Detection and degradation of Alzheimer's disease

Niharika Sinha
ns4451@nyu.edu
New York University
New York, New York, USA

Rachit Jain
rj2219@nyu.edu
New York University
New York, New York, USA

Maria Lorena Carlo Unda
mlc9971@nyu.edu
New York University
New York, New York, USA

ABSTRACT

Approximately 6.5 million Americans are diagnosed with Alzheimer's in 2022. Alzheimer's disease results in loss in brain capacities in patients, which can ultimately start interfering with their everyday life. It can lead to severe memory impairment. Early detection of Alzheimer's can help in identifying the correct procedures required for its treatment and cure. We measure 7 features related to brain activities for a person across various time intervals, starting from the patient's first visit. In this regard, we developed various classification models for predicting whether a person, given this time series data for the values of the 7 features recorded at uniform testing intervals, is likely to get affected with Alzheimer's disease. We have implemented Synthetic Minority Oversampling Technique (SMOTE) for removing data skewness which seems to play a major role in generalization of the program. As per our findings, Random Forest Classifer performs the best with SMOTE applied to it with an accuracy of 84.72%.

KEYWORDS

healthcare, random forest, SMOTE, machine learning, alzheimer's

ACM Reference Format:
Niharika Sinha, Rachit Jain, and Maria Lorena Carlo Unda. 2018. Detection and degradation of Alzheimer's disease. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Alzheimer's disease has a severe impact on a person's memory and problem solving capacities, and can start affecting the daily life of a patient. It results in a decline in brain abilities, leading to severe memory impairment and inability to carry out everyday tasks [5]. Over 6 million Americans over 65 suffer from Alzheimer's disease. If researchers don't find new ways to prevent or treat it, the number of people with Alzheimer's and dementia aged 65 and older is expected to reach 12.7 million by 2050 [2]. By keeping an eye on the patient's brain activity, it is possible to track the disease's early onset. Early detection allows for quicker examinations and quicker access to beginning treatment with cutting-edge studies. In order to help doctors diagnose their patients, it is our goal to develop a classification model for predicting the onset of Alzheimer's and visualization dashboards for uncovering data insights and tracking patients' physical and mental health.

We have a real-time time-series data set from Kasturba Medical Hospital, Manipal [3] of about 1705 patients whose brain activities were tracked over a period of about 15 revisits. The data is divided into 4 modals depicting cognitive, MRI and CSF biomarkers values for the patients tracked across visiting periods. The dataset does not necessarily have all the results of all the patients across all the testing periods, and due to this, we had to perform a number of data wrangling steps. Using the existing values of the tracking features provided by the modal- *CogPtMem, EcogPtTotal, Entorhinal, Hippocampus, TAU, Ptau* and *Abeta* values, and the demographic information, we have proposed classification models to detect whether a new patient is likely to get affected by this disease or not. We have also studied the correlation between the tracking features. We have conducted exploratory data analysis through visualizations in Tableau. We propose an efficient dashboard for clinicians and doctors to view patient history data quickly and make comparisons between the symptoms of an affected patient and an unaffected one. We have also studied the trend of the disease onset and progress with age, gender, and years of education, and the fluctuation in the brain activities across testing periods.

2 EXPLORATORY DATA ANALYSIS

Preliminary data analysis was done using basic Python commands and some very insightful Tableau visualizations. Analysis for null and missing values, data types, length of the data set, the range of all values, number of data points available for all classes of the Alzheimer's disease and class imbalances was conducted through basic Python commands. Tableau Visualizations helped us understand the distribution of data across various categories. This was particularly useful as our data is three dimensional.

Figure 1 is a histogram analysis of the age and gender distribution of all the patient data available to us across the various stages of Alzheimer's that patients were detected with. For all stages of the disease, the maximum number of patients who visited the hospital lie in the age bracket 70-75 years. While female patients are more in comparison to male patients in the earlier age bracket of 50-70 years, the male gender is dominant in the age bracket of 70-95 years. Highest number of the patients who got detected with Alzheimer's lie in the age bracket 70-80 years. More males than females got detected with Alzheimer's across almost all age groups.

Figure 2 shows the analysis of the years of education received and gender distribution of all the patient data available to us, across the various stages of Alzheimer's that patients were detected with. Almost 97% of the patients who visited the hospital during the study had received more than 12 years of education, and 63% had received more than 16 years of education for every stage of the disease. The distribution of male and female patients across years of education

Conference'17, July 2017, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

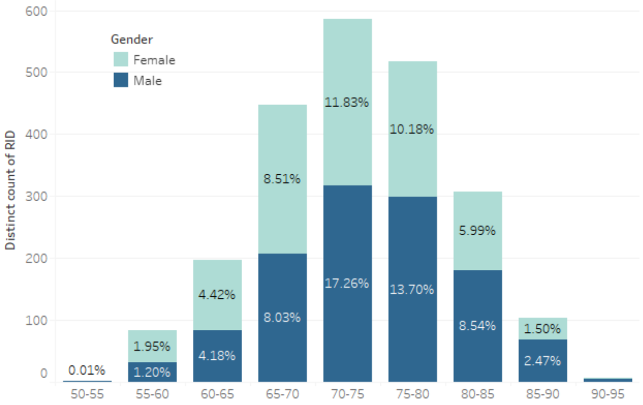


Figure 1: Patient Age and Gender Distribution across the different stages of Alzheimer's

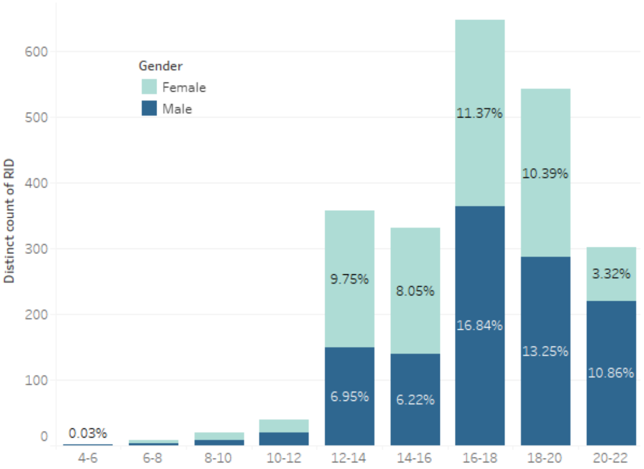


Figure 2: Patient years of education received and gender distribution across the different stages of Alzheimer's

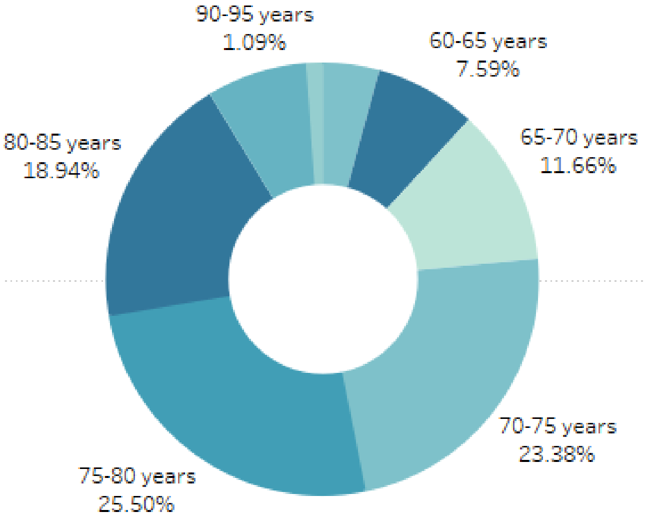


Figure 3: Percent of Total Patients affected with Alzheimer's based on Age

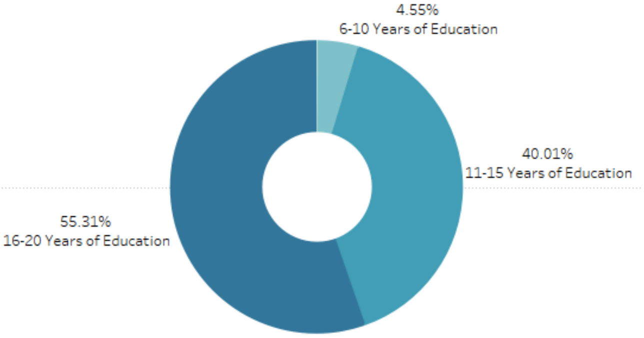


Figure 4: Percent of Total Patients affected with Alzheimer's based on Years of Education received

received are comparable in most cases, and males dominate above 16 years. These visualizations also show the trend of these 2 variables across the different stages of Alzheimer's that patients got detected with.

While the Figures 1 & 2 show the variation of patient age and years of education received with gender for all stages of Alzheimer's disease, Figures 3 and 4 focus on the demography analysis of only those patients who get detected with Alzheimer's. Figure 3 shows the percent of total patients detected with Alzheimer's across ranges of years of education received. This doughnut chart indicates that 55.31% of the total patients detected with Alzheimer's had received 16-20 years of education, implying that high education probably may not be a guard against the disease. On similar lines, figure 4 shows the percent of total patients affected with the disease across various age groups. Almost 50% of the total patients affected by the disease were in the age group 70-80 years.Very few patients above 90 years of age were detected with the disease, but this could also be because of lack of data for this age group. This chart indicates that 70-80 age bracket is the most susceptible to the disease, based on the data collected for the study. Such information could be very useful while prescribing medicines where age could be a cautionary factor.

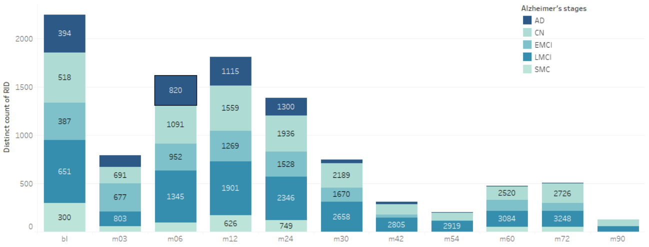


Figure 5: Number of Patient Visits and Revisits for testing intervals across all stages of Alzheimer's disease

Figure 5 shows the number of patients who visited the hospital and how much they frequent in the subsequent testing intervals.For each testing interval, the chart also indicates the running total of the patients in various stages of the disease. This graph clearly shows that most of the patients got detected with Alzheimer's by the *m30*, i.e. 30 th month from their first visit. This graph gave good insights while selecting the testing intervals for the classification model. As indicated by the graph, the classifiers achieved the best results when testing intervals only till the 30th month from the first patient visit were included. The plot also shows a decline in the number of patient revisits. It is human tendency to not follow up with the doctor at testing intervals of just 6 months. Also, patients who do not get affected by Alzheimer's also tend to slack off on their regular check ups. The decline in the number of patients revisits corroborates this human nature.

Figures 6, 7, 8 are gantt charts that plot the values of the test features for all patients across all the testing intervals, differentiating the patients that are at various stages of the disease through color. Figure 6 plots the CSF biomarkers values, figure 8 plots the MRI test values and figure 7 plots the cognitive test values. These figures confirm the scarcity in data, since every patient tends to not complete every revisit on time. Most of the patients who returned beyond 48 months from the baseline were not ultimately detected with Alzheimer's. This implies that patients who got affected by the disease stopped revisiting to record their feature values eventually. Few patients, during a revisit, may get a few sets of tests done and leave a few others. As we can see, for the testing interval of 3 months, there are barely any records for CSF Biomarkers (*PHENO*), while there are many patient records for MRI features. If we observe the baseline value of the MRI feature *Entorhinal*, we see an Alzheimer's patient with this value as an outlier. This is because a single test is not sufficient to make a prediction for the disease. Rather, a combination of all these features, together with the patient demography, determines whether a patient is likely to get affected by this disease or not. Due to these observations, we limited our analysis to only up to the 30th month testing interval. When we combined the data for all the feature values for the patients selected for model training, there were null values in the data since few test values were missing. To overcome this, we imputed the missing data using the KNN imputation algorithm.

3 METHODOLOGY

We propose the methodology of detection of Alzheimer's disease by using different 4 kinds of models.

3.1 Dataset

Data from 1705 patients' initial and subsequent visits are included in the dataset. A RID, or a unique identifier, is given to each patient. The data is in a time series format which is depicted by the VISCODE column. The VISCODE column is divided into categories which denote the testing periods where 'bl' corresponds to baseline, 'm06' corresponds to month 6 testing after baseline, 'm12' corresponds to month 12 testing after baseline is recorded, similarly for the subsequent months. The information is separated into four categories: cognitive modality [8], CSF biomarkers [7], MRI modality, and demographic modality. All of these models have information about the brain activity that was seen, such as data from the MRI modality for the entorhinal, hippocampus, etc. and from the cognitive modality for the ecogPtMem and ecogPtTotal, etc. These serve as the tracking sites for research into the causes of Alzheimer's disease.

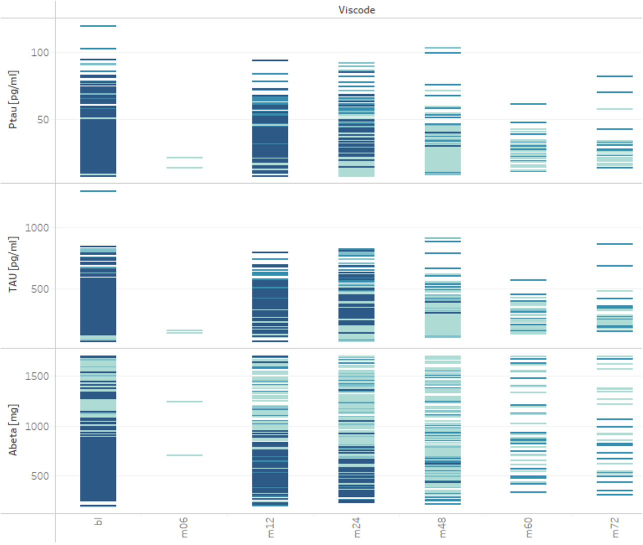


Figure 6: Values of CSF Biomarkers of all patients across all testing intervals

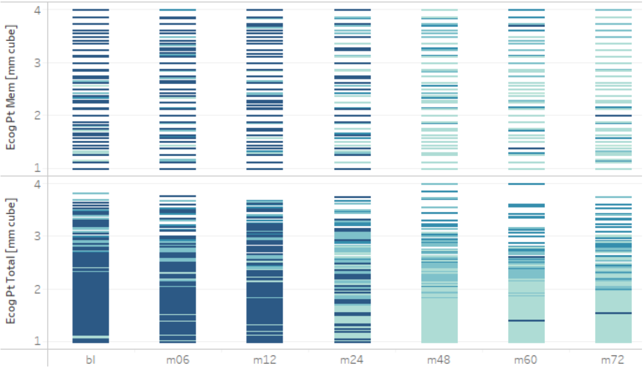


Figure 7: Values of Cognitive test features of all patients across all testing intervals

The data has some missing values which could be for a modality for a testing period or it could be that a patient did not get tested during a specific testing period. The data has a class imbalance in which the overall samples of the Alzheimer's Detected (AD) class are less compared to the other. Due to this reason, we chose to classify the samples as AD or non-AD. Table 1 gives us a good idea of how the data looks like and the units and the modality it belongs to.

Table 1: Dataset Description

Model Type	Value, Unit	Meaning
CSF Biomarkers	ABETA, mg	amyloid-β 1–42 peptide
CSF Biomarkers	TAU, pg/ml	Total tau protein
CSF Biomarkers	PTAU, pg/ml	Total tau phosphorylated protein
Cognitive	EcogPtMem, units	Everyday cognitive functioning memory
Cognitive	EcogPtTotal, units	Everyday cognitive functioning total
MRI	Hippocampus, mm^3	Hippocampal volume
MRI	Entorhinal, mm^3	Entorhinal cortical thickness
Demographic	AGE, years	Age of person
Demographic	PTEDUCAT, years	Number of years of education
Demographic	APOE4, protein/mg	APOE ϵ4 status
Demographic	PTGENDER, binary	Birth gender

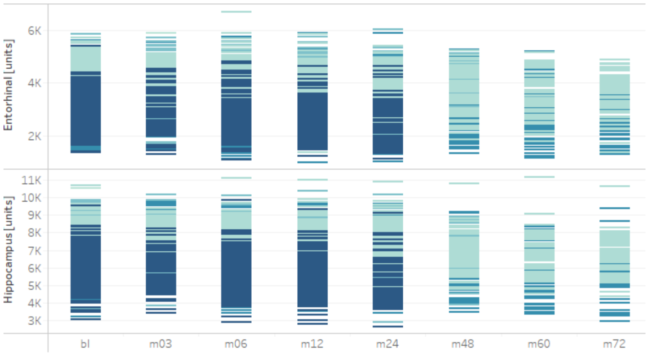


Figure 8: Values of MRI test features of all patients across all testing intervals

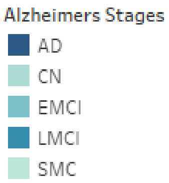


Figure 9: Legend for Figures 6, 7, 8

The output classes for the classification are specified by DX_BL as in the dataset:

(1) AD – Alzheimer's disease
(2) MCI – Mild cognitive impairment
(3) CN – Cognitively normal
(4) LMCI – Late mild cognitive impairment
(5) EMCI – Early mild cognitive impairment
(6) SMC – Significant memory concerns

The final output classes that we chose due to data density for the classification are:

(1) AD – Alzheimer's disease
(2) Not AD – Not Alzheimer's disease

3.2 Feature Selection

To analyze how a variation in one variable cause a variation in another, we performed correlation analysis using Pearson Correlation Coefficient. We performed correlation analysis of each feature with the other features and visualized it using heat maps depicted in Figure 10. As we can see, among the CSF biomarkers, FEATURES TAU and *PTAU* have a very strong correlation. Due to this, one of them can be gropped for our modeling. Similarly, *EcogPtMem* and *EcogPtTot* have a high correlation coefficient of 0.7, and hence one of them could be dropped. All the other correlation values for other biomarkers are shown in the heatmaps. The trend of one variable compared to another can also be observed in the scatter plot matrix as shown in figure 11.

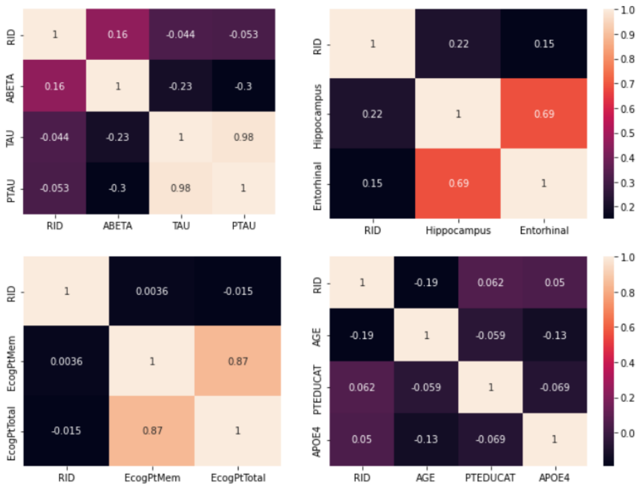


Figure 10: Correlation matrix heatmap of all features

3.3 Data Merging

We wanted to build a classifier that could predict the chances of a patient getting affected with Alzheimer's by incorporating the results of all the biomarkers into the model. For this purpose, the 4 separate datasets that were avilable to us had to be merged. The four

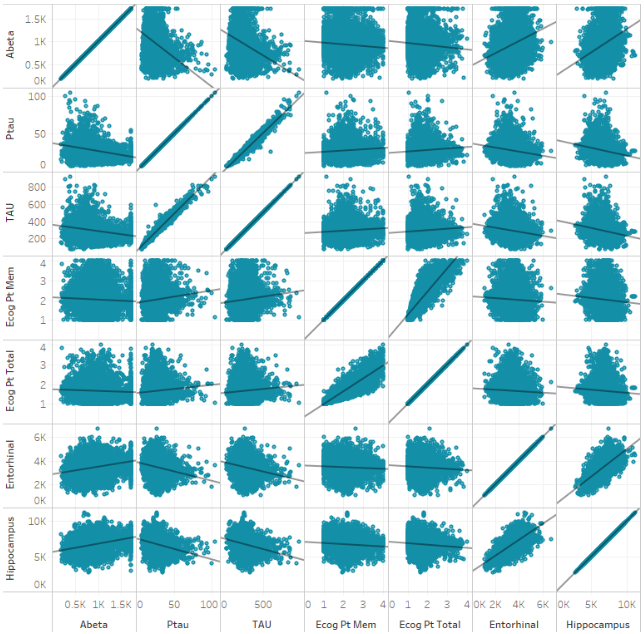


Figure 11: Scatter Plot Matrix between MRI, Cognitive and CSF Biomarkers

datasets were joined using outer join on *RID* and *VISCODE*. Due to an outer join, the resultant data had a few missing labels. The missing labels were inserted with the assumption that the diagnosis at the baseline is the ground truth.

3.4 Data Transformation

Upon plotting the data of number of samples in each class from Figure 5, we realized that most patients are coming for regular testing come from the baseline (bl) to month 24 (m24). Based on Figure 5, we kept the data points which were in these month categories only. This would help in solving class imbalance as there is not much data present for patients with Alzheimer's after these months. Columns in the dataset were converted from their string counterparts to categorical variables. The data was recorded in a way that there were a few duplicate rows. This is possible because of a data entry issue due to which some rows could have been recorded multiple times.

3.5 Data Selection

When a few patients came for testing, they got tested for a few parameters but not for the other parameters. For example - a patient got tested for his MRI values but did not get tested for his CSF Biomarkers in month 6 of testing. Due to this, some of the data points had missing columns. We carefully analyzed and removed the rows which had more than 3 NaN values in the row since this data would not give us a good definition of the data. The sequences were also padded for all the testing periods.

3.6 Data Imputation

Similar to section 4.2.4, when a patient does not get tested for a few parameters, there will be NaN values present. It is important that we deal with the missing values in the correct manner. Dropping all the rows with missing columns does not help model the overall problem. This can be dealt with by substituting the columns which have missing values with data. We initially modeled the missing data with the average of the values belonging to that class but that did not give good results. That could be because the data was not intelligently inserted into the data. An average might not be a good indicator of the data that is to be inserted.

Instead, we used K-Nearest Neighbours (KNN) imputation. KNN is an algorithm that works by when a new point comes, it checks the K-nearest neighbours and then calculates the distance between them and chooses the k-points which are closest. For imputing the data, we chose 28 neighbours which were the closest and the weights were decided by the distance parameter. This gave us an imputed dataset which has no missing values.

3.7 Oversampling using SMOTE

The procedure for balancing dataset skewness is called Synthetic Minority Oversampling Technique (SMOTE). Because there are so few samples of the textbfAD class in the dataset we worked on, there is a significant class imbalance. We oversample the data using the KNN algorithm, just like in section 4.3. Since the sample points in our situation are close together, we apply the textitSMOTE function, which is SMOTE for Regression, using the nearest neighbor idea. A random point is chosen, its closest neighbors are checked, a random neighbor is chosen from that sample space, and finally a sample data point is created between the two randomly picked points. The minority class is thus oversampled in favor of the balanced side. This data balancing process will help the final model generalize on unseen data. We use the imblearn library for this.

3.8 Support Vector Machine

We classified the data into two categories using a supervised classification model called the Support Vector Machine (SVM) [1]. Support vectors, or decision boundaries, are used by SVMs to optimally separate the data into the categories. Each data point is plotted onto an N-dimensional space, and the optimum support vectors are then determined. We can obtain a better support vector by plotting the vectors in a higher dimension because doing so provides us with more information. Using a linear kernel, we use this model from the scikit-learn library.

3.9 Random Forest Classifier

A supervised learning system called the Random Forest classifier [4] uses the ensemble learning method to identify classes. This method creates a final model by combining various methods. The final output is produced by taking the mean across all the models after each decision tree has been performed independently. Given that it uses the average of the models, this model will prevent overfitting. To create smaller datasets for each forest in the model, feature sampling is carried out. When selecting the data for the decision trees, the model randomly samples the data, providing good accuracy. We use this model from scikit-learn library.

3.10 Stochastic Gradient Descent Classifier

This classifier uses the Stochastic Gradient Descent (SGD) algorithm [6] for classifying the data points. The objective function is set to be some convex function which is then optimized by using the SGD loss function. This model is an efficient way of optimizing the classifier.

4 RESULTS

Results from our experiments are shown in Table 2. After final processing of the data, the training data had 1906 samples and the testing data had 818 samples. As can be seen from Table 2, the model performed best with the Random Forest Classifier. The reason why random forest performs so well is that it uses ensemble learning to predict the final output which is based on a voting mechanism.

While SVD and SGD models achieved decent performances too, their low accuracies could be due to the non linearity present in the data. The data fed to these models are actually three dimensional, which had to be flattened in order to be fed into these classifiers. The classifiers were trained on a set of 9 features, having correlation as shown in the heat map of figure 10. These 2 models are not able to capture the features as well as random forests. SVM achieved an accuracy of 67%, while SGD achieved an accuracy of about 65%. The high accuracy of Random Forest could be because of its ability to capture the non linearities in the data.

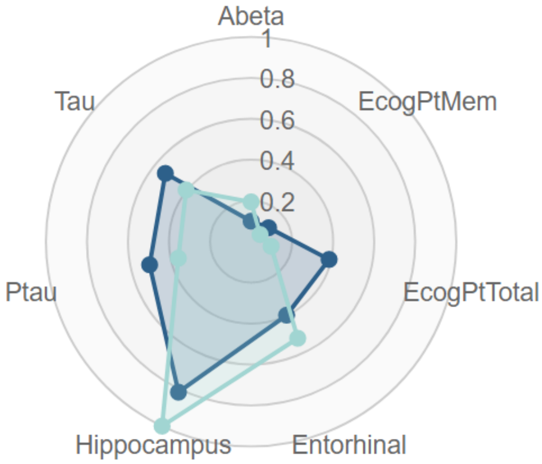


Figure 12: Average Normalized Values of Features for patients with and without Alzheimer's

The difference between the values of the biomarker features for Alzheimer's vs non Alzheimer's is as shown in figure 12. It is interesting to note that for MRI features, values for non Alzheimer's is more than for Alzheimer case. The opposite is true in case of cognitive biomarkers. For CSF biomarkers *Tau* and *Ptau*, the Alzheimer's value is more.

5 DISCUSSION

The classifiers achieved a maximum accuracy of 84.72% with Random Forest classifier. Analyzing data visualizations prior to running the classifiers gave insights that helped achieve a better accuracy score for the classifier. One such insight was to consider testing intervals only upto 30 months since the first patient visit. This was because figure 5 indicated that most of the Alzheimer's cases get detected within the first 30 months of the first visit. Similarly, plotting the values for CSF biomarkers for all the patients available from the study showed the scarcity in data and highlighted the fact that every patient did not necessarily attend follow ups at every testing interval. These plots also shed light on the human tendency of not following through every medical appointment, especially when there is no cause for concern (as we see for patients who never got detected by Alzheimer's). The most susceptible age to get affected by Alzheimer's, as indicated by figure 3 is 70-80 years. Although, among the patients > 90 years of age the percentage of Alzheimer's affected patients was high, but this number compared to the rest of the population still stands low. This is because not many people greater than 90 years of age are expected to show up for their baseline recordings in the first place. The trend of declining patient revisits can be seen for all stages of the disease. For Alzheimer's patients, this can be attributed to the fact that once affected by this illness, very few patients would have continued visiting the hospital at every testing interval simply to give a recording of their biomarker values. For non Alzheimer's patients, this behavious can be attributed to the human tendency of stopping medical visits for re check ups if your health condition is perfectly fine.

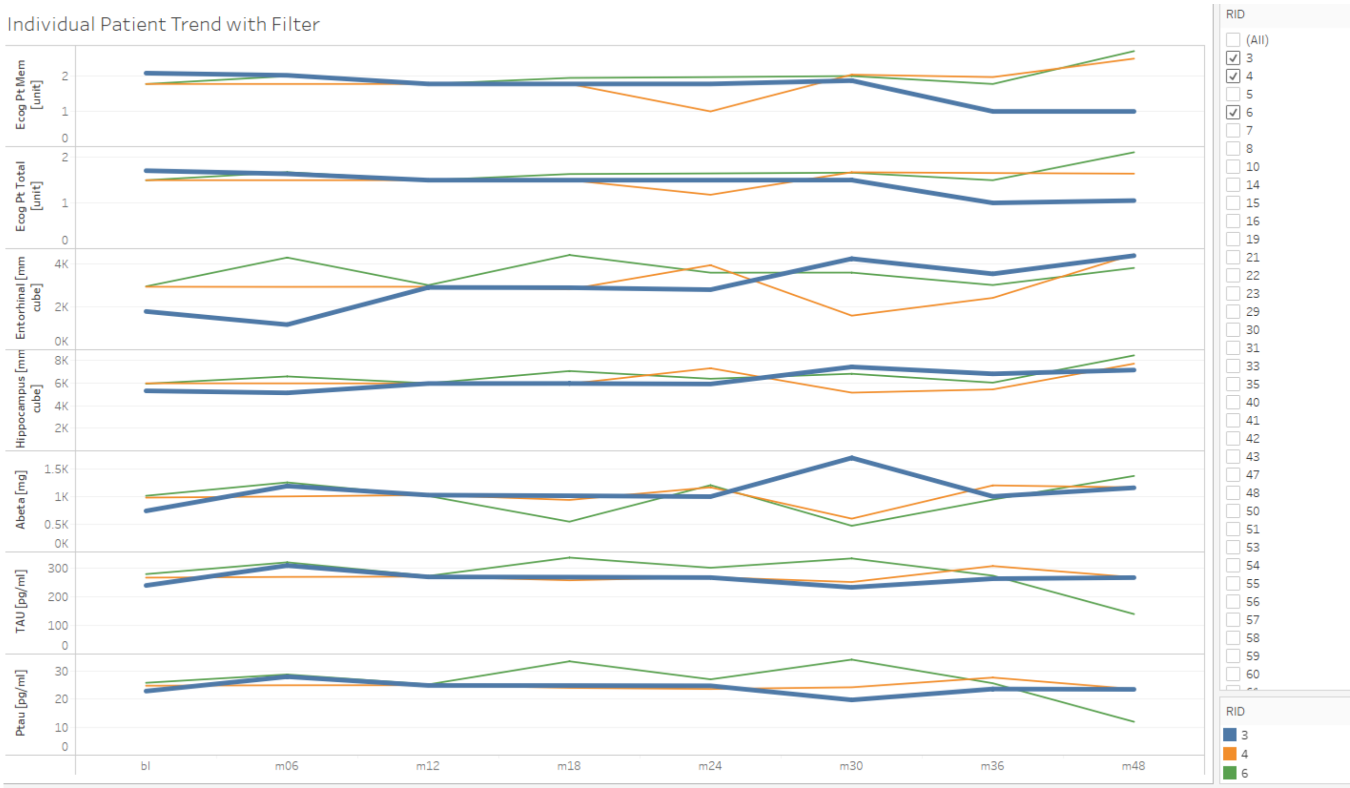
Figure 13 shows the trend of the feature values for MRI, cognitive and CSF biomarkers tests for all patients during each of their subsequent revisit. This is a great tool to get patient history, compare the results of several patients and even compare the traits of an affected patient with an unaffected one. The trend lines of an unaffected patient differs from an affected one in terms of line thickness. This helps in parallel comparison of the test values of all types of patients. Since our target audience for this project is doctors and healthcare professionals, we want to provide a historical view of all the cases that have been submitted so far.

Figure 12 is a radar chart that indicates the average relative values of all the test features for an Alzheimer's and a non Alzheimer's patient. For cognitive, MRI and CSF biomarker indicators, the average of all values were recorded separately for Alzheimer's and non Alzheimer's patients. These values were then normalized with respect to each other. Due to the huge difference in scale, the values of the same features were scaled up so that all features could be represented at the same level of granularity. For *Tau, Ptau* CSF biomarkers and *EcogPtTotal, EcogPtMem* cognitive indicators, higher value indicates chances of Alzheimer's. For features *Abeta* of CSF biomarker and *Hippocampus, Entorhinal* MRI indicators, a lower value is more indicative of Alzheimer's onset.

This project has definitely shed importance on the power of good data visualizations. Also, it shed light on the fact that a disease as complicated as Alzheimer's cannot be affected by one factor alone. For this study itself, there were 7 tests that were being tracked. An outlier in the expected range of even one of the features may result in abnormalities.

Table 2: Comparison of different Machine Learning Models

Model Name, Accuracy	Precision	Recall
Support Vector Classifier, 67.36%	70.70%	62.47%
Random Forest Classifier, 84.72%	83.79%	87.17%
Stochastic Gradient Descent Classifier, 65.77%	67.85%	63.66%



RID
☐ (All)
☑ 3
☑ 4
☐ 5
☑ 6
☐ 7
☐ 8
☐ 10
☐ 14
☐ 15
☐ 16
☐ 19
☐ 21
☐ 22
☐ 23
☐ 29
☐ 30
☐ 31
☐ 33
☐ 35
☐ 40
☐ 41
☐ 42
☐ 43
☐ 47
☐ 48
☐ 50
☐ 51
☐ 53
☐ 54
☐ 55
☐ 56
☐ 57
☐ 58
☐ 59
☐ 60

Figure 13: Dashboard for tracking patient medical history. Can be used for comparison of multiple patients and analyzing the trend of biomarkers for Alzheimer's

6 CONCLUSIONS AND FUTURE WORK

Our paper outlines the major classification algorithms that can be used for the detection of Alzheimer's disease based on multiple feature inputs. We draw the following conclusions:

- We develop an algorithm to show that just taking one feature, even if it is highly correlated with the output variable, cannot accurately detect Alzheimer's in a patient.
- We use KNN imputation for imputing some of the data points and use KNN based oversampling techniques called SMOTE for solving class imbalance.
- Refinement using SMOTE helped the precision of the models significantly proving that skew data could hamper the performance of the model.
- Considering 4 different types of modals gives the model more accurate results and is in line with emerging research areas for Alzheimer's.
- We take into account all new areas of emerging research where it could help in better predictions.
- As per the results, Random Forest Classifier performs the best and is a good model when compared to other classification models.
- Even though the Long Short Term Memory Networks (LSTMs) are highly accurate, they were unable to give us good predictions. This is due to the data scarcity.

This paper provides a proof of concept of how Alzheimer's can be detected and also gives various insights about the shortcomings of the current approach. In future, we plan on gathering more data from the patients for training neural networks which can result in more accurate predictions. The current testing framework is such that a lot of patients do not get tested for all parameters. This should be taken care of with more precision so that the model has complete data. While this approach is exciting, a lack of a more structured database decreases its utility. We plan on reducing model granularity and including more features for a generalizable model. Having developed a model that predicts whether a patient may get affected by Alzheimer's, the next step would be to be able to determine exactly when a patient may get affected by the disease. This will provide great benefits in terms of relative preparedness that a patient could be provided with.

ACKNOWLEDGEMENTS

We would like to thank Professor Maria Lorena Carlo Unda and New York University for providing us with the opportunity to work and explore this line of project. We would also like to thank the doctors from Kasturba Medical College for providing us with this real world data of patients.

REFERENCES

[1] [n. d.]. 1.4. Support Vector Machines. https://scikit-learn.org/stable/modules/svm.html

[2] [n. d.]. Alzheimer's disease: Facts amp; figures. https://www.brightfocus.org/alzheimers/article/alzheimers-disease-facts-figures

[3] [n. d.]. Kasturba Hospital, Manipal. https://khmanipal.com/

[4] [n. d.]. Sklearn.ensemble.randomforestclassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[5] [n. d.]. Why get checked? https://www.alz.org/alzheimers-dementia/diagnosis/why-get-checked

[6] 2019. Introduction to SGD classifier - Michael Fuchs Python. https://michael-fuchs-python.netlify.app/2019/11/11/introduction-to-sgd-classifier/

[7] Harald Hampel, Katharina Buerger, Raymond Zinkowski, Stefan J. Teipel, Alexander Goernitz, Niels Andreasen, Magnus Sjoegren, John DeBernardis, Daniel Kerkman, Koichi Ishiguro, and et al. 2004. Measurement of phosphorylated tau epitopes in the differential diagnosisof alzheimer disease. *Archives of General Psychiatry* 61, 1 (2004), 95. https://doi.org/10.1001/archpsyc.61.1.95

[8] Fadi Thabtah, Robinson Spencer, and Yongsheng Ye. 2020. The correlation of everyday cognition test scores and the progression of alzheimer's disease: A Data Analytics study. *Health Information Science and Systems* 8, 1 (2020). https://doi.org/10.1007/s13755-020-00114-8