Longitudinal Analysis of Alzheimer's Disease Features to Aid in Forecasting and Prediction

Siddarth Satuluru (404838847), Saachi Kudtarkar (604745507), Pratyusha Majumder (404797253)

I. Introduction

An estimated 44 million people worldwide are living with Alzheimer's disease (AD), including approximately 5.5 million people in the United States, where it is currently the sixth leading cause of death [1]. Alzheimer's disease is caused by tissue abnormalities called amyloid plaques and neurofibrillary tangles. The amyloid plaques are abnormal deposits of beta-amyloid proteins, which is hypothesized to be the earliest signs of the disease process. This protein can accumulate between neurons and disrupts cell function. Neurofibrillary tangles are abnormal clumps of tau proteins that also contribute to Alzheimer's disease. The tau proteins in a normal brain bind to and stabilize microtubules, but in an Alzheimer's patient, these proteins detach from microtubules and clump together, eventually forming tangles inside the neurons and blocking synaptic communication between neurons [2].

The beta-amyloid proteins and tau tangles are interconnected; after a certain tipping point of beta-amyloid proteins, there is a spread of tau tangles in the brain. Alzheimer's disease is also caused by a faulty blood-brain barrier. This prevents the abnormal clumps of beta-amyloid and tau proteins being cleared away and also prevents glucose from reaching the neurons in the brain. A combination of these factors disrupts the communication between neurons and breaks down the connections in neuronal networks over time, first affecting the entorhinal cortex and hippocampus (regions responsible for memory), then the cerebral cortex (region responsible for language, reasoning and social behavior), before it continues to spread to other areas of the brain. As the regions begin to shrink (atrophy), there is a loss of brain volume, resulting in cognitive decline [2]. This is monitored by various neurophysiological tests such as CDRSB, MMSE, ADAS11, ADAS13, Ecog and RAVLT.

Research currently exists in detecting biomarkers from Alzheimer's patients, both through the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and through external analysis. A current deep learning approach exists using f-18 PET scans that classifies scans with Alzheimer's disease based on a saliency map [3]. The analysis however does not use human-interpretable signs, using a whole image analysis instead. Further blood based analysis has been performed to screen for early indication of AD with studies showing evidence for the nerve growth factor precursor protein and amyloid imaging markers as biomarkers for AD. It is also shown that variation in structural MRI markers cannot always be directly attributed to AD [4][5]. Salivary metabolomics biomarkers in conjunction with other risk markers such as functional health have been shown to be a robust method of diagnosis, however, as the other risk markers are not always present/quantifiable, we believe a quantitative analysis would be beneficial [6].

The Alzheimer's Disease Neuroimaging Initiative (ADNI) has collected large amounts of data related to the progression of Alzheimer's disease in elderly people, classified as being of age 65 or older [7]. The institution began data collection in 2004, and has since managed to collect data on individuals without Alzheimer's, individuals with mild memory impairment, and individuals with Alzheimer's or dementia.

The data archive consists of data of five types: clinical, genetic, MRI image, PET image, and biospecimen. The clinical data includes information such as demographics and physical and cognitive assessments. The genetic data includes genotyping and sequencing data, and the biospecimen data includes information such as blood and urine samples from participants. It's also important to note that for the vast majority of data, including the MRI and PET image data, ADNI conducted periodic data collection on participants every 3 months, starting with an initial screening and ending at month 48 or beyond.

We've chosen to use PET scans and MRI images, alongside other markers available to us, to determine the progression of Alzheimer's in patients. We believe MRI and PET scans can help show the progression of Alzheimer's Disease (AD), as ADNI houses data on patients that has been collected every 3 months for at least 2 years. Our project hopes to find the correlation of biomarkers, either genomic or imaging-based, and the progression of AD, if any exists, to help physicians determine subsequent stages of the disease.

While previous research has focused on detecting the presence of Alzheimer's Disease in patients, this project will aim to track the longitudinal progression of Alzheimer's Disease with the use of genomic and imaging data. Our goals align with The Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE) challenge. The TADPOLE Challenge aimed to forecast AD indicators for the short to medium term (1-5 years) using the ADNI database [8]. Participants are asked to find out if at-risk individuals' progression to AD can be predicted, the data, processing pipelines and predictive models that best predict AD progression, and if these methods can improve cohort selection for clinical trials.

Following the TADPOLE Challenge's guidelines, we further refined our research goals to answer the following questions:

Which biomarkers are the most important in determining clinical status? How accurately can we classify a patient's clinical status, given specific features? How accurately can we forecast the progression of AD in patients? Our focus on addressing the progression of Alzheimer's Disease in this way coincides with the natural progression of Alzheimer's Disease in patients. As Alzheimer's Disease is a neurodegenerative disease that can't be cured, we aim to determine the progression of the disease at an early stage to be able to administer specialized treatment as soon as possible.

II. Methods

Our goal was to track the progression of the top biomarkers that contribute to a patient's diagnosis and see if it corresponds to the actual progression that a given patient experiences. This consisted of a few steps: preprocessing the data, selecting key features, building and selecting a regression model, tracking the progression of the 'average' patient, grouped by baseline diagnosis, and testing the progression by evaluating the results using the original classification model.

The data used was provided by the organizers of the TADPOLE challenge. This data is a standardized subset of the ADNI database that has been pre-processed. In particular, we used the "D1_D2.csv" file, which contained data from both D1 and D2. D1 is a comprehensive longitudinal data set intended for training, and D2 is a comprehensive longitudinal data set intended for forecasting. The data included preprocessed image data that extracted key pieces of information from the ADNI PET and MRI images, such as measures of the amyloid beta protein using cerebrospinal fluid puncture or amyloid positron emission tomography or atrophy. The data was organized based on the patient's de-identified ID, which enabled easier additional preprocessing and analysis. It also included the examination date and months from the baseline examination, which allowed for great ease when conducting the longitudinal study of patients.

A. Preprocessing data

In order to build the regression models with selected features, we needed to first preprocess the data. The original dataset had approximately 12,000 entries and 2,000 columns for each of the patients tracked in the ADNI database. This included demographic data, results from cognitive assessments, and extracted features from MRI and PET scans. The extracted features were, in general, aggregated results (volume, surface area, etc.) from regions of interest (ROI) in the brain. Missing data arose from ROI algorithms that were only applied to certain subsets of the patient population. To account for this, the dataset was padded using forward fill. This was done per patient in order to ensure that any missing data did not come from a different patient. All exams were not conducted for every date of visit for some patients. To account for missing data on any exam date, the field was filled with value 0.

The following categories were given numerical representations: PTGENDER (gender), PTETHCAT (ethnicity), PTRACCAT (race), PTMARRY (marital status) and DX (diagnosis). The values assigned for DX were: 1 for CN (control) ; 2 for CN to MCI (mild cognitive impairment) or MCI to CN ; 3 for MCI ; 4 for MCI to Dementia or Dementia to MCI ; 5 for Dementia. The added DX_NUMERICAL column is composed of these numerical representations.

B. Feature selection

The selectKBest method from the scikit-learn library was used to select the 20 features most important in determining the clinical status of a patient. The f_regression() scoring function was deemed the most suitable as we were testing the target, DX_NUMERICAL, against multiple regressors. The scoring function calculated the Pearson correlation coefficient between the features and the target and the selectKBest method then returned the 20 features with the highest scores. Pearson correlation coefficient (PCC) ensured selected features, when isolated, had a significant correlation with the diagnosis. PCC determines values between -1 and +1 to indicate negative linear correlation and positive linear correlation respectively. Value 0 indicates no linear correlation [9][10]. It is calculated by the formula below:

$$r = \frac{n\sum xy - \sum x \sum}{\sqrt{(n\sum x^2 - (\sum x)^2)}\sqrt{(n\sum y^2 - (\sum y)^2)}} \quad (1)$$

C. Building and Selecting Regression model

After the selection of the 20 features with the k highest scores from the selectKBest method, the data was split into a training and testing set, with 67% of the data allocated for training and 33% of the data allocated for testing. This split was done with a random seed to ensure that the data chosen was not a cofactor. Three different regression models were tested : XGBoost Classification, Random Forest Regression, and Linear Regression.

XGBoost is a Gradient Boosting Algorithm which sequentially adds classification trees that account for residuals of the previous model. It has a focus on performance, cache awareness, and scalability and uses block structure for parallel learning. This results in faster computation in comparison to other gradient boosting algorithms [11][12]. XGBoost minimizes the loss function shown below:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (2)$$

Random Forest is a supervised learning algorithm which aggregates multiple random decision trees. As each tree classifies the data, the model classifies based on the majority. Contrary to XGBoost that makes use of trees with a smaller number of splits, Random Forest uses trees that are grown to their maximum extent [13][14]. Random Forest Regression makes a prediction at a new point x using the function shown below:

$$Regression : \hat{f}_{rf}^{B} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{3}$$

Linear regression models the linear relationship between an output variable and a set of input variables. The 'best fit' for the data is calculated through a process of minimizing the loss. A drawback of the model is that it assumes a linear relationship between variables [15]. Linear regression is described by the following algorithm:

$$\hat{\theta} = argmin_{\theta} J(\theta) = \|A\theta - y\|^2 \tag{4}$$

$$minimizer : \hat{\theta} = (A^T A)^{-1} A^T y \tag{5}$$

To train the models and test the models, XGBoost libraries were used for XGBoost Classification, sklearn.ensemble libraries for Random Forest Regression, and sklearn.linear_model libraries for linear regression. Patients of AD are diagnosed into discrete classes so a classification model was chosen as it aligns closely to the nature of the progression and diagnosis of Alzheimer's Disease. Based on overall accuracy, XGBoost was finalized as our classification model of choice.

D. Tracking Patient Progression of Select Features

Our next goal was to track the progression of AD based on the top 20 features. Due to the variation in the data resulting from patients entering the study at different stages of AD, different dates and being examined with varying consistency, the data was split into five datasets based on patients' baseline (initial) diagnosis: Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), Significant Memory Concern (SMC), and Alzheimer's Disease (AD). This was informed by our data visualizations that showed differences in progressions of patients based on baseline diagnosis, which can be seen in Figures 1-5 in the Appendix. To prevent the model from tracking each patient's progression individually and being unable to track patients separate from the ADNI database, the data was further simplified by using the 'average' patient in each category. This is represented in Figures 6-10 in the Appendix.

Then, the Seasonal Autoregressive Integrated Moving Average with eXogenous regressors model (SARIMAX) was used on each of the 20 features for all five groups. SARIMAX is one of the most widely used methods for time series data forecasting [16]. It is based on the idea that previous data alone can help determine future states. 'Auto Regressive' means it is a linear regression model, and the Y value depends on its own lags. 'Moving average' means the Y value depends only on the lagged forecast errors. The two estimates of Y are calculated and combined linearly, alongside an error term, to estimate the final value of Y. The 'Seasonal' aspect allows SARIMAX to handle data that has white noise. While ARIMA (Autoregressive Integrated Moving Average) cannot handle data with a seasonal component, SARIMAX can. It uses three hyperparameters to specify the differencing, autoregression, and moving average. SARIMAX is described using the following formulas:

$$y_t = x_t + u_t \tag{6}$$

$$u_t = \mu + \eta_{t-1} + \zeta_t \tag{7}$$

All of the data except for the final visit, for each of the five groups, was used to train this time series model. The model was then used to predict the results of the final visit of the 'average' patient of each category.

E. Testing Progression Model with Regression Model

Once the predicted values for each of the 20 features for each group was obtained, the results were classified using the XGBoost model developed previously. This resulted in 'predicted' diagnoses for the 'average' patient, which could be compared to the actual 'average' diagnosis for each group.

III. Results

A. K Best Values

Using PCC, we found that the top 20 features that corresponded with the diagnosis were: APOE4 (a gene that is a risk factor for developing AD), the neuropsychological test scores of the following: CDRSB (Clinical Dementia Rating Sum of Boxes), ADAS11 (The 11-point Alzheimer's Disease Assessment Scale), ADAS13 (the 13-point Alzheimer's Disease Assessment Scale), MMSE (Mini Mental State Exam), RAVLT_Immediate (Rey Auditory Verbal Learning Test for immediate recall), RAVLT_learning (Rey Auditory Verbal Learning Test, learning component), RAVLT_percent forgetting (Rey Auditory Verbal Learning Test, forgetting component), Functional Assessment Questionnaire, ECOG_SP_Total, the following baseline scores: CDRSB baseline, ADAS11 baseline, ADAS13 baseline, MMSE Baseline, RAVLT_Immediate baseline, RAVLT_learning baseline, RAVLT_percent forgetting baseline, and FAQ baseline, and the Hippocampus volume and Entorhinal Volume.

B. Classification Results

Classifiers were analyzed based on the Jaccard index [17]. The Jaccard Index is defined as the size of the intersection of the two sets, divided by the size of the union of the two sets. Our XGBoost classifier had an accuracy of 88.19%, our Random Forest Regressor had an accuracy of 87.66%, and our Linear Regression had an accuracy of 72%. This, as stated above, drove XGboost to become the chosen classification model.

C. SARIMAX Results

The results of testing SARIMAX on the final visit of each averaged patient (baseline CN, baseline EMCI, baseline LMCI, baseline SMC, baseline AD), alongside the actual results and the percentage error are shown in Tables 1-5. The results show a large variation in error between features with some, namely APOE4 in EMCI showing upwards of 263% error between the real and predicted value, while others, namely MMSE_bl, showing an error of around 0.1%. As some data is largely constant among months, such as features detailing baseline values, while other data is subjective, such as results from questionnaires, this is reasonable.

D. Predicted Diagnosis Accuracy

Finally, we tested the results of SARIMAX by using them to create a predicted diagnosis. The predicted diagnosis was generated by the XGBoost model we had previously created and each category was based on the baseline diagnosis, as stated above. For the CN category, the average patient had a clinical status of 2.5, corresponding to MCI. and predicted diagnosis for this patient was 3, corresponding to LMCI. For the EMCI category, the average patient had a clinical status of 2.333, corresponding to MCI and the predicted diagnosis for this patient was 3, corresponding to LMCI. For the LMCI category, the average patient had a clinical status of 3, corresponding to LMCI and this exactly matched the predicted diagnosis. For the SMC category, the average patient had a clinical status of 2, corresponding to EMCI. The predicted patient had a diagnosis of 1, corresponding to CN. Both the average patient and predicted patient in the AD category had a diagnosis of 5, corresponding to AD. These results are shown in Table 6 in the Appendix.

IV. Discussion

Our models were able to successfully predict the clinical status of a patient and track the progression of the top 20 features relating to AD but there were some limitations that were encountered. As discussed in the results, the XGBoost classifier resulted in integer-valued predicted diagnoses. Though individual patients had a diagnosis with integer value, the mean patient for each group was not expected to have an integer-valued diagnosis. Thus, there was some inherent error when comparing the actual and predicted diagnosis of the patients. Even accounting for this error, all predicted values were within one integer value, signifying that the same clinical status would be determined.

We decided to use the top 20 features in our model as deemed them representative of MRI, PET, and neurophysiological assessment results and were satisfied with the accuracy that resulted from them. To further improve our findings, one can optimize the classifier by setting the number of features as a hyperparameter and finding the optimal number of features to use.

We tested multiple models based on classification trees or linear regression, but encountered long run times and lack of accuracy. XGBoost was determined to be the most suitable when classifying patients based on clinical status because it accounted for the long run times of other gradient boosting algorithms and had the highest accuracy of the tested models.

For time series forecasting, we aimed to predict the average patient overall, as this would allow us to aid in the overall patient treatment. However, there was a strong variation in data between patients. Patients entered the study with different baseline diagnoses, on different dates, and were tested with varying consistency. Visualizing the data showed that patients also had different rates of progression. To account for these differences, the population was segmented to minimize the variation between patients in the group. Patients were categorized based on their baseline diagnosis, which was a decision motivated by both the data visualizations and the knowledge that Alzheimer's Disease progresses more rapidly in later stages.

Although our predictions and XGBoost model had high accuracy, it was simplified as we only used data visualizations to determine the progression was different for patients who started with different baseline diagnoses. To further improve our findings, one could either more strongly prove the distinction between segmentations or adapt the model to allow for all individual patients to be represented, rather than create the average patient in a given group. Further research should also be conducted to find if a more general trend in the progression of Alzheimer's Disease exists, using all of the data provided, instead of finding trends of progression in patients grouped by their baseline diagnosis. This could result in a more comprehensive view on the progression of AD in patients.

The errors in our final predictions may not be fully attributed to an incorrect SARIMAX forecast, and could be influenced by the error in our XGBoost model itself. Considering both the SARIMAX forecast and the XGBoost model have a margin of error, one could separate the two so that the accuracy of the forecasting can be more correctly determined.

Finally, simplifying the model to use the 'average' patient for each category resulted in less data for forecasting purposes. Of our five groups, three groups had fewer than 15 time series data points that could be used to train the SARIMAX models. This resulted in higher errors in the forecasting for those groups, compared to the other groups with more time series data points. This can be mitigated by using an estimation algorithm to determine the progression of each feature in the years between the assessments. Currently, the data is collected in 3 year intervals, so the 1- and 2- year marks can be estimated and filled in, which would result in each group having three times as many data points as previously. This could potentially result in a more accurate time series forecasting model.

V. Conclusion

We were able to successfully build regression models that determine a patient's diagnosis. We first pre-processed the data, condensing it from a data set of around 12000 entries

and 2000 features to one of 20 features, determined by SelectKMeans filtering, and data organized by patient ID and date of visit. We then fit various regression and classification models, including XGBoost Classification, Random Forest Regression, and Linear Regression, to the data and tested their accuracies. We found that the XGBoost classifier had the highest accuracy at 88.19% and chose to use this as our classifier moving forward.

The data was then prepared for time series forecasting. The dataset was separated into five distinct groups, grouped by their baseline diagnoses (i.e. the diagnosis they had when they first joined ADNI). The 'month' at which patients entered the study was used as a common time scale between all patients and the data was averaged for all patients at a given month timestamp. This was done by reducing all 20 features and hundreds of rows into one row per month timestamp. Thus, the 'average' patient grouped by baseline diagnosis was created.

Finally we conducted time series forecasting and testing. The SARIMAX time series forecasting model was applied on all 20 features for all 5 groups and trained using all of the data points except those of the last timestamp. We then tested the model by predicting the progressions of the 20 features in the final timestamp and used our classifier to determine the estimated diagnosis based on the 20 feature predictions. Comparing his estimated diagnosis to the actual 'average patient' diagnosis, gave us desired results with fairly high accuracy, especially for the groups that had more data points.

Present methods of diagnosing AD progression categorize patients into five stages: preclinical AD, MCI due to AD, mild dementia due to AD, moderate dementia due to AD and severe dementia due to AD [18]. These are primarily based on neurophysiological assessments. However, these assessments often use different thresholds to classify patients, who may not perform uniformly across all these scales. In this case, clinicians must use their best judgement to determine the patient's stage of AD. Our regression models can simplify this process as they use an aggregate of the 20 most relevant features to predict the patient's stage of AD. The model we have can fairly accurately determine the stage a patient may be in, given those data points.

Furthermore, our progression models were able to track the progression of select features that determine patients' cognitive state. This can help clinicians determine a patient's subsequent cognitive states based on their current cognitive assessments, neuroimaging results and genetic factors. Additionally, it can help them estimate the rate at which a patient may progress given their current state. This can also help clinicians administer individualized treatment at an earlier stage, which could help prolong the development of cognitive decline in patients with AD. Given that the progression of AD occurs at different rates for patients, to be administered targeted treatment at an earlier stage could help patients maintain their present condition for a longer period of time.

VI. Appendix

A. Figures

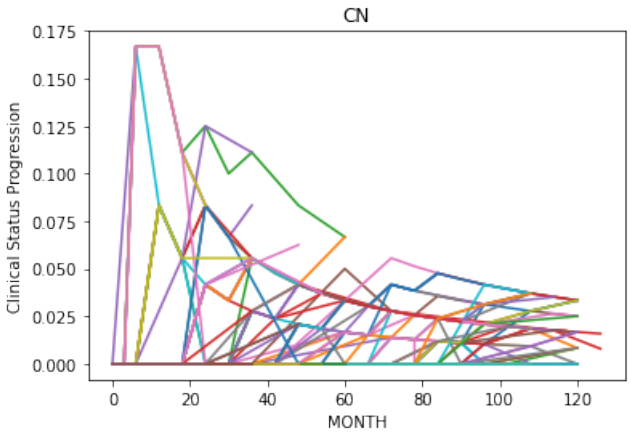

Fig. 1: Clinical Status progression rate of baseline CN patients observed over months from baseline measurement.

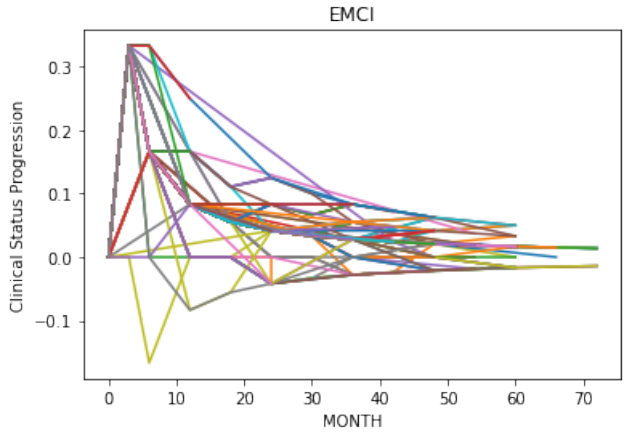

Fig. 2: Clinical Status progression rate of baseline EMCI patients observed over months from baseline measurement.

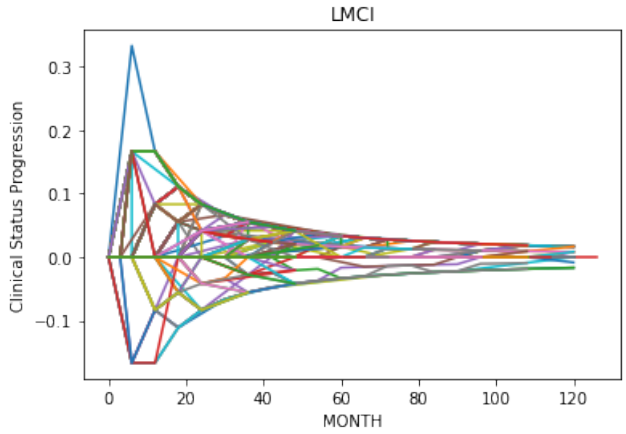

Fig. 3: Clinical Status progression rate of baseline LMCI patients observed over months from baseline measurement.

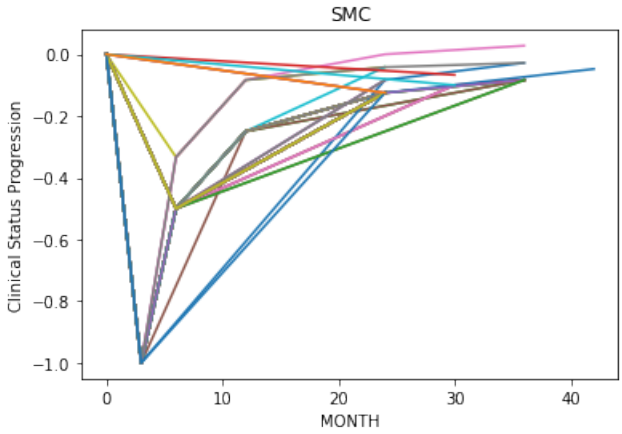

Fig. 4: Clinical Status progression rate of baseline SMC patients observed over months from baseline measurement.

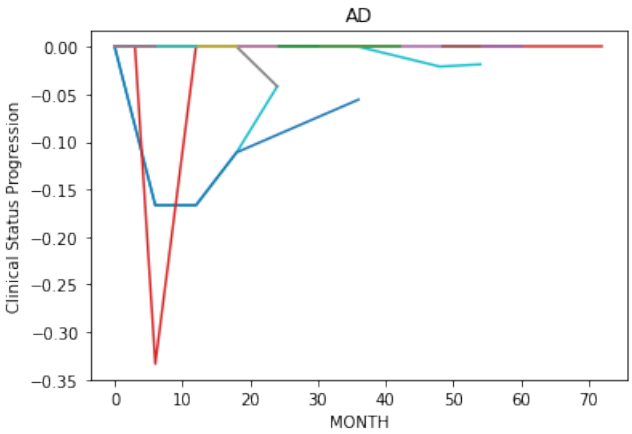

Fig. 5: Clinical Status progression rate of baseline AD patients observed over months from baseline measurement.

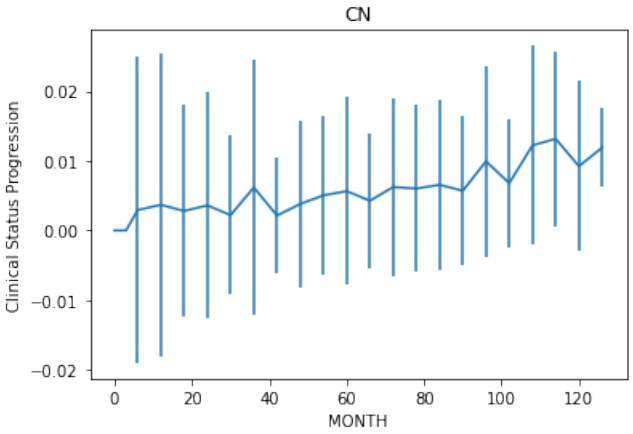

Fig. 6: Clinical Status progression rate of average baseline CN patient observed over months from baseline measurement. The lines represent standard deviation.

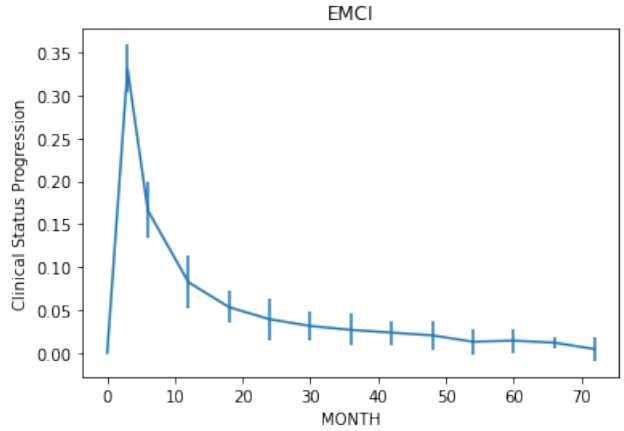

Fig. 7: Clinical Status progression rate of average baseline EMCI patients observed over months from baseline measurement. The lines represent standard deviation.

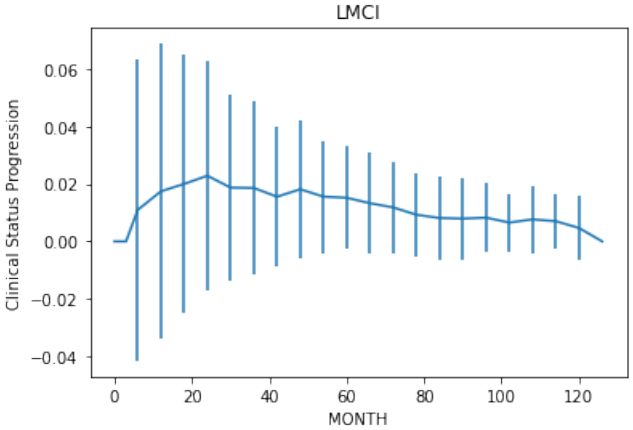

Fig. 8: Clinical Status progression rate of average baseline LMCI patient observed over months from baseline measurement. The lines represent the standard deviation.

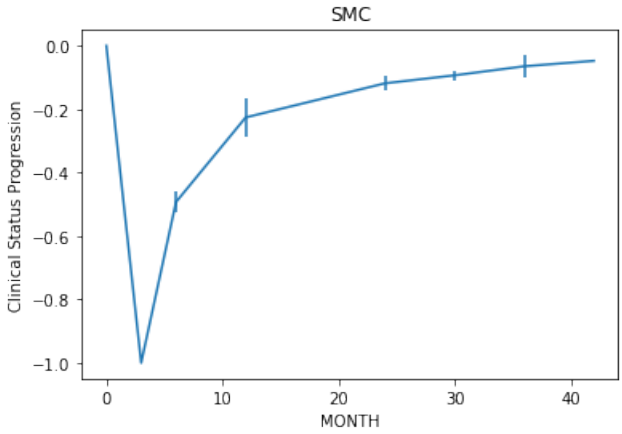

Fig. 9: Clinical Status progression rate of average baseline SMC patient observed over months from baseline measurement. The lines represent the standard deviation.

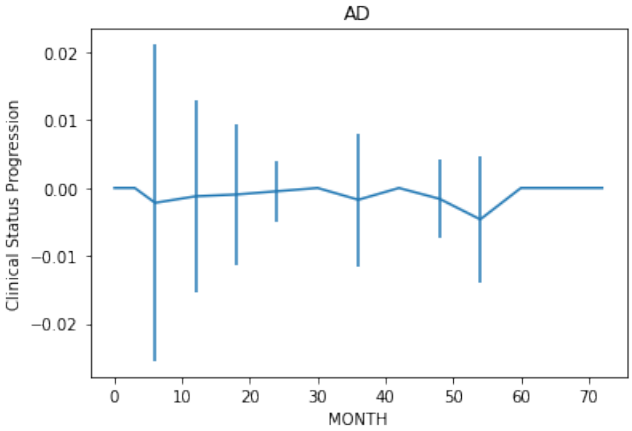

Fig. 10: Clinical Status progression rate of average baseline AD patient observed over months from baseline measurement. The lines represent the standard deviation.

B. Tables

	Actual	Predicted	Percentage_Error
APOE4	0.212766	0.174990	17.7546
CDRSB	1.648936	1.117618	32.2219
ADAS11	9.553191	9.057013	5.19385
ADAS13	14.978723	14.876513	0.682368
MMSE	27.404255	28.309888	3.30472
RAVLT_immediate	38.829787	33.820302	12.9011
RAVLT_learning	4.276596	4.288399	0.276
RAVLT_perc_forgetting	57.142135	71.567885	25.2454
FAQ	4.787234	2.861832	40.2195
EcogSPTotal	1.653798	1.504630	9.01973
Hippocampus	6774.297872	6343.364704	6.3613
Entorhinal	3486.787234	3093.093844	11.291
CDRSB_bl	0.021277	0.004418	79.2339
ADAS11_bl	5.681064	6.437070	13.3075
ADAS13_bl	8.617234	9.544788	10.7639
MMSE_bl	29.063830	29.098547	0.119451
RAVLT_immediate_bl	45.255319	46.662772	3.11003
RAVLT_learning_bl	6.148936	8.193577	33.2519
RAVLT_perc_forgetting_bl	32.606729	32.588891	0.0547062
FAQ_bl	0.085106	0.060782	28.5808

TABLE I: Predicted values of the top 20 features from the selectKBest method as a result of the SARIMAX model for the average baseline CN patient. This is compared with the actual values of the last month's measurements and the percentage error between them.

EMCI

	Actual	Predicted	Percentage_Error
APOE4	0.111111	0.403427	263.085
CDRSB	1.000000	1.665895	66.5895
ADAS11	7.444444	6.014643	19.2063
ADAS13	10.444444	12.796458	22.5193
MMSE	28.111111	27.433028	2.41215
RAVLT_immediate	46.555556	36.929746	20.676
RAVLT_learning	3.555556	4.857995	36.6311
RAVLT_perc_forgetting	44.849401	58.564712	30.5808
FAQ	1.333333	3.011111	125.833
EcogSPTotal	1.496552	1.696070	13.3318
Hippocampus	7104.222222	7939.769040	11.7613
Entorhinal	3972.888889	3634.611107	8.51465
CDRSB_bl	1.333333	1.072715	19.5464
ADAS11_bl	6.333333	7.190031	13.5268
ADAS13_bl	9.333333	11.961954	28.1638
MMSE_bl	28.333333	28.280352	0.186992
RAVLT_immediate_bl	43.888889	41.063924	6.43663
RAVLT_learning_bl	6.000000	5.219291	13.0118
RAVLT_perc_forgetting_bl	36.326751	45.570841	25.4471
FAQ_bl	1.000000	1.039348	3.93481

TABLE II: Predicted values of the top 20 features from the selectKBest method as a result of the SARIMAX model for the average baseline LMCI patient. This is compared with the actual values of the last month's measurements and the percentage error between them.

	Actual	Predicted	Percentage_Error
APOE4	0.00000	0.412355	100
CDRSB	1.00000	3.244761	224.476
ADAS11	19.00000	16.066208	15.441
ADAS13	30.00000	23.863064	20.4565
MMSE	27.00000	24.503723	9.24547
RAVLT_immediate	13.00000	30.229402	132.534
RAVLT_learning	4.00000	3.366699	15.8325
RAVLT_perc_forgetting	80.00000	77.750376	2.81203
FAQ	2.00000	13.732877	586.644
EcogSPTotal	1.63158	2.643950	62.0484
Hippocampus	7029.00000	5945.023544	15.4215
Entorhinal	3864.00000	3330.948670	13.7953
CDRSB_bl	1.50000	1.286194	14.2537
ADAS11_bl	15.00000	10.481156	30.1256
ADAS13_bl	24.00000	16.765496	30.1438
MMSE_bl	26.00000	27.880323	7.23201
RAVLT_immediate_bl	23.00000	38.407503	66.9891
RAVLT_learning_bl	7.00000	4.175319	40.3526
RAVLT_perc_forgetting_bl	25.00000	56.271835	125.087
FAQ_bl	0.00000	1.747152	100

TABLE III: Predicted values of the top 20 features from the selectKBest method as a result of the SARIMAX model for the average baseline SMC patient. This is compared with the actual values of the last month's measurements and the percentage error between them.

SMC

	Actual	Predicted	Percentage_Error
APOE4	1.0	0.379037	62.0963
CDRSB	3.5	0.569568	83.7266
ADAS11	12.0	7.208931	39.9256
ADAS13	18.0	11.652043	35.2664
MMSE	28.0	28.109164	0.389871
RAVLT_immediate	37.0	41.416608	11.9368
RAVLT_learning	5.0	4.896962	2.06075
RAVLT_perc_forgetting	100.0	36.969855	63.0301
FAQ	6.0	2.438051	59.3658
EcogSPTotal	2.0	1.476670	26.1665
Hippocampus	5059.0	7528.533928	48.8147
Entorhinal	2731.0	3531.704717	29.3191
CDRSB_bl	0.5	0.038835	92.2331
ADAS11_bl	8.0	8.227157	2.83947
ADAS13_bl	12.0	12.947299	7.89416
MMSE_bl	30.0	29.474214	1.75262
RAVLT_immediate_bl	41.0	37.795085	7.81687
RAVLT_learning_bl	7.0	4.937673	29.4618
RAVLT_perc_forgetting_bl	100.0	60.267517	39.7325
FAQ_bl	1.0	0.464568	53.5432

TABLE IV: Predicted values of the top 20 features from the selectKBest method as a result of the SARIMAX model for the average baseline EMCI patient. This is compared with the actual values of the last month's measurements and the percentage error between them.

AD

	Actual	Predicted	Percentage_Error
APOE4	1.00000	0.935733	6.42668
CDRSB	11.00000	6.073737	44.7842
ADAS11	20.33000	26.662411	31.1481
ADAS13	32.67000	35.086129	7.39556
MMSE	16.00000	18.623604	16.3975
RAVLT_immediate	12.00000	15.304186	27.5349
RAVLT_learning	3.00000	1.953666	34.8778
RAVLT_perc_forgetting	100.00000	98.942425	1.05758
FAQ	28.00000	14.612878	47.8111
EcogSPTotal	3.97436	2.993198	24.6873
Hippocampus	4349.00000	5675.658885	30.5049
Entorhinal	2332.00000	2567.014903	10.0778
CDRSB_bl	3.00000	3.096457	3.21524
ADAS11_bl	16.67000	19.066602	14.3767
ADAS13_bl	29.67000	28.571204	3.70339
MMSE_bl	22.00000	23.293926	5.88148
RAVLT_immediate_bl	14.00000	17.664123	26.1723
RAVLT_learning_bl	1.00000	3.569706	256.971
RAVLT_perc_forgetting_bl	100.00000	71.791965	28.208
FAQ_bl	12.00000	-5.661163	147.176

TABLE V: Predicted values of the top 20 features from the selectKBest method as a result of the SARIMAX model for the average baseline AD patient. This is compared with the actual values of the last month's measurements and the percentage error between them.

	Predicted Diagnosis	Real Diagnosis
Clinical Status	3	2.5

Early Mild Cognitive Impairment

	Predicted Diagnosis	Real Diagnosis
Clinical Status	3	2.333333

Late Mild Cognitive Impairment

	Predicted Diagnosis	Real Diagnosis
Clinical Status	3	3.0

Severe Memory Concern

	Predicted Diagnosis	Real Diagnosis
Clinical Status	1	2.0

Alzheimer's Disease Positive

	Predicted Diagnosis	Real Diagnosis
Clinical Status	5	5.0

TABLE VI: Predicted values of the top 20 features from the selectKBest method as a result of the SARIMAX model for the average baseline AD patient. This is compared with the actual values of the last month's measurements and the percentage error between them.

REFERENCES

[1] "Alzheimer's disease statistics." Alzheimer's News Today. https://alzheimersnewstoday.com/alzheimers-disease-statistics (accessed May 2, 2020).

[2] "What happens to the brain in Alzheimer's disease?" National Institute of Aging. https://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease (accessed May 2, 2020).

[3] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. M. Aparici, S. C. Behr, R. R. Flavell, S.-Y. Huang, K. A. Zalocusky, L. Nardo, Y. Seo, R. A. Hawkins, M. H. Pampaloni, D. Hadley, and B. L. Franc, "A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain," *Radiology*, vol. 290, no. 2, pp. 456–464, Nov. 2018.

[4] V. Mantzavinos and A. Alexiou, "Biomarkers for Alzheimers Disease Diagnosis," *Current Alzheimer Research*, vol. 14, no. 11, Jun. 2017.

[5] H. Zetterberg and S. C. Burnham, "Blood-based molecular biomarkers for Alzheimer's disease," *Molecular Brain*, vol. 12, no. 1, 2019.

[6] S. Sapkota, T. Huan, T. Tran, J. Zheng, R. Camicioli, L. Li, and R. A. Dixon, "Alzheimer's Biomarkers From Multiple Modalities Selectively Discriminate Clinical Status: Relative Importance of Salivary Metabolomics Panels, Genetic, Lifestyle, Cognitive, Functional Health and Demographic Risk Markers," *Frontiers in Aging Neuroscience*, vol. 10, Feb. 2018.

[7] "ADNI: Data Types."ADNI. http://adni.loni.usc.edu/data-samples/data-types (accessed May 4, 2020).

[8] TADPOLE - Grand Challenge," TADPOLE Grand Challenge. [Online]. Available: https://tadpole.grand-challenge.org/Home/ (accessed June 16, 2020).

[9] S. Glen, "Correlation Coefficient: Simple Definition, Formula, Easy Calculation Steps," Statistics How To, 14-Apr-2020. [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula (accessed June 16, 2020).

[10] J. Brownlee, "How to Choose a Feature Selection Method For Machine Learning," Machine Learning Mastery, 14-May-2020. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data (accessed June 16, 2020).

[11] J. Brownlee, "A Gentle Introduction to XGBoost for Applied Machine Learning," Machine Learning Mastery, 21-Apr-2020. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning (accessed June 16, 2020).

[12] "XGBoost Documentation," XGBoost Documentation - xgboost 1.2.0-SNAPSHOT documentation. [Online]. Available: https://xgboost.readthedocs.io/en/latest (accessed June 16, 2020).

[13] L. Breiman and A. Cutler, "Random Forests Leo Breiman and Adele Cutler," Random forests - classification description. [Online]. Available: https://www.stat.berkeley.edu/ breiman/RandomForests/cc_home.html (accessed June 16, 2020).

[14] J. Brownlee, "How to Implement Random Forest From Scratch in Python," Machine Learning Mastery, 13-Aug-2019. [Online]. Available: https://machinelearningmastery.com/implement-random-forest-scratch-python/ (accessed June 16, 2020).

[15] "Linear Regression Example," scikit. [Online]. Available: https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html (accessed June 16, 2020).

[16] J. Brownlee, "A Gentle Introduction to SARIMA for Time Series Forecasting in Python," Machine Learning Mastery, 21-Aug-2019. [Online]. Available: https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/ (accessed June 16, 2020).

[17] "sklearn.metrics.accuracy_score," scikit. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html (accessed June 16, 2020).

[18] "What to know about the stages of Alzheimer's," Mayo Clinic, 19-Apr-2019. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-stages/art-20048448#: :text=There%20are%20five%20stages%20associated%dementia-%20due%20to%20Alzheimer's%20disease (accessed June 16, 2020).