.net/resources/alzheimers-statistics/

[4] Kloeppel S et al (2008) Automatic classification of MR scans in Alzheimer's disease. Brain 131:681–689

[5] https://www.synapse.org/#!Synapse:syn2527678/wiki/69937

[6] https://www.synapse.org/#!Synapse:syn2775285/wiki/70018

[7] https://www.synapse.org/#!Synapse:syn2631754/wiki/69912

[8] http://adni.loni.usc.edu/data-samples/access-data/

[9] https://scikit-learn.org/stable/

[10]https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html

[11]https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f 0812effc72

[12] https://scikit-learn.org/stable/modules/grid_search.html

[13]https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[14] Gaussian Processes for Classification, Amir Atiya Dept Computer Engineering, Cairo University

[15] https://en.wikipedia.org/wiki/Bootstrap_aggregating

[16] https://en.wikipedia.org/wiki/Boosting_(machine_learning)
[17]https://www.alz.org/media/Documents/alzheimers-facts-and-figures-infographic.pdf:
[18] Siqi Liu et al (2014) Early diagnosis of Alzheimer's disease with deep learning
[19] Ortiz et al Ensembles of Deep Learning Architectures for the Early Diagnosis of the Alzheimer's Disease