**Alzheimer’s Disease Classification using ML Pipeline on**

**Fast Fourier Transformed EEG Data**

Jacob Brawer

In collaboration with Tyler Yoshihara

Pomona College

Machine Learning w/ Neural Signals (NEUR182)

Professor Spezio

**Abstract:**

Alzheimer’s Disease (AD) is the most common neurodegenerative disease and sixth leading cause of death in the US (Alzheimer's Association, 2019). EEG screening has shown potential as a noninvasive diagnostic tool for AD.  Machine learning allows us to make inferences about changes in frequency bands within EEG data and how these changes relate to neural code. The EEG data obtained for this project was from a 2014 paper written by Fiscon et al. There were patients with AD, mild cognitive impairment (MCI), and healthy controls (HC). The data was already preprocessed using a fast fourier transform (FFT). After running four classification techniques, relevance vector classifier (RVC), random forest classifier (RF), and Fisher’s discriminant analysis (FDA) were the most successful in classifying AD vs. HC and MCI vs. AD. Due to inconsistent feature importances across models, conclusions about important frequency bands for classification were not able to be made with certainty. Similarly, different frequencies were not able to be localized to different regions of the brain. Further research is necessary to develop more interpretable models for classification.

**Introduction:**

Alzheimer’s disease is an irreversible, progressive neurodegenerative disorder marked by memory issues associated with dementia, language problems, and erratic behavior. Currently, 1 in 3 seniors die with either Alzheimer’s or a related dementia, yet only 16% of seniors receive regular cognitive assessments. By 2015 it is projected that 14 million Americans will be diagnosed with Alzheimer’s disease. The biological manifestation of the disease is characterized by beta amyloid plaques and tau protein tangles in the brain and progressive cognitive decline. Most current diagnostic techniques center around the identification of biomarkers-- amyloid-βeta peptide 1-42, total tau protein and hyperphosphorylated tau at threonine 181 have all be identified as certified markers of the disease in human cerebrospinal fluid (CSF) (Robinson et al., 2017).  While biomarkers are successful marker of a patient’s proclivity to developing Alzheimer’s, inaccuracy is a problem that plagues a lot of these approaches (Tarnanas, Tsolaki, Wiederhold, Wiederhold, & Tsolaki, 2015). Additionally, the development of these biomarkers is a sign of later stages of the disease, which makes mitigation of the symptomatology of the disease limited. Invasive, posthumous brain examination is the only guaranteed diagnostic tool at this point in time. As a result, the application of machine learning to EEG data could be an extremely useful and cost-effective method to non-invasively screen for Alzheimer’s disease and could potentially serve as an incredibly beneficial and life-saving medical protocol (Ding et al., 2018). Often Alzheimer’s symptoms are associated with later stages of the disease, which is why early screening could allow for intervention prior to neurodegeneration running its course. The classification of Alzheimer’s disease necessitates cross-subject analysis, as intra-subject classification would not provide any useable results.

The EEG data used for the body of this study comes from a 2018 study by Fiscon et al., utilizing classification techniques derived in an earlier 2014 study by the same group (Fiscon et al., 2014). Our hope in expanding upon Fiscon et al. 2018’s work is to improve upon their classification techniques and provide interpretable results that can be localized to specific frequency bands and channels for diagnostic purposes (Fiscon et al., 2018).  Fiscon et al. employed decision trees (DT), support vector machines (SVC) and rule-based classifiers, ultimately citing DTs as their most effective algorithm in their 2018 paper. In selecting our models, the primary goal was to maximize sensitivity and specificity, while retaining parsimony and being highly cognizant of overfitting and decreases in accuracy due to features derived from noise that is not consistent across cases (Hebart & Baker, 2018). Our classification algorithms include Relevance Vector Classifier (RVC), Ridge Regularized Linear Regression (RLR – L2), Fisher’s Discriminant Analysis (FDA), and Random Forest (RF) (Liaw et al., 2002). Four primary EEG frequency bands were examined for salience in our interpretation of feature importance: delta (0.5-4 Hz), theta (4-7 Hz ), alpha (8-13 Hz), beta (13-30 Hz). It was hypothesized that alpha bands would be downregulated across channels, while beta and delta frequency would increase for AD patients relative to healthy control. MCI patients were expected to have less degradation of alpha frequency compared to AD subjects.

**Existing Data Set:**

The existing data was taken from Fiscon et al. 2018’s study—It is pre-processed EEG recording from 86 participants with Alzheimer’s Disease (AD) or Mild Cognitive Impairment (MCI) and 23 Healthy Controls (HC). The data was derived from a resting-eyes closed (EC) state for 300 seconds, with the central three-minutes of data (60 sec to 240 sec) being the focus of analysis. A 19 electrode array with a sampling rate of 256 Hz was used. Fiscon et al. preprocessed the data utilizing two techniques—Daubechies Wavelets and Fast Fourier Transform (FFT). However, only FFT will be the focus of this follow-up study. FFT was conducted on the central-three minutes of data in order to transform the signal from the time domain to the frequency domain. The three minutes of data was divided into six, 30 second epochs. 16 Fourier coefficients representing difference frequency ranges were extracted from each epoch, which was presented to us in the form of csv files with cases represented and labeled on the y-axis and features in the form of fourier coefficients on the x-axis (304 columns). Three csv files were produced, one for the AD vs. HC grouping, one for the AD vs. MCI grouping, and one for the MCI vs. HC grouping. Of the 16 features for each electrode, the first four correspond to the Delta band (0-4Hz), 5-8 correspond to the Theta band (4-7Hz), 9-12 correspond to the Alpha band (8-12Hz), and 13-16 correspond to the Beta band (13-30 Hz).

**Methods:**

Since the data was already pre-processed the natural first step was to look at the existing methodology that Fiscon et al. used in their 2018 paper and examine potential ways we could improve their results with our own classification techniques. The authors utilized support vector machines (SVM) and decision-based classifiers, before deciding on decision trees as their best classification technique. Fiscon et al. used k-fold cross-validation with the “leave one out” strategy. Despite generally successful performance in classification, one area in which Fiscon et al.’s decision trees did not perform was in terms of specificity within the AD vs. HC classification. This result is indicative of a model that errs in the direction of false negatives (AD diagnosis in cases of HC).

Four different classification techniques were chosen for the data. First, we wanted to use a more parsimonious kernel-based classifier, as Fiscon et al. observed overfitting in their SVM implementation, a result that they attributed the model being susceptible to nonstationary noise in the data. As a result, we chose a relevance vector classifier as one of our models, a Bayesian kernel method, which creates a separatory hyperplane using probabilistic measure in order to maximize the minimum distance between the data points of different classes (Bishop, 2006). In choosing a kernel method with less support vectors than an SVM, we hoped that our RVC model would demonstrate more robustness (Figure 1). Next, we chose Random Forest classification, which we believed would build on the success of the decision tree models used by Fiscon et al., a method that they found best handed the noise and nonstationarity of the data. Random Forests are an ensemble method that builds and combines a multitude of decision trees, utilizing bagging (bootstrap aggregation) and feature bagging (random ensemble method) in order to avoid overfitting during the training step associated with decision trees (Figure 2) (Liaw & Wiener, 2001).

The final two techniques were linear classification methods, both of which attempt to only weight features that have salient signal as opposed to noise. The first of which was Fisher’s Linear Discriminant Analysis (FDA) (Figure 3), utilizing single value decomposition in order to transform feature vectors that maximize separability. The second was regularized (l2) linear regression (RLR), a technique with a linear basis function that also weights a maximum penalized likelihood estimation, mitigating the “ridge” that occurs under features that demonstrate multicollinearity (Bishop, 2006).

We first centered and scaled our data sets and divided them into training (80%), validation (8%), and testing sets (12%) (for RLR) and training (80%) and testing (20%) sets for the other three classifiers using the shuffle split function. We also utilized the imblearn function, SMOTE, in order to control for imbalanced sampling pools (much greater number of AD patients than HC subjects), using k-nearest neighbors to equalize training, validation and test sets. Before running any classification, we examined the multicollinearity of the data to determine if we were going to utilize principal component analysis (PCA) prior to FDA and RLR classification. We discovered with the use numpy’s corrcoef function that there was very little multicollinearity within any of our data sets, so we decided not to continue with the PCA analysis. We then trained hyperparameters for RLR and RF with a gridsearch in order to achieve optimization. Our resulting models were analyzed using four different measures of accuracy (regular accuracy, ROC (AUC), sensitivity and specificity). In order to interpret our feature importance information, we had associate our given features with corresponding frequency bands. This involved using Welch’s method to get the power spectral density from the fourier coefficients. Next, took the power spectral density for all of the 30 second epochs, collapsing the data into one, easily interpretable importance vector. Following feature importance calculation for all four of our models, we examined cross-model correlation of feature importance, frequency-band analysis, and spatial analysis by frequency and channel.

**Results:**

Atop the hierarchy of importance for diagnosis is the Alzheimer’s (AD) vs. Healthy Control (HC) classification. After running all four of our models, the two most effective classifiers were the relevance vector classifier and the random forest classifier. Evaluation metrics for RVC demonstrated high sensitivity (0.9) (good classification of true HC), but was slightly less impactful in terms of specificity (0.6) (poorer identification of true AD cases). The AUC for the generated ROC curves was 0.8 for AD and HC cases, with an overall accuracy of 75% (Figure 4). The random forest classifier boasted a perfect measure of specificity on the test set (1.0) (classification of true AD cases), but had lesser sensitivity (0.4) (poorer identification of true HC). The AUC values were 0.88 for both AD and HC cases, with a reported accuracy of 70% altogether (Figure 5).

A secondarily important diagnostic criterion is distinguishing between cases of mild cognitive impairment (MCI) and Alzheimer’s (AD). Our two most robust models for this classification step were relevance vector classifiers and fisher’s linear discriminant classifier. RVC demonstrated a high specificity (0.81) (correctly identifying AD), but much less sensitivity (0.55) (correctly classifying MCI). The AUC value for the ROC curve was 0.70 for both AD and MCI, with an overall accuracy of 68% in classification of the test set (Figure 6). FDA did not perform nearly as well for specificity (0.64) (poorer identification of true AD cases), but had identical sensitivity (0.55) (distinguishing true MCI patients) to the RVC model. The AUC value was 0.65 for AD and MCI, and an accuracy of 59% (Figure 7).

While less critical in terms of diagnostic importance, we also found that random forests and relevance vector classifiers were the best models for the classification of mild cognitive impairment (MCI) vs. healthy controls (HC). RF had a respectable specificity (0.8) and solid sensitivity as well (0.71), with AUC values of 0.6 both for HC and MCI and accuracy of 75%. RVC has a similar specificity (0.8), but much poorer sensitivity performance (.43) (classifying true HC correctly). AUC values for HC and MCI were 0.66, while accuracy was 58% overall (Figure 8).

In further analyzing feature importance, we utilized numpy’s corrcoef to determine the cross-model feature importance correlation for our top two models within each respective classification task (AD vs. HC, AD vs. MCI, MCI vs. HC). We did not find any significant correlational relationship feature importances across all classification tasks. However, we were able to graphically and spatially represent activation across channels and frequencies using heat signature maps and the given feature importances for each model. Notably, across RVC and RF, in the AD vs. HC classification, the T3 channel in the delta frequency and the O2 channel in the theta frequency both showed increased importance and activation relative to other channels (Figure 9). Additionally, in MCI vs. AD classification the FP2 channel in the delta frequency and the F4 channel in the theta frequency both demonstrated increased activation and importance compared to other channels in the same frequency bands (Figure 10). One surprising but important result that the heat maps allowed us to visualize was overfitting in our RVC model, evident by lots of orange and red regions across the electrodes, compared to the RF model, which had much more localized feature importance (Figure 9). This was unexpected, given the traditionally parsimonious nature of RVC. We had a similar observation for RF in the MCI vs. HC classification, compared to FDA which had more noticeable locality (Figure 10). Lastly, we graphed feature importance vs. frequency (averaged across channels) within each of our models (Figure 11, 12). Using confidence intervals, we determined that there were no individually, significantly important frequencies in classification any of our models in AD vs. HC classification and MCI vs. AD classification. However, consistent with the Fiscon et al. 2018 frequencies within the alpha and beta range showed relatively (albeit insignificant) feature importance relative to other bands for the AD vs. HC classification (Figure 11).

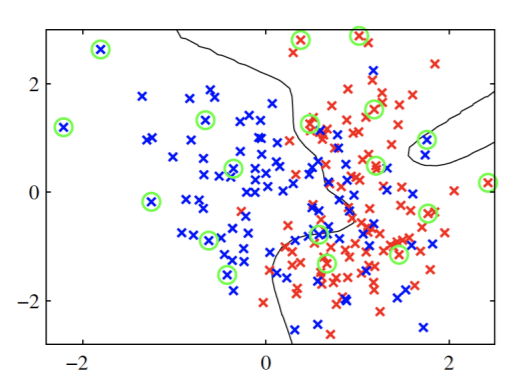
**Conclusion:**

Within the AD vs. HC classification task, RVC was the most successful model according to our evaluation metrics, while RF performed the best for MCI vs. AD and HC vs. MCI. Unfortunately, our results do not support any strong conclusions about the neural underpinnings of Alzheimer’s Disease. Correlations between feature importance for different algorithms within classification tasks between models was quite low. Additionally, channel activation and frequency band patterning was inconsistent across models and classification tasks, evident in the heat maps and frequency by feature importance graphics (Figures 9-12). While, we did notice patterns of activation in the delta and theta bands within the AD vs. HC classification, they were not strongly correlated and the feature importance related to these specific channels did not show significantly different importance compared to the grouping of features allotted for each classification. Despite the fact that the data sets were not multicollinear, PCA could be a beneficial technique in future work using this data set in order to generate features that are more interpretable. If we were receiving raw data, instead of preprocessed data, using a different form of preprocessing step such as a wavelet transform, could be beneficial in improving our classification results. Fiscon et al. noted in their results that their Daubechies Wavelet Transform classification results were significantly better than the FFT results, due to the transformation techniques ability to handle noise and nonstationarity in signal (Fiscon et al., 2016). These additional techniques could aid in finding consistent feature importance across different classification models, which is not a result that we witnessed in our work. Future research is necessary in order to draw more interpretable conclusions pertaining to locality and frequency band specificity when it comes to Alzheimer’s classification.

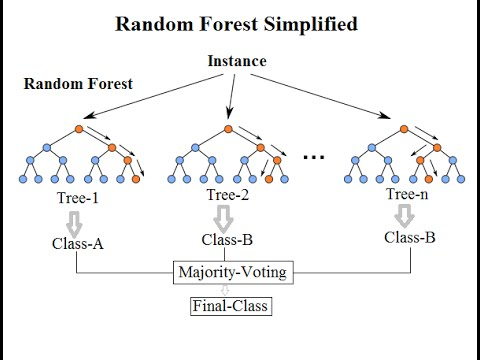
**Acknowledgements:**

Thank you to Professor Spezio for guidance in this project and thank you to Fiscon et al. for providing the preprocessed data and original study that inspired this work. Lastly, thank you to Isabelle and Fernanda for collaborating with us in the early stages of this project.

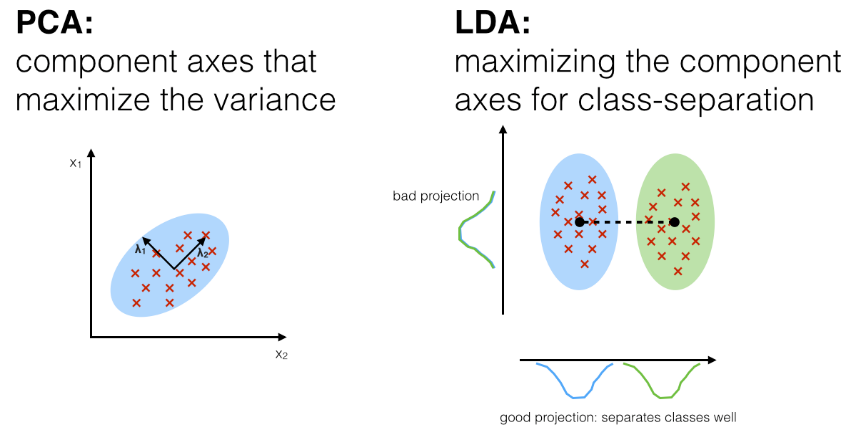
**Figures:**



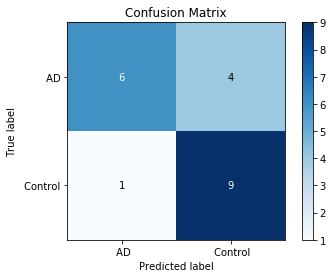
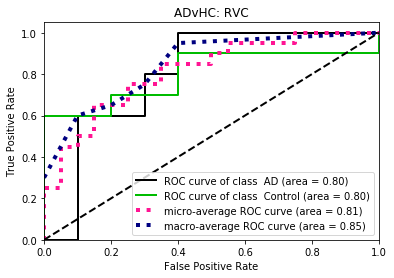
**Figure 1 : Relevance Vector Classifier** – pictured withdrawn decision boundary and circled relevance vectors



**Figure 2 : Random Forest Classifier** - Disjoint set of decision trees that are combined to form an ensemble method for classification

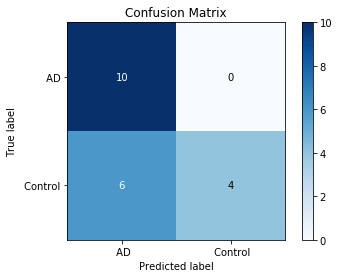
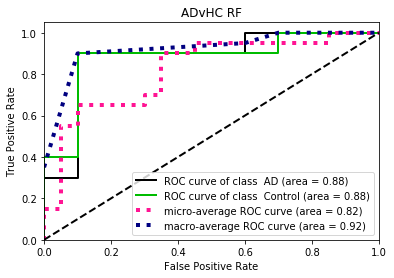


**Figure 3: Principle Components Analysis and LDA –** left graph is demonstrative of PCA grouping, while right figure demonstrates decision boundary for Fisher’s Discriminant Analysis



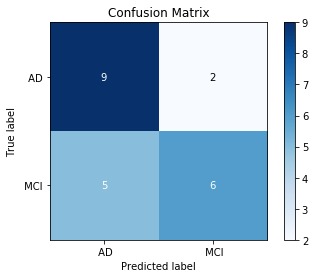
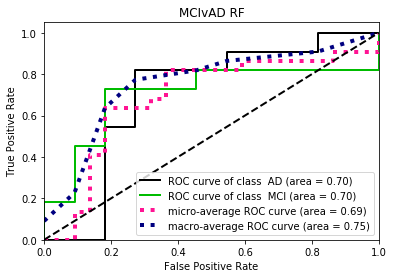
|  |  |
| --- | --- |
| ***Classification Measure*** | **Performance** |
| *Specificity* | 0.6 |
| *Sensitivity* | 0.9 |
| *AUC* | .80 (AD) - .80 (HC) |
| *Accuracy* | 75% |

**Figures 4a,b,c: RVC Evaluation Metrics for AD vs. HC Classification** – ROC Curve, Confusion Matrix, Metrics Table



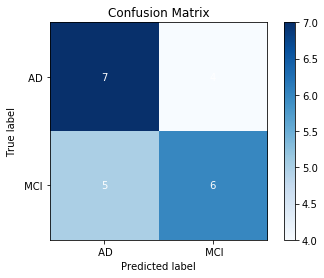
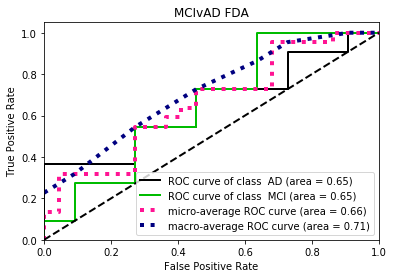
|  |  |
| --- | --- |
| ***Classification Measure*** | **Performance** |
| *Specificity* | 1.0 |
| *Sensitivity* | 0.4 |
| *AUC* | .88 (AD) - .88 (HC) |
| *Accuracy* | 70% |

**Figures 5a,b,c: RF Evaluation Metrics for AD vs. HC Classification** – ROC Curve, Confusion Matrix, Metrics Table



|  |  |
| --- | --- |
| ***Classification Measure*** | **Performance** |
| *Specificity* | 0.81 |
| *Sensitivity* | 0.55 |
| *AUC* | .70 (AD) - .70 (MCI) |
| *Accuracy* | 68% |

**Figure 6: RF Evaluation Metrics for MCI vs. AD Classification** – ROC Curve, Confusion Matrix, Metrics Table



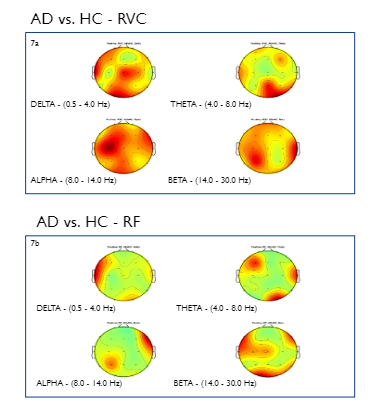
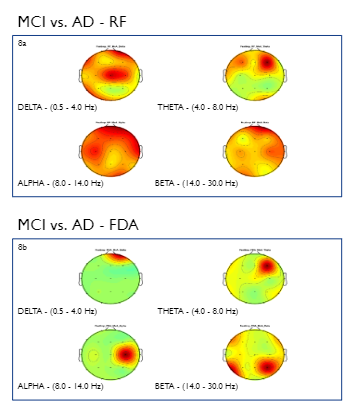
|  |  |
| --- | --- |
| ***Classification Measure*** | **Performance** |
| *Specificity* | 0.64 |
| *Sensitivity* | 0.55 |
| *AUC* | .65 (AD) - .65 (HC) |
| *Accuracy* | 59% |

**Figure 7: FDA Evaluation Metrics for MCI vs. AD Classification** – ROC Curve, Confusion Matrix, Metrics Table

|  |  |
| --- | --- |
| ***Classification Measure*** | **RF - Performance** |
| *Specificity* | 0.8 |
| *Sensitivity* | 0.71 |
| *AUC* | .60 (HC) - .60 (MCI) |
| *Accuracy* | 75% |

|  |  |
| --- | --- |
| ***Classification Measure*** | **RVC - Performance** |
| *Specificity* | 0.8 |
| *Sensitivity* | 0.43 |
| *AUC* | .66 (HC) - .66(MCI) |
| *Accuracy* | 58% |

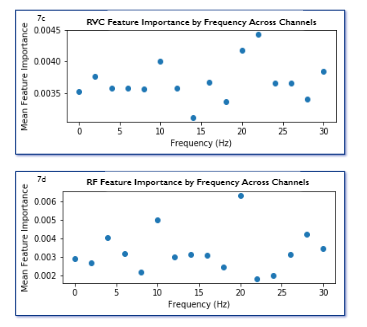
**Figure 8: RF & RVC Evaluation Metrics for MCI vs. HC Classification** – no inclusion of ROC curves or confusion matrix because this is a secondary objective to the project and therefore less important.



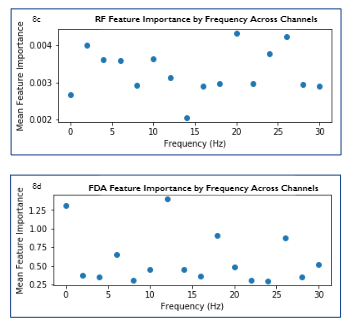
**Figure 9: AD vs. HC Heat Map for RVC & RF** – **Figure 10: MCI vs. AD Heat Map for RVC & RF** -

O2 (Theta) and T3 (Delta) activation across F4 (Theta) and FP2 (Delta) activation across

models models



**Figure 11: AD vs. HC Feature Importance by Frequency Across Channels for RVC and RF** - Relative feature importances did not demonstrate any consistency across channels and frequencies



**Figure 12: MCI vs. AD Feature Importance by Frequency Across Channels for FDA and RF** - Relative feature importances did not demonstrate any consistency across channels and frequencies

**Works Cited:**

Anon (n.d.) Facts and Figures. Alzheimers Dis Dement Available at: https://alz.org/alzheimers-dementia/facts-figures [Accessed May 16, 2019].

Robinson RAS, Amin B, Guest PC (2017) Multiplexing Biomarker Methods, Proteomics and Considerations for Alzheimer’s Disease. Adv Exp Med Biol 974:21–48.

Fiscon, G., Weitschek, E., Cialini, A., Felici, G., Bertolazzi, P., De Salvo, S., … De Cola, M. C. (2018). Combining EEG signal processing with supervised methods for Alzheimer’s patients classification. *BMC Medical Informatics and Decision Making*, *18*(1), 35. https://doi.org/10.1186/s12911-018-0613-y

Fiscon, G., Weitschek, E., Felici, G., Bertolazzi, P., De Salvo, S., Bramanti, P., & De Cola, M. C. (2014, December 10). *Alzheimer’s disease patients classification through EEG signals processing*. https://doi.org/10.1109/CIDM.2014.7008655

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, *180*, 4–18. https://doi.org/10.1016/j.neuroimage.2017.08.005

Liaw, A., & Wiener, M. (2001). Classification and Regression by RandomForest. *Forest*, *23*.

Tarnanas, I., Tsolaki, A., Wiederhold, M., Wiederhold, B., & Tsolaki, M. (2015). Five-year biomarker progression variability for Alzheimer’s disease dementia prediction: Can a complex instrumental activities of daily living marker fill in the gaps? *Alzheimer’s & Dementia : Diagnosis, Assessment & Disease Monitoring*, *1*(4), 521–532. <https://doi.org/10.1016/j.dadm.2015.10.005>