National College of Ireland

Classification of Alzheimers Disease Stages using Radiology Imaging and Longitudinal Clinical Data

MSc Research Project
Data Analytics

Piush Vaish
Student ID: x17122449

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Piush Vaish
Student ID:	x17122449
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	16/09/2019
Project Title:	Classification of Alzheimers Disease Stages using Radiology Imaging and Longitudinal Clinical Data
Word Count:	13634
Page Count:	47

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	7th March 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Classification of Alzheimers Disease Stages using Radiology Imaging and Longitudinal Clinical Data

Piush Vaish
x17122449

Abstract

Alzheimer's disease is an irreparable, degenerative disease with ongoing loss of functions of the brain. Currently, there is no medicine or treatment present to stop or slow down the progression. The identification of different stages for diagnosis require a combination of clinical data, complex cognitive tests, radiology imaging, demographic information, time and highly skilled physicians. Recent machine learning techniques can help to provide a process to extract insights and improve the quality of life for the patients and assist the physicians. In this project, various machine learning techniques such as feature selection, feature engineering, dealing with imbalanced data, imputation of missing values and standardization are applied. Multiple algorithms are also compared before performing random grid search to tune hyperparameters for classifiers and developing an ensemble learner to classify three clinical stages (normal, mild cognitive impairment and dementia). Cognitive tests, magnetic resonance imaging of left hippocampus and cortical thickness of right entorhinal are discovered to be important features for prediction. This finding is similar to that reported in a number of studies. The model can equally distinguish between a class and other classes with an average area under the receiver operating characteristics score of 0.83. This is within the range of the evaluation metric of existing state-of-the-art models. A web-based application is developed and deployed to the cloud to address the gap for the user to benefit from the developed model. This result in an end-to-end pipeline that will empower the user with a practical application and contribute to the active research in the area.

1 Introduction

Life expectancy among humans has increased globally through effective diagnosis and medicine. However, medicine or treatment cannot cure Alzheimer's disease. The disease affects more than 47 million people globally (Nanni et al., 2016). Furthermore, the people affected by the disease are increasing annually and deaths attributed to the disease have been continuously increasing while deaths from other diseases have been decreasing. For example, in the U.S.A., people affected by the disease are expected to rise to 13.8 million by 2050 from 5.4 million diagnosed in 2016. The rate of diagnosis is on top going to double from 66 seconds in 2016 to 33 seconds by 2050. In 2013, 84,767 deaths were due to Alzheimer's disease. Deaths due to chronic diseases such as stroke and heart disease decreased between 2000 and 2013 while deaths from the disease increased by 71% (Alzheimer's Association, 2016). In Ireland, the number of patients is expected to rise

to 150,000 by 2046 from 48,000 in 2011 (O'Kelly, 2016). It equally affects the health, financial security and time spent in love and care by the family members for the patient.

Alzheimer's disease is an unrecoverable disease with ongoing damage to memory and other cognitive functions. The changes in the brain happen years before the diagnosis of dementia. The pace is also not same among patients (Bilgel and Jedynak, 2018). Moreover, the specific cause is still unknown with only a few patients diagnosed in a correct and timely manner (Kruthika et al., 2019a). Hence, changes in biomarkers (biological feature used to measure the absence or presence/risk of developing the disease) and cognitive markers need to be discovered at the initial stage to intervene, manage symptoms and offer effective care. The active management of the disease leads to improved quality of life for both patient and their family members. It also increases coordination among the physicians and the patient.

1.1 Motivation and Background

Currently, there is no test available to diagnose Alzheimer's disease except for brain biopsy upon death. There is also no pharmacological treatment or medicine available which can stop or slow down the progression. A limited number of drug trials are successful because of the high cost of development of drugs and long observation time for the progression. The disease can progress rapidly but cannot be assigned to it only. The symptoms vary among patients with no clear reason as to why some people progress to advance stages. It is challenging to distinguish between age-related cognitive decline and other neurological disorders (Moscoso et al., 2019).

Alzheimer's disease is progressive, irreversible disease and leads to loss of functions of the brain. The decline occurs because the nerve cells affecting cognitive functions are either damaged or destroyed. The person's ability to sustain essential functions like walking, reasoning and swallowing are affected. Patients in the final stage need around the clock care and are bed-bound (Zhang et al., 2017). Examples of some of the typical symptoms are inability to perform routine tasks, unable to solve problems, confusion about time, place or relationship, poor judgement, misplacing things and unable to retrace the steps to recover the items. Additionally, the health care costs are greater than any other disease e.g., in the U.S.A., the total cost of care amounted to $259 billion in 2017. It is also estimated that $341,651 represent the cost of care for a patient in the last five years of their life (New York State Coordinating Council, 2017). Furthermore, 18.1 billion hours of care are contributed by approximately 15 million family members and other unpaid caregivers (Alzheimer's Association, 2016).

The techniques for diagnosis include gathering information about family and medical history, feedback from relatives or friends about the changes in skills or behavior, physical and cognitive tests, blood tests and brain imaging. The use of cognitive tests with magnetic resonance imaging (MRI) of the brain is the most popular method to identify the deterioration of the brain (Moscoso et al., 2019). The distinct stages of Alzheimer's disease are:

- Pre-clinical phase before the symptoms occur.
- Mild cognitive impairment (MCI) includes more cognitive decline than at the patient's age. It does not affect routine life.
- Dementia (Alzheimer's Association, 2016).

1.2 Research Question

Although identified 100 years ago, Alzheimer's disease is being identified as the primary cause of dementia and a major cause of death in the last 30 years. However, researchers believe that rapid detection remains the primary key that can help in the disease being prevented, slowed or stopped (Alzheimer's Association, 2016), (Zhang et al., 2017).

RQ: *"Can identification and classification of different Alzheimer's disease stages (normal, mild cognitive impairment, dementia) using several machine learning techniques and algorithms (logistic regression, support vector machine, ensemble of classifiers, etc.) help to improve research and support practitioners to provide early intervention and care to patients?*

Identifying changes at a developmental stage of the disease helps in managing patient's treatment and care. It also increases coordination among the physicians and the patient and improves quality of life through all stages of the disease for both patient and family members (Alzheimer's Association, 2016). Further, the area is being researched significantly and still many discoveries need to be done like the precise cause of biological changes, rate of progression and how it can be stopped or slowed down.

Web-based application helps access to the user to determine the stage of the disease given certain parameters.

Sub-RQ: *"Can a web-based application framework help enhance the user experience for the identified stages in the disease progression?"*

To solve the research question, the project implements and evaluates different machine learning techniques to find the best performing model which can classify the various stages of the disease. A model is additionally chosen to develop an open-source web-based application that can be used by anyone. The following objectives are implemented to obtain answers to the research question.

1.3 Research Objectives and Contribution

The first objective is a critical review of the literature on Alzheimer's disease between 2004 and 2019. The goal is to recognize the problem and identify the gaps. These gaps are addressed in the two tables as the objectives of the project to contribute to the research by replicating some of the literature, implementing machine learning techniques to develop a model for classifying the various stages of the disease i.e., normal, mild cognitive impairment or dementia. The evaluation metrics for measuring the performance are normalized confusion matrix, average multiclass AUROC (Area Under the Receiver Operating Characteristics) score, multiclass AUROC dictionary and AUROC curve. Furthermore, a model is used for developing a web-based application.

The main objectives stated in Table 1 are implementing, evaluating and reporting result of multiple machine learning algorithms on monthly changes in biomarkers. Furthermore, tree-based models are trained, evaluated and compared.

Table 1: Objectives

Objective	**Aim**	**Details**
2	Implement multiple machine learning algorithms on features	Build a classification model using monthly changes in radiology images and clinical data (section 4.7)
2(a)		Extract features to help identify and model the stages of progression (section 4.7)
2(b)		Implementation, evaluation and result of Logistic Regression (section 4.7.1)
2(c)		Implement, evaluate and result of Linear Discriminant Analysis (section 4.7.2)
2(d)		Implementation, evaluation and result of K-nearest Neighbors (section 4.7.3)
2(e)		Implementation, evaluation and result of Decision Tree (section 4.7.4)
2(f)		Implementation, evaluation and result of Random Forest (section 4.7.5)
2(g)		Implementation, evaluation and result of Gaussian Naive Bayes (section 4.7.6)
2(h)		Implementation, evaluation and result of Support Vector Machine (section 4.7.7)
2(i)		Implementation, evaluation and result of Neural Network (section 4.7.8)
2(j)		Comparison of developed models (section 4.7.9)
3	Train tree-based models using selected features. The output from the model is also explained	Build a classification model using tree-based algorithms (section 4.8)
3(a)		Implementation, evaluation and results of Decision Tree using different numbers of leaf nodes (section 4.8.1)
3(b)		Implementation, evaluation and result of Random Forest (section 4.8.2)
3(c)		Implementation, evaluation and result of XGBoost (section 4.8.3)
3(d)		Comparison of developed models (section 4.8.4)
3(e)		Interpreting machine learning model (section 4.8.5)

The main objectives stated in Table 2 are creating and deploying a web-based application, tuning the hyperparameters using random grid search, using different feature selection techniques and developing an ensemble of classifiers.

Table 2: Objectives (continued)

Objective	**Aim**	**Details**
4	Create a web-based application. The code is open-source and the application is made public so that anyone can use it	Create a web-based application using XGBoost (section 4.9)
4(a)		Implementation, evaluation and result of XGBoost (section 4.9.1)
4(b)		Interpreting machine learning model (section 4.9.2)
4(c)		Implementation, design and deployment of web-based application (section 4.9.3)
5	Improve robustness over a single model through ensemble learning	Fine tuning of an ensemble of classification models using random grid search (section 4.10)
5(a)		Run grid Search to find the most acceptable hyperparameters (section 4.10.1)
5(b)		Generate feature importance of the classifiers (section 4.10.2)
5(c)		Implementation, evaluation and result of ensemble of classifiers (section 4.10.3)
6	Find the important features from the data set using feature selection techniques	Use Feature Selection Techniques and Build an Ensemble of Classification Models (section 4.11)
6(a)		Feature selection using variance threshold, random forest and univariate selection (section 4.11.1)
6(b)		Use Principal Component Analysis to understand the variance explained (section 4.11.2)
6(c)		Implementation, evaluation and result of Decision Tree using 50 leaf nodes (section 4.11.3)
6(d)		Implementation, evaluation and result of Random Forest (section 4.11.4)
6(e)		Implementation, evaluation and result of XGBoost (section 4.11.5)
6(f)		Implementation, evaluation and result of Ensemble of Classifiers (section 4.11.6)
6(g)		Comparison of developed models (section 4.11.7)
6(h)		Interpreting machine learning model (section 4.11.8)
7		Comparison of the developed model with existing state-of-the-art models (section 5.1)

The major and minor contributions resulting from this research can be summarized as:

- A generalized model that handles radiology images and longitudinal clinical data

to distinguish different stages of Alzheimer's disease i.e., normal, mild cognitive impairment and dementia.

- The model handles an imbalanced data set and missing values to improve the performance.
- Chronological dependencies are modeled using an ensemble of Extreme Gradient Boosting (XGBoost) and other classifiers. To the best of the researcher's knowledge, this is the first time this technique has been applied using radiology images, characteristics of the patient and longitudinal clinical data.
- An end-to-end pipeline to develop a web-based application to predict the stages of the disease. It is a practical application of the process.

The scope for the project is the implementation, evaluation and presentation of the results for several machine learning techniques. The techniques are further investigated to determine the factors which contribute to the output of the model. After a thorough explanation, a model is used to build a web-based application. Certain machine learning techniques such as recurrent neural networks and convolution neural networks are out of the scope of this project because of the lack of access to graphics processing unit (GPU).

The report is divided into different chapters. Chapter 2 presents an investigation of existing literature in Alzheimer's disease, followed by chapter 3. It discusses the scientific methodology approach and architecture design, Chapter 4 states the implementation, evaluation and results of the different machine learning techniques and development of web-based application. Chapter 5 presents a discussion, followed by conclusion and acknowledgement as Chapter 6 and 7.

2 Literature Review of Alzheimers Disease Progression

2.1 Introduction

This section discusses the papers published between 2004 and 2019 regarding Alzheimer's disease progression. It starts by reviewing the application of machine learning on radiology images followed by different tests that are used to measure the clinical stage of the progression. Clinical data sets are also discussed to find the most suitable data set for the project. Various features, feature engineering and feature selection techniques used by the researchers are examined to gain knowledge. Finally, a critique of machine learning algorithms, techniques and evaluation metrics is conducted to identify the gaps. The key findings include Alzheimer's Disease Neuroimaging Initiative (ADNI) is the most common data set for longitudinal studies, biomarkers from neuroimaging are good to develop a model, multiple machine learning algorithms are applied to classify different stages of the disease. However, ensemble of Extreme Gradient Boosting (XGBoost) and other classifiers has not been applied. Further, none of the models developed are deployed as a web-based application.

2.1.1 A Review of Radiology Images and Identified Gaps

A recent review of more than 30 papers shows the use of machine learning on radiology images for determining and conversion to Alzheimer's disease (Nanni et al., 2016). The

techniques recognize and classify complex patterns from various images to carry out clinical decisions with comparable performance to human. Furthermore, the application of machine learning to radiology images is estimated to grow in the next 5 - 10 years because of the active research (Zhang and Sejdić, 2019). Different neuroimaging techniques help to provide the primary markers of the brain pathology (Masdeu et al., 2005). It is also demonstrated that early detection using MRI help in treatment that delays the progression of Alzheimer's disease (Lahmiri and Shmuel, 2018). MRI is a non-invasive and most sensitive imaging scan of the brain. It is employed to visualize the anatomical structure of the brain in a routine clinical environment. It also produces high spatial resolutions and image details (Wang et al., 2018). However, using only MRI can result in ignoring subtle abnormalities due to time. Hence, a longitudinal analysis of MRI is important (Cui and Liu, 2019). Moreover, existing models require assumptions regarding trajectories or discount relationship between patient's trajectories and multiple biomarkers. Only a few papers researched non-image data with image data to properly capture the interaction. The approaches equally failed to quantify the degree of abnormalities on a normal scale (Aditya and Pande, 2017) and ignored chronological dependencies.

2.1.2 A Review of Tests to Measure Alzheimers Disease Progression

This section discusses the different tests used to measure Alzheimer's disease progression to ensure the tests are recognized and then used as features.

Different tests act as tools to evaluate the progression. Tests use the state of the health, events like death and institutionalization and different mathematical approaches such as hazard ratios and probabilities. Some models e.g., Consortium to Establish a Registry for Alzheimer's Disease (CERAD), Fenn and Gray measure evolution between disease severity or changes in cognitive functions. Other models e.g., CERAD_Mini-Mental State Exam (MMSE) and Kinosian include placement in an institution or Assessment of Health Economics in Alzheimer's Disease (AHEAD) model consider full-time care. Statistical models in addition use a diverse range of measurement scales e.g. MMSE, Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-cog) demonstrates a natural progression in Alzheimer's disease.

ADAS-Cog uses multiple biomarkers like language and memory and is the most applied test (Skinner et al., 2012). However, cognition solely is not an indicator as it fails to consider the effect of healthcare and costs (Green et al., 2011). Papers use ADAS-Cog and other biomarkers to calculate the progression at each time point (K. Fisher et al., 2018), (Young et al., 2014a), (Yang et al., 2011). Other tests performed are MMSE (Doody et al., 2010), Global Staging Clinical Dementia Rating (CDR) (Wang et al., 2018) or own metrics from the age of the patient to measure the start of the symptoms of Alzheimer's disease (Bateman et al., 2012). However, these tests divide the progression from continuous to discrete stages and precise duration cannot be measured. To overcome the issue a paper uses the exponential-shaped trajectory of the ADAS-Cog score as a continuous factor (Schmidt-Richberg et al., 2016). In conclusion, there is no sole test to measure the progression. Tests can be performed in conjunction with other biomarkers to predict the progression across the timeline.

2.1.3 A Review of Data Sets Used

Longitudinal data evolve with time and have no shape. It is used to gain information regarding one-to-one change from successive two time-points. The researchers utilize a

variety of data for predicting the progression of Alzheimer's disease. Nevertheless, there is a need for longitudinal data upon which machine learning can be done (Fisher et al., 2018). Longitudinal data are also incomplete (Mehdipour Ghazi et al., 2019). The studies also research on a small subset of data leading to a less accurate characterization of stages than from large data sets (Goyal et al., 2018). The data used for research are as follows:

- Creating one's own customized cohort such as a study selected participants using neuropsychological evaluations (Pereira et al., 2017).
- Competitions like Computer-Aided Diagnosis of Dementia (CADDementia) or Kaggle (Vieira et al., 2017).
- Collaborating with a single facility such as The Baylor Alzheimer's Disease and Memory Disorders (Doody et al., 2010) or National Alzheimer's Coordinating Center (Bateman et al., 2012) or from UCSF Memory and Aging Center (Zhou et al., 2012) or multiple centers e.g., Coalition Against Major Diseases (CAMD) database (Fisher et al., 2018) or The Alzheimer's Disease Neuroimaging Initiative (ADNI) (Young et al., 2014a), (Goyal et al., 2018), (Venkatraghavan et al., 2019), (Kruthika et al., 2019b).

ADNI is the most common data set used for the longitudinal studies for tracking the disease progression (Zhang et al., 2017). Nevertheless, the data sets extracted from ADNI vary for separate studies with a varying number of patients and features. For example, a study uses longitudinal scans in cerebral cortex thickness that are 1 year apart for 132 patients with MCI (Lee et al., 2016) while other studies use T1-weighted MRI of 830 patients (Cui and Liu, 2019) or 445 patients with various stages of the disease (Wang et al., 2019).

There is a shortage of a common data set for determining the stages for Alzheimer's disease. The multi-sharing initiative and the TADPOLE[1] challenge can provide a large enough sample size and a common data set to gain a piece of consistent information.

2.1.4 An Investigation of Features, Feature Engineering and Feature Selection

This section discusses the various features, feature engineering and feature selection techniques that are used by different papers. The aim is to develop a knowledge of the features which can be used to develop the model.

MRI provides information about the brain tissues e.g., grey and white matter in a non-invasive manner. Studies utilize features from MRI such as brain matter and specific thinning of cortex (Aditya and Pande, 2017) or "volumes of ventricles, hippocampus, whole brain, fusiform, middle temporal gyrus, and entorhinal cortex" (Mehdipour Ghazi et al., 2019) to build models. Another study suggests using a measurement of cerebrospinal fluid, amyloid and tau, neural injury, etc. using MRI or positron emission tomography (PET) images (Nanni et al., 2016).

However, there is a difference in MRI, computed tomography (CT) and PET images. For example, a white area in MRI is the subcutaneous fat while it is the skull in CT images. PET images demonstrate both the biochemical and physiological changes while MRI and CT capture anatomical changes. Hence, it is significant to be aware of these differences while using these scans. The features are not limited to color and shape. Gabor filter

[1] `https://tadpole.grand-challenge.org/`

remains the most common method to extract features from the medical images. Further, the quality of the medical images needs to be good to extract the features. Transfer learning is, in addition, an effective technique which can reduce bias among equipment while producing an image. It is equally possible to have abnormalities from the same data source with various scenarios. It results in imbalanced data and the way to handle the imbalance is still an open research issue (Zhang and Sejdić, 2019). A study states that changes in specific nerve cells i.e., N-methyl-D-aspartate receptor results in the severity of the disease (Mishizen-Eberz et al., 2004). Another paper states that deterioration of hippocampus and entorhinal can establish the onset of dementia. It uses data from ADNI1 and ADNI2 and informs that both data sets used MRI scanners with different field strength. Hence, it combined both data sets and affected their study (Moscoso et al., 2019).

Papers also suggest the use of other features in addition to radiology imaging e.g., patient's age, age of the patient's parents, short term follow-up data (Bilgel and Jedynak, 2018), education and socio-economic status (Wang et al., 2019). A study uses "fractals from MRI of cerebral cortex, cortical thickness, gyrification index and ADAS-Cog test scores" to distinguish between healthy and patients with Alzheimer's disease (Lahmiri and Shmuel, 2018). Another paper uses biomarkers such as genetics, cognitive measurements in conjunction with the cerebral cortex. The cerebral cortex inter-connects the cerebral hemisphere and longitudinal structural callosal changes extracted from MRI help to determine conversion from mild cognitive impairment (MCI) to dementia (Lee et al., 2016). Nevertheless, it is a challenge to associate imaging features with the static feature at multiple time points.

A study addresses the association of feature selection on single-task learning and multi-task learning. In single-task learning, progression is estimated separately at different time points while multi-task learning focuses on multiple related tasks. It finds that single task learning is suboptimal in predicting progression because each task is treated separately. In multi-task learning, the various tasks share a subset of features and each task is related equally. However, the learning results in sparse data (Goyal et al., 2018).

In conclusion, features from radiology imaging such as the hippocampus and cerebral cortex are good to build a model in addition to clinical data and information regarding the patient such as age, socio-economic status and cognitive score.

2.2 A Critique of Machine Learning Algorithms, Techniques and Evaluation Metrics Used and Identified Gaps

There are multiple algorithms or methods used to extract patterns with a correlation between the stages of Alzheimer's disease and features. This section discusses some of the techniques to critique, compare and evaluate them. The aim is to gain an understanding to reproduce the techniques and identify gaps.

A study uses logistic regression and a defined common template for binary classification of a patient having the disease or not. Logistic regression is a simple, linear model and provides co-efficient weights to localize the deformations related to the disease for each voxel. The study applies both kinds of regularization, least absolute shrinkage and selection operator (LASSO) and Ridge, to handle a wide data set with more features than the number of observations (Fiot et al., 2014). Additionally, another paper uses logistic regression model with fused LASSO regularization to predict the annual changes in callosal thickness. It states that the gender of the patient influences the accuracy of the

prediction. The prediction is 84% accurate in females and 61% in males. Furthermore, the annual changes in callosal atrophy predict conversion from MCI to dementia in females more accurately than males. The use of MMSE, ADAS or Rey Auditory Verbal Learning Test (RAVLT) at baseline did not help the prediction (Lee et al., 2016). However, the study is limited to the data set and does not ensure the patient in a group of normal or MCI will not convert to dementia after the follow-up period. Survival analysis may benefit the study. Multitask exclusive learning is applied by another paper to predict markers for the disease. It utilizes information from adjacent time points to understand the intrinsic relationship among multiple cognitive measures without knowing them in advance. Least square regression with LASSO regularization is applied to the data from each time point to accurately identify image markers. It states certain biomarkers are highly associated with MMSE scores at multiple data points (Wang et al., 2019). However, the size of the participants is limited and the study does not have complete information for each patient. The model is excellent on MRI data and other types of radiology imaging e.g. PET and CSF can improve the performance.

Event-based modelling is equally suitable to understand the dynamics of progression. A study employs a discriminative approach to estimate an ordering of events for each subject and a central ordering for all subjects to create a longitudinal timeline (Venkatraghavan et al., 2019). Still, short term trajectories for imaging and non-imaging parameters after disease progression are important to consider. Another paper develops a multivariate Bayesian model and a quantitative template to compute trajectories as a function of time and measure the similarity between longitudinal biomarkers. It prepares the model after aligning short term longitudinal data and estimates quantitative template for various stages. It can learn long-term trajectories with mean error in the onset of dementia to less than 1.5 years from short-term clinical data and known risk factors (Bilgel and Jedynak, 2018). However, it assumes a single pathway of biomarker changes and finds it difficult to generalize as each patient's biomarkers are different.

A study uses non-image data to quantify abnormality and explore inter-feature relationships using similarity indices. It builds a reference knowledge base for normal patients and those with dementia to contrast a patient and label normal and dementia according to the affinity to each class. Multifactor affiliation analysis is used to compare the feature value of training subjects and quantify the severity of the disease (Aditya and Pande, 2017). However, the technique measures quantified distances and is suitable for numerical data only. Support vector machine (SVM) is also used to detect the disease in computer-aided-diagnosis systems. The kernel used for SVM is linear because it provides co-efficient and is simpler than non-linear SVM (Lahmiri and Shmuel, 2018). Nevertheless, SVM is difficult to modify and add spatial regularization. Another paper uses an ensemble SVM to find that it performs better than standalone SVM on five different data sets. Furthermore, certain feature selection approaches work differently for different data (Nanni et al., 2016). Although it needs to validate the findings by testing on unseen data and try new techniques for building an ensemble.

Some papers experiment with different machine learning algorithms to find the best model. For example, four algorithms (linear discriminant analysis (LDA), k-nearest neighbors algorithm (kNN), naive Bayes (NB) and second-order polynomial SVM) are used to classify normal or demented patients (Lahmiri and Shmuel, 2018) or learning algorithms such as SVM, random forest, regression and neural networks (NN) to predict the stages of the disease (Zhang and Sejdić, 2019).

Additionally, a multistage classifier using multiple machine learning methods e.g., NB,

SVM and NN is applied to classify Alzheimer's disease more efficiently and effectively by a study (Kruthika et al., 2019a). It uses correlation to eliminate redundant features and focuses on improving image retrieval from the smallest number of features by using transfer learning and capsule networks on MRI. Capsule networks require a limited set for training and have a lower learning curve. The main feature is the extraction of subtle changes in texture and contour of the hippocampus employing diverse techniques. It states that capsule nets provide better accuracy (82.45%) to estimate the stages than other NN architectures (Kruthika et al., 2019b). However, the prediction has more missing results than other methods and the use of more capsule layer can improve the accuracy. The fault alarm rate for the disease is equally smaller than other classes and requires advanced biomarkers and biochemical information to get better performance.

A study uses a modified Long short-term memory (LSTM) model to address the issue of disease progression models neglecting chronological dependencies of multiple biomarkers and making assumptions about patient's trajectories. It additionally uses LDA classifier and gets Area Under the Receiver Operating Characteristics (AUROC) score of 0.90 (Mehdipour Ghazi et al., 2019). Another study uses a combination of a convolution neural network (CNN) and bidirectional gated recurrent unit (BGRU) for longitudinal analysis of MRI images. CNN learns the spatial features of MRI for classification and three cascaded BGRU are trained on the output from CNN at multiple time points for extracting longitudinal features. BGRU also handles the issue of incomplete longitudinal data through processing varying length image sequences. The architecture achieves an accuracy of 91.33% for Alzheimer's disease vs. normal patients (NC) and 71.71% for progressive MCI subjects (pMCI) vs. stable MCI subjects (sMCI) (Cui and Liu, 2019). In addition, it finds the use of high-dimensional and longitudinal data challenging.

In summary, the area is highly researched and multiple machine learning techniques are used to model the disease progression. However, there seems to be no evidence of the application of one of a popular technique called Extreme Gradient Boosting (XGBoost) or an ensemble of classifiers including XGBoost. Some of the studies reviewed use accuracy as a metric. As discussed in the section 2.1.4, the data present for predicting the progression of the disease is generally imbalanced and accuracy is not a good metric for imbalanced data. There is also a lack of literature utilizing novel techniques like Shapley Additive Explanations[2] (SHAP) to explain the output of the model in an interpretable manner. This project implements techniques to handle an imbalanced data and develop ensemble learning models. It uses normalized confusion matrix, average multiclass AUROC score, multiclass AUROC score dictionary and AUROC Curve as metrics to measure the performance of the model and SHAP for explaining the output of the model.

2.3 Comparison of Features, Algorithms and Results from Literature Review

Features, machine learning techniques and results of some of the literature reviewed in section 2.2 are summarized in Table 3. The studies primarily use MRI biomarkers along with other features such as clinical, laboratory, genetic and demographic data (Goyal et al., 2018) to develop models using different machine learning algorithms e.g., hidden Markov model, L2-regularized logistic regression (Goyal et al., 2018) or support vector machine (SVM) (Kruthika et al., 2019a). The studies also implemented different ways to

[2] `https://github.com/slundberg/shap`

measure the performance of the models e.g. AUROC score and accuracy.

Table 3: Comparison from Literature Review

Features	**Techniques**	**Results**	**Authors**
Clinical, imaging, laboratory, genetic, and demographic data	Hidden Markov Model, L2-regularized logistic regression	AUROC score = 0.85	(Goyal et al., 2018)
MRI biomarkers	LSTM + LDA classifier	AUROC score = 0.90	(Mehdipour Ghazi et al., 2019)
MRI biomarkers	SVM + kNN	Accuracy = 89.22	Kruthika et al. (2019a)
MRI biomarkers	Logistic Regression	AUROC score = 0.84	(Moscoso et al., 2019)

2.4 A Review of Web-based Application and Identified Gaps

A review of the literature and web search shows that there is a huge research gap for the end-user to gain benefits from the developed models through a web-based application. Only one study is discovered that developed a web-based application for Alzheimer's disease prevalence model for the state of Maryland, USA. The application enables the user to define the parameters and any interventions to create a burden projection for each calendar year until 2050. It also calculates the costs of the potential interventions that may either reduce or slow the progression (Colantuoni et al., 2010). There are far more web-based applications being developed for other diseases e.g. monitor patients with Parkinson's disease remotely and support decision making for practitioners (Patel et al., 2010), (Memedi et al., 2011). The web-oriented expert system is also being used in other domains such as predicting results for hurdle races (Przednowek et al., 2018) or measurement of chemical toxicity (Alves et al., 2018).

A huge amount of research is being done to classify the disease stages. Nevertheless, none of the reviewed literature has deployed the models to the web to enable the end-user to access them easily. The project is developing a user-friendly web-based interface as a proof of concept and can be used to classify the stages of the disease based on a few parameters. It is to develop a social cause and benefit the research in the future.

2.5 Conclusion

The literature review identifies the gap and a need to further research the progression and answer the research question (section 1.2). The review also highlights a need for the development of a web-based application and hence answers the sub-research question (section 1.2). The following chapter presents a scientific methodology and the architecture design to successfully complete the project.

3 Scientific Methodology and Architecture Design

3.1 Introduction

Alzheimer's disease typically progresses gradually in stages namely normal, mild cognitive impairment (MCI) and dementia. The project represents an end-to-end functional approach to implement machine learning techniques and develop a web application. Alzheimer's disease methodology (an adaptation of Knowledge Discovery in Databases (KDD) process) is applied to complete the project. The project's architecture design is two-tier.

3.2 Alzheimers Disease Progression Methodology

The analysis and modelling are done using an adapted KDD process (Fawcett, 2005). The methodology (Figure 1) involves gathering data from ADNI, followed by data preparation which involves cleaning, handling missing values, removing duplicates and creating new features.

Analysis and visualization are also done to understand the data set. There are many columns and few rows and therefore feature selection is done to prevent the curse of dimensionality. The data set with selected features is divided into training and test set. Machine learning algorithms are applied on training data set to develop models and tested on the test data set to ascertain the generalizability. The patterns are identified to represent knowledge and the relationship between features. The models are evaluated using different metrics e.g., normalized confusion matrix.

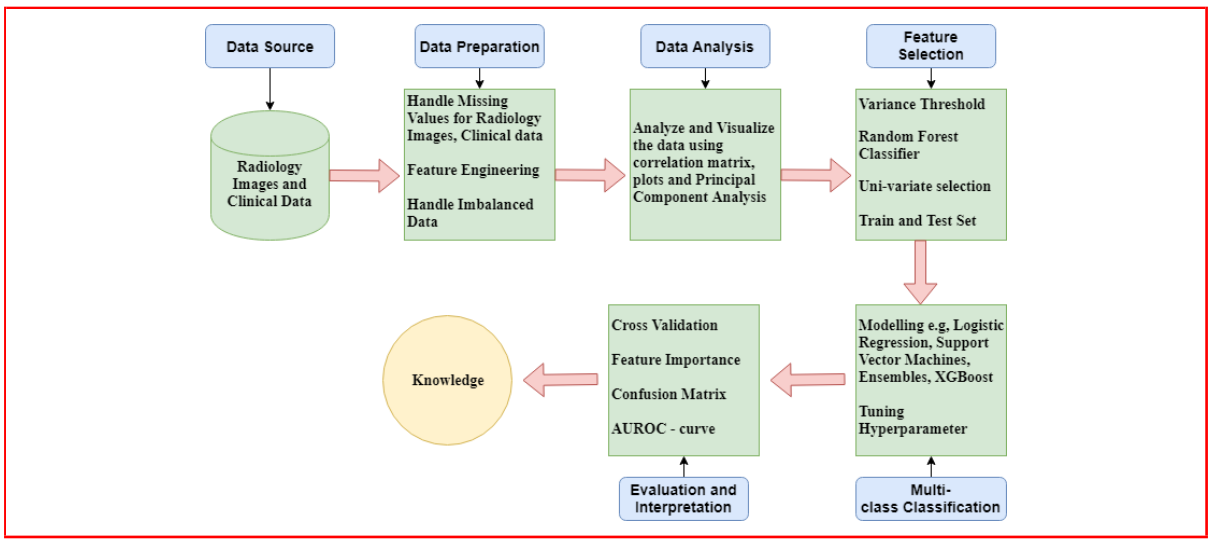

Figure 1: Alzheimer's Disease Progression Methodology

3.3 Architecture Design Specification

The architecture design for the project (Figure 2) is two-tier and consists of:

1. Client tier. Data Visualization and explanation of the output of the machine learning models are presented using Jupyter notebooks and the predictions from the model are displayed using a web-based application. The development involves the creation

of the architecture and framework, defining the input features that feed into a model to return the prediction values. The application is published with all the functionality to a web service to enable a user to access it. The interface is aimed to be simple and easy to use.

2. Business logic tier. It involves loading the data from different files, feature extraction, transformation and selection to train the models and comparing these models to obtain the most efficient model that predicts the different stages of Alzheimer's disease.

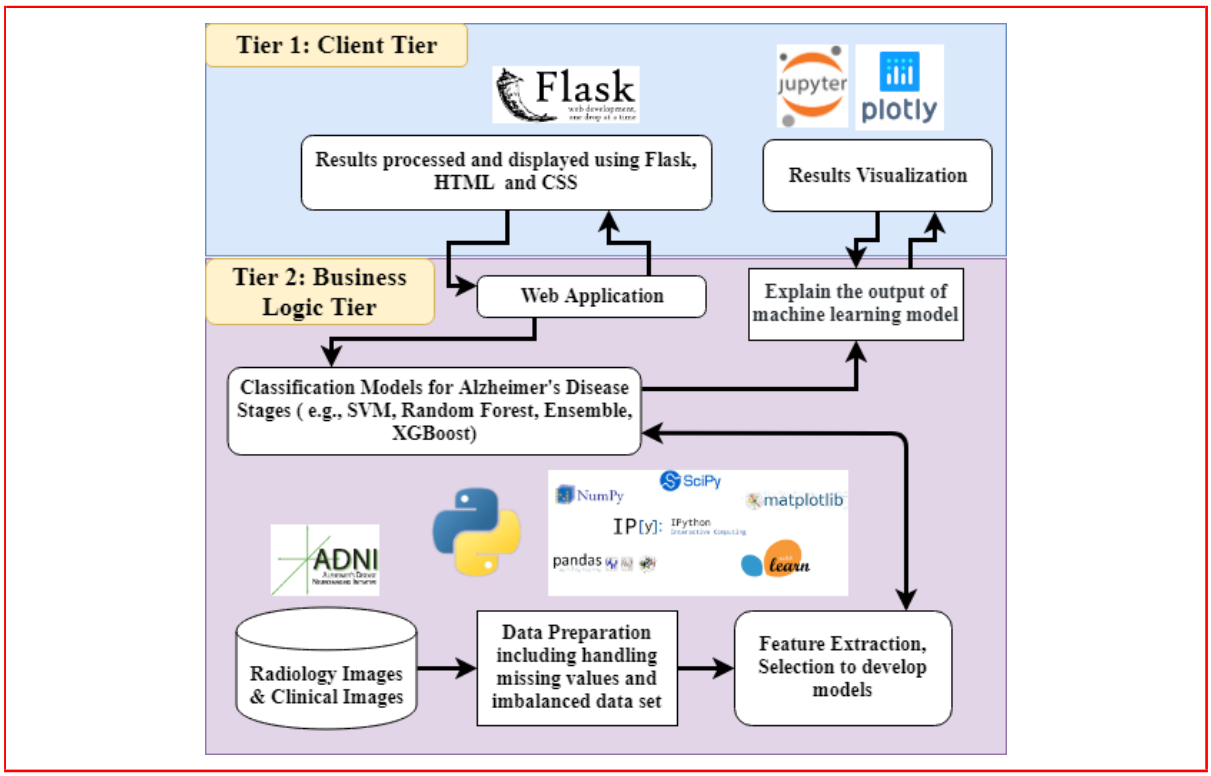

Figure 2: Architecture Design for Alzheimer's Disease Progression

3.4 Conclusion

Alzheimer's disease methodology and architecture design are enough to successfully deliver the project. The following section includes implementation, evaluation and results for classification models using selected features, multiple machine learning techniques and metrics. A model is also trained for developing the web-based application and answer the research questions as stated in section 1.2.

4 Implementation, Evaluation and Results of Alzheimers Disease Progression Models and Development of Web-based Application

4.1 Introduction

There are few conditions and perquisites for building classification models e.g., sample size of data, type of data, correlation, number of classes and type of problem. The studies suggest building multiple models and select the model with the best performance (Kruthika et al., 2019b), (Goyal et al., 2018). This section provides a short description about the data set which consists of subjects from the U.S.A. and Canada before detailing preparation for modelling. The multiple clinical stages are combined into three stages i.e, normal, mild cognitive impairment (MCI) and dementia. Machine learning techniques are applied to handle data imbalance, missing and high magnitudes values. Feature selection, feature engineering e.g., monthly changes in specific biomarkers, random grid search to tune hyperparameters and multiple algorithms are implemented before evaluating using normalized confusion matrix, AUROC (Area Under the Receiver Operating Characteristics) score and AUROC curve. AUROC can also be written as AUC (Area Under The Curve)-ROC (Receiver Operating Characteristics). The output from the models is explained and a web-based application is developed and deployed to the cloud.

4.2 Tools and Languages

The project is implemented using the following hardware:

- Central Processing Unit (CPU) - Intel(R) Core (TM) i7-6700HQ CPU @ 2.60GHz
- RAM - 64 GB
- Cores 4
- Logical Processors - 8
- Operating System - Microsoft Windows 10

The project is completed using Python. The web-based application is developed using Hypertext Markup Language (HTML) and Cascading Style Sheets (CSS), Flask[3] framework and deployed using the Heroku[4] platform.

4.3 Description and Analysis of the Data Acquired

The project utilizes data from The Alzheimer's disease Prediction of Longitudinal Evolution (TADPOLE)[5] challenge to predict the future outcome of clinical stages (normal, MCI or dementia) for a patient. The data set is gathered from the Alzheimers Disease Neuroimaging Initiative (ADNI). ADNI was launched in 2003 and so far, there have been three stages for data collection namely ADNI followed by ADNI-GO and ADNI-2. The

[3] https://www.fullstackpython.com/flask.html
[4] https://www.heroku.com/
[5] https://tadpole.grand-challenge.org/Data/

subjects are recruited from over 50 sites across the U.S.A. and Canada to help research the application of neuroimaging e.g., magnetic resonance imaging (MRI) and other biomarkers. The code to generate the standard data set is openly available in a GitHub[6] repository.

The data collected from ADNI are merged into a comma-separated values (CSV) file and consists of cerebrospinal fluid (CSF) markers of amyloid-beta and tau deposition, different types of radiological images such as MRI, positron emission tomography (PET) and diffusion tensor imaging (DTI), cognitive tests e.g., The Alzheimer's Disease Assessment Scale (ADAS), the Mini-Mental State Examination (MMSE) acquired in the presence of a clinical expert, genetic information such as apolipoprotein E4 (APOE4) status and general information such as age, gender and education, depression and head injuries. Each row in the data represents a visit of the subject and each column either represents the information about the subject or the biomarker from the visit. The duplicated rows are removed to include recent information. There is a total of 1737 subjects of which 957 are males and 780 are females. Subjects with APOE4 values of zero, one and two are 522, 298, 62 respectively. 204, 171 and 128 subjects have 16, 18 and 20 years of education and have different relationship status e.g., 653 subjects are married, 33 subjects never married and 99 subjects are widowed.

4.4 Data Preparation

The project is a multiclass classification problem and therefore presents different challenges from binary classification. It also has missing data and the classes are not distributed equally. The following section discusses the common approach used before modelling to answer the research question (section 1.2).

4.4.1 Combining Multiple Clinical Stages to Three Classes

The data set includes different clinical stages in addition to normal, MCI and dementia. There are also missing values for the clinical stage due to different reasons e.g., subject not returning for another examination. 3,837 rows are not selected because the stage of the disease is not recorded. The distribution of the stages is stated in Table 4. MCI is most common stage followed by normal in the data set.

Table 4: Count of Different Clinical Stages

Clinical Stages	**Count**
Normal	2668
Mild Cognitive Impairment	3932
Dementia	1732
Mild Cognitive Impairment to Dementia	372
Normal to Mild Cognitive Impairment	108
Mild Cognitive Impairment to Normal	77
Dementia to Mild Cognitive Impairment	12
Normal to Dementia	3

The values for “normal to mild cognitive impairment” and “dementia to mild cognitive impairment” are replaced by “mild cognitive impairment”. Similarly, the values for “mild cognitive impairment to dementia” and “normal to dementia” are substituted with

[6] `https://github.com/noxtoby/TADPOLE`

"dementia", and "mild cognitive impairment to normal" is replaced with "normal". The count of clinical stages after renaming is shown in Figure 3. MCI is nevertheless the most common stage followed by normal.

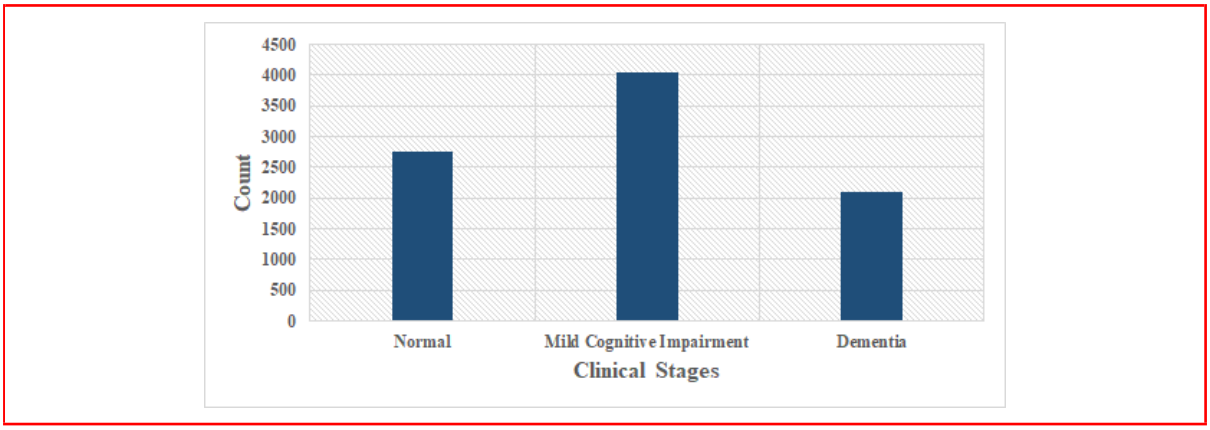

Figure 3: Count of Clinical Stages

Dealing with Imbalanced Data

Figure 3 shows an imbalanced data. Each class is not equally represented and hence certain machine learning techniques are difficult to implement. It is also an open research area (section 2.1.4). Synthetic Minority Over-sampling Technique (SMOTE) (Lemaitre G. et al., 2017) is applied to handle class imbalance. SMOTE[7] implements a nearest-neighbor algorithm to generate new synthetic data for the training set. The new samples are not generated for test data set to ensure the model generalizes well.

Dealing with Missing Data

Missing values is a common issue when working with longitudinal data (Mehdipour Ghazi et al., 2019) and results in an error from scikit-learn estimators as most of the algorithms expect numeric values. Incomplete rows/columns are deleted or the missing values are imputed from the known values of the data to handle the problem. SimpleImputer[8] is applied to utilize the strategy of mean or most frequent value of the column in which missing values are present.

Dealing with Features Varying in Order of Magnitude

Features with high magnitude bear more weight than features with low magnitude. Feature scaling is performed to normalize the range of features implementing standardization to prevent the influence of variation on machine learning algorithms. StandardScaler[9] from scikit-learn library replaces the values with their Z-score and the features with a mean of zero and standard deviation of one.

$$x' = \frac{x - \hat{x}}{\sigma}$$

x' is standardization, x is the original feature vector, $\hat{x}$ is the mean of the feature vector and σ is the standard deviation.

[7] `https://imbalanced-learn.readthedocs.io/en/stable/`

[8] `https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html`

[9] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

Applying One-vs-the-rest Strategy on Multiclass Classification

The strategy adopted for multiclass classification is different from binary classification. Multiclass assumes that each sample is assigned to one and only one class e.g. a stage can either be MCI or dementia but not both at the same time. One-vs-the-rest[10] (OvR) strategy is applied and fits one machine learning algorithm per class and a class is fitted against all other classes for each algorithm. The advantages include interpretability, efficient computation and gain information by inspecting the corresponding classifier.

4.5 Interpreting Machine Learning Model

Modelling is built upon the principle that minor differences can be exploited to discover patterns to determine the various classes. However, the models developed in an ideal situation do not produce the same real-life results (Zhang et al., 2017). Models with good performance e.g., ensemble of classifiers are often complex and difficult to explain. Further, a proper understanding of the reasons why a model makes a prediction is crucial to gain confidence of humans. A framework called Shapley Additive Explanations[11] (SHAP) is applied for interpreting model output. SHAP assigns each feature an importance value for a prediction based on cooperative game theory (Lundberg and Lee, 2017). Shapley value is defined as the average of marginal contribution for a feature across all the possible combinations.

4.6 Metrics to Evaluate the Performance of Machine Learning Algorithm

The project is a multiclass classification problem with an imbalance data set. Therefore, accuracy is not a proper evaluation metric. Confusion matrix is a better evaluation technique to summarize the performance of a model that classifies different classes. Metrics used to evaluate are used by other studies e.g., (Mehdipour Ghazi et al., 2019), (Cui and Liu, 2019) and include normalized confusion matrix, average multiclass AUROC (Area Under the Receiver Operating Characteristics) score, multiclass AUROC score dictionary and AUROC curve. Receiver operating characteristics is a probability curve and area under the curve tells measure of separability.

Confusion matrix (Figure 4) is a table to describe the performance of a classification model. It[12] is good for calculating other metrics like F1-score and AUROC curve for classification problems. It can also be normalized so that the numbers are between zero and one. It enables to obtain percentage of correctly classified samples.

[10] `https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html`

[11] `https://github.com/slundberg/shap`

[12] `https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62`

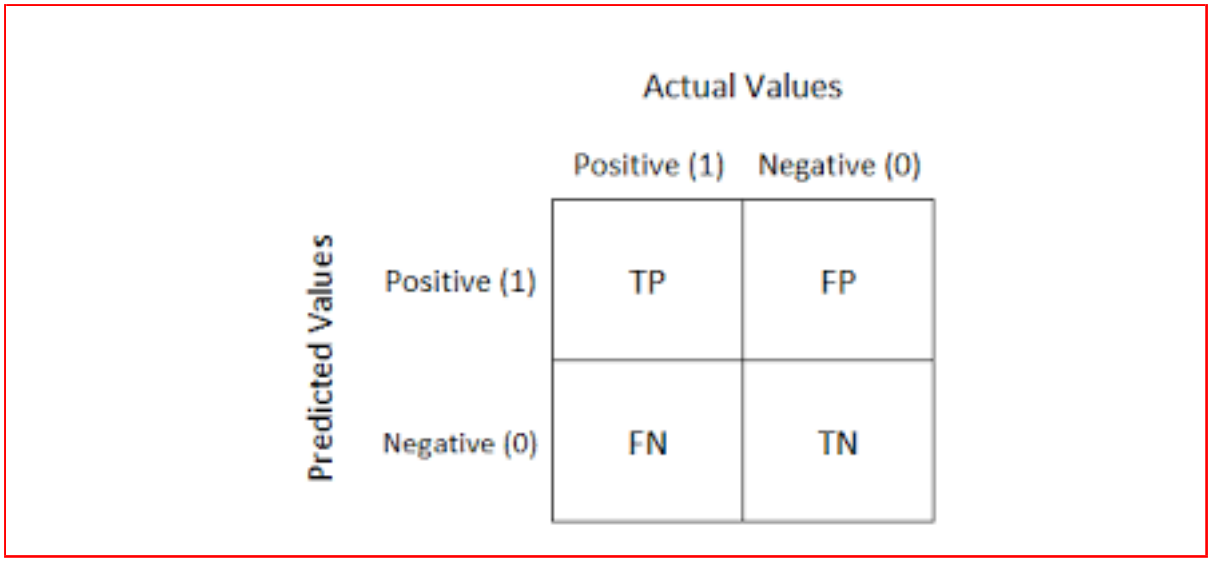

Figure 4: Confusion Matrix

where

- TP (true positives): Cases that are predicted as true and are actually true
- TN (true negatives): Cases that are predicted as false and are actually false
- FP (false positives): Cases that are predicted as true but are actually false
- FN (false negatives): Cases that are predicted as false but are actually true

$$\hat{A}(c_i|c_j) = (S_i - n_i(n_i + 1)/2)/(n_i n_j)$$

is the equation for AUROC score for one class (c_i) against class (c_j) where n is the "number of points belonging to each class and S_i is the sum of ranks of the class i test points after ranking all the class i and j data points in increasing likelihood of belonging to class i" (Azvan et al., 2018), (Mehdipour Ghazi et al., 2019). To use the metrics for multiclass, "micro" averaging or "macro" averaging can be used after implementing One-vs-the-rest strategy. The project uses "macro" averaging because it treats all classes equally by calculating the the metric independently for each class and then taking an average. "Macro" averaging[13] is:

$$PRE_{macro} = \frac{PRE_1 + ... + PRE_k}{k}$$

where PRE is performance of each individual class. Average AUROC score is after using "macro" averaging.

AUROC curve is a 2D graph where the x-axis is the measure of the true positive rate (TPR) or recall[14] while the y-axis is a measure of false positive rate (FPR) or miss rate.

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}$$

[13] https://sebastianraschka.com/faq/docs/multiclass-metric.html
[14] https://en.wikipedia.org/wiki/Sensitivity_and_specificity

The model with high performance has higher AUROC score with top left corner of the plot as an ideal point. It is where FPR is zero and TPR is one. It is also ideal to maximize TPR while minimizing FPR. It is commonly used to visualize the performance of the binary classifier. The project is multiclass classification problem and hence it is necessary to binarize the output. One AUROC curve is produced per class [15].

The challenge provides longitudinal data to train called "D1" and test the predictions on another longitudinal data called "D2". "D2" includes subjects who rolled over from previous ADNI studies to prospective ADNI-3 study. (Azvan et al., 2018).

4.7 Build a Classification Model using Monthly Changes in Radiology Imaging and Clinical Data

Multiple studies such as (Lahmiri and Shmuel, 2018) and (Zhang and Sejdić, 2019) typically apply various machine learning algorithms to find the best model which correctly classifies stages of Alzheimer's disease. In this implementation, features are selected based on two papers ((Goyal et al., 2018) and (Kruthika et al., 2019a)) which are analyzed in section 2.1.4. The main categories of features are radiology images e.g., MRI, PET, CSF measure and clinical data e.g., cognitive tests and information about patient. New features are created to determine the monthly change in the selected features and are based on the literature (Young et al., 2014b) which also used changes in biomarkers across time for modelling. Missing values are replaced with 0.0 and values are scaled. The categorical variables e.g., gender, marital status are converted into dummy variables. The data is then divided into training and test data sets. The models are from scikit-learn[16] library and trained using new features and dummy variables. Machine learning algorithms use OvR strategy to train and test data set is used for evaluation.

4.7.1 Implementation, Evaluation and Result of Logistic Regression

Logistic regression is implemented to calculate the probability of an event occurrence for categorical target variables. It is implemented using scikit-learn library and the function is LogisticRegression()[17]. The multiclass option is set to one-vs-rest (OvR). Logistic regression implementation from the library uses LASSO regularization as used by the study (Lee et al., 2016). The model resulted in average AUROC score of 0.595.

4.7.2 Implementation, Evaluation and Result of Linear Discriminant Analysis

Linear discriminant analysis estimates the probability of an input belonging to every class and is used by a paper (Mehdipour Ghazi et al., 2019). It is implemented using scikit-learn library and the function is LinearDiscriminantAnalysis()[18]. Singular value decomposition is applied as the solver because it does not compute the co-variance matrix. The model resulted in average AUROC score of 0.593.

[15] https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

[16] https://scikit-learn.org/stable/supervised_learning.html

[17] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[18] https://scikit-learn.org/stable/modules/lda_qda.html

4.7.3 Implementation, Evaluation and Result of K-nearest Neighbours

K-nearest neighbours assign an input to the class of which it is the nearest neighbor. It is implemented using scikit-learn library and the function is KNeighborsClassifier()[19]. It uses five nearest neighbours and all points are weighted equally in each neighborhood. The model resulted in average AUROC score of 0.612.

4.7.4 Implementation, Evaluation and Result of Decision Tree

Decision tree split the data into cascading questions based on significant contrast in input. It is implemented using scikit-learn library and the function is DecisionTreeClassifier( )[20]. "Best" strategy is adopted to select the split at each node and "gini" represent the function to measure the quality of split. The model resulted in average AUROC score of 0.651.

4.7.5 Implementation, Evaluation and Result of Random Forest

Random forest is an ensemble of a decision tree. It is implemented using scikit-learn library. The function used to implement is RandomForestClassifier()[21]. Number of trees is 10 and the maximum depth of the tree is set to 1. The model resulted in average AUROC score of 0.581.

4.7.6 Implementation, Evaluation and Result of Gaussian Naive Bayes

Naive Bayes predicts a class for the input using conditional probability. Gaussian naive Bayes is implemented because some features are continuous and categorical values are present in the data. It is implemented using scikit-learn library. The function used to implement is GaussianNB()[22] and uses default parameters. The model resulted in average AUROC score of 0.524.

4.7.7 Implementation, Evaluation and Result of Support Vector Machine

Support vector machine (SVM) finds a hyperplane before dividing into classes. It is implemented using scikit-learn library and the function is SVC()[23]. It uses default parameters from the library. The model resulted in average AUROC score of 0.60.

Figure 5 is a normalized confusion matrix for SVM. It is a visual representation in which the row denotes an instance of true class whereas column denotes instance of predicted class. It measures the performance over a fixed threshold. The values of the diagonal elements denote the degree of correctly predicted class i.e., 0.72 for normal (NL), 0.16 for MCI and 0.59 for dementia. The off-diagonal elements are mistakenly confused with the other classes. SVM is not an appropriate model for the selected features because the diagonal values are low when the threshold for the classifier is fixed at 0.5.

[19] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[20] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[21] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[22] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

[23] https://scikit-learn.org/stable/modules/svm.html

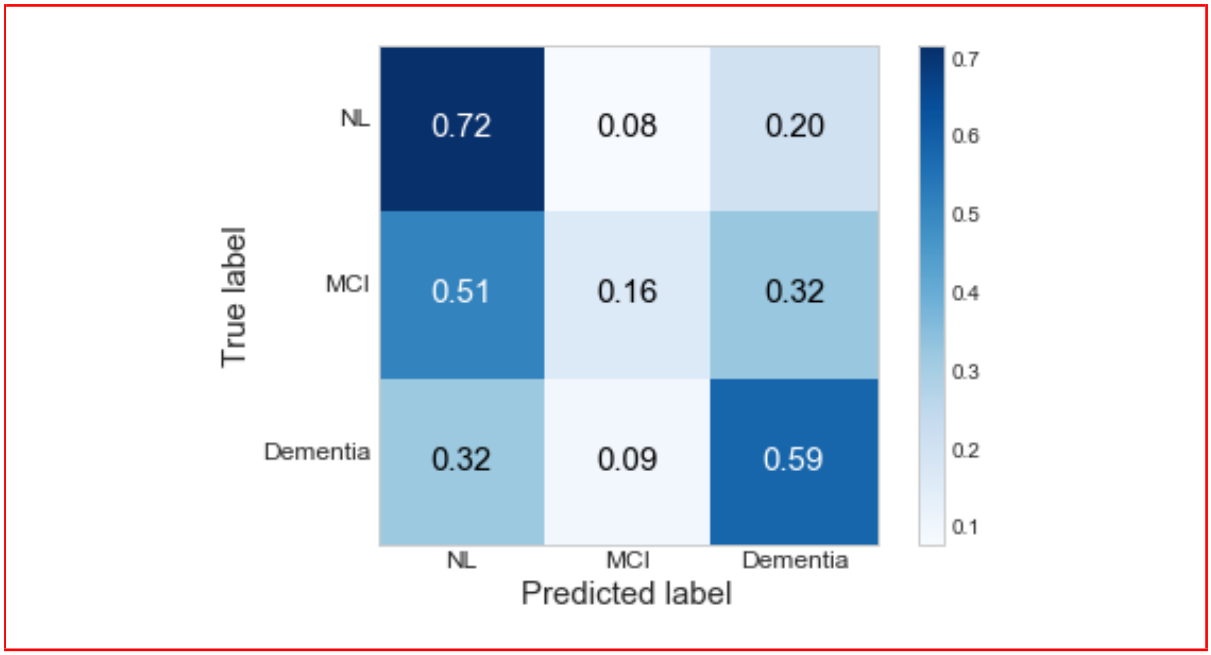

Figure 5: Normalized Confusion Matrix for SVM

Figure 6 is AUROC curve for SVM and shows the measure of the random positive sample ranked before random negative sample. A random model obtains an AUROC score of 0.5 and hence the classifiers should perform better than 0.5. It helps to measure the performance of the model without fixing the threshold. It plots a point for every possible threshold and is helpful to select the threshold of the model depending on the use case. SVM is better in predicting dementia with AUROC score of 0.66 against normal and MCI, normal with AUROC score of 0.64 against MCI and dementia when the threshold is unfixed. However, it is only little better than random method to predict MCI against normal and dementia because AUROC score is 0.54.

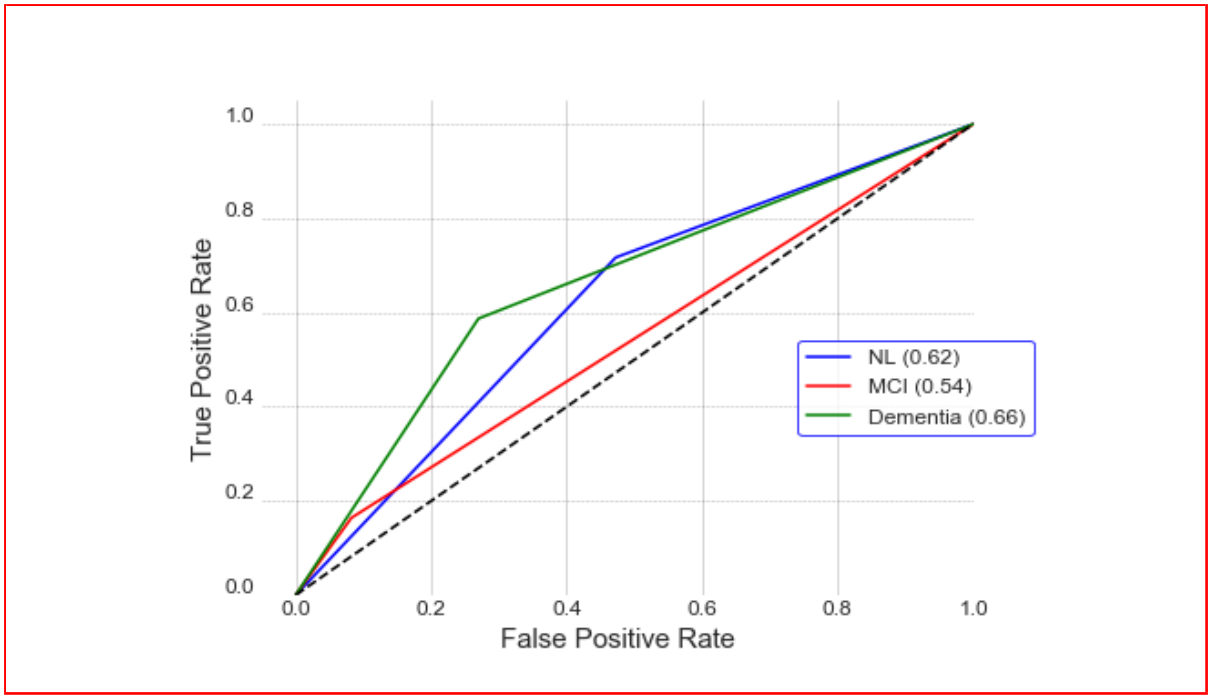

Figure 6: AUROC Curve for SVM

4.7.8 Implement, Evaluate and Result of Neural Network

Neural network attempts simulating the human brain. It is implemented using scikit-learn library. The function used to implement is MLPClassifier()[24]. It has 100 numbers of neurons and activation function is "relu". The model resulted in average AUROC score of 0.619.

4.7.9 Comparison of Developed Models

Figure 7 shows average multiclass AUROC score (CADDementia, 2014) for different machine learning algorithms. The metric assigns equal weight to the classification of each class. It shows decision tree, neural network and SVM are the top three performing models with decision tree as the best performing model. Decision tree maps non-linear relationships and is easy to interpret. However, it tends to overfit and does not handle non-numeric data well. SVM also separates classes in a multi-dimensional space but is equally likely to overfit.

[24] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

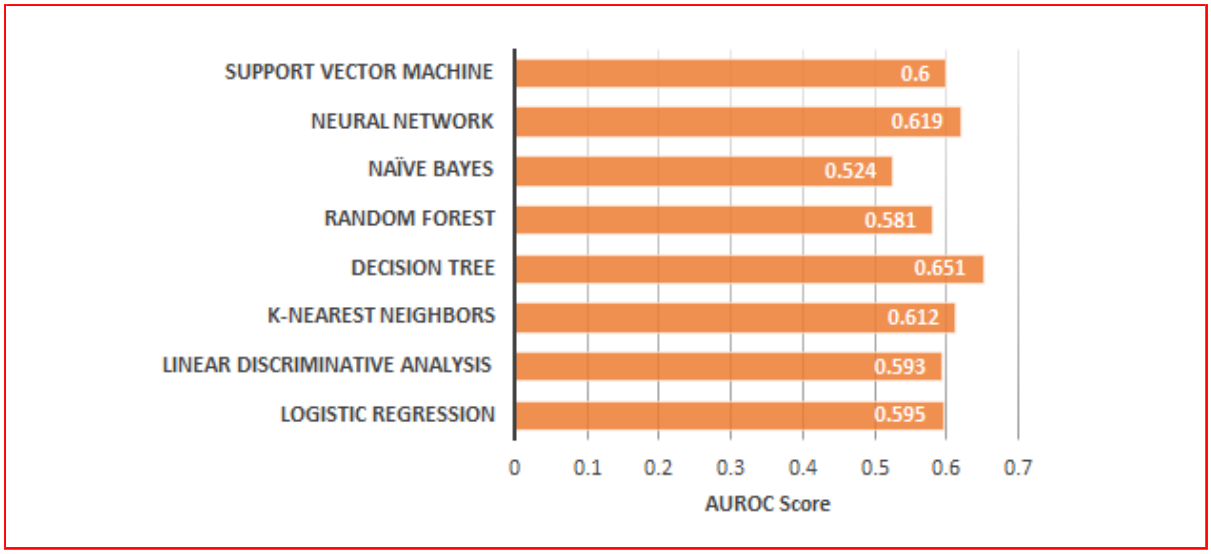

Figure 7: Average Multiclass AUROC Score

Figure 8 shows AUROC score per class for multiple algorithms. AUROC score is a measure of how well a model can distinguish between a class and other classes. For example, SVM can distinguish normal patients from other two classes with AUROC score of 0.605. It also shows that all the algorithms have a higher AUROC score for normal and dementia than MCI.

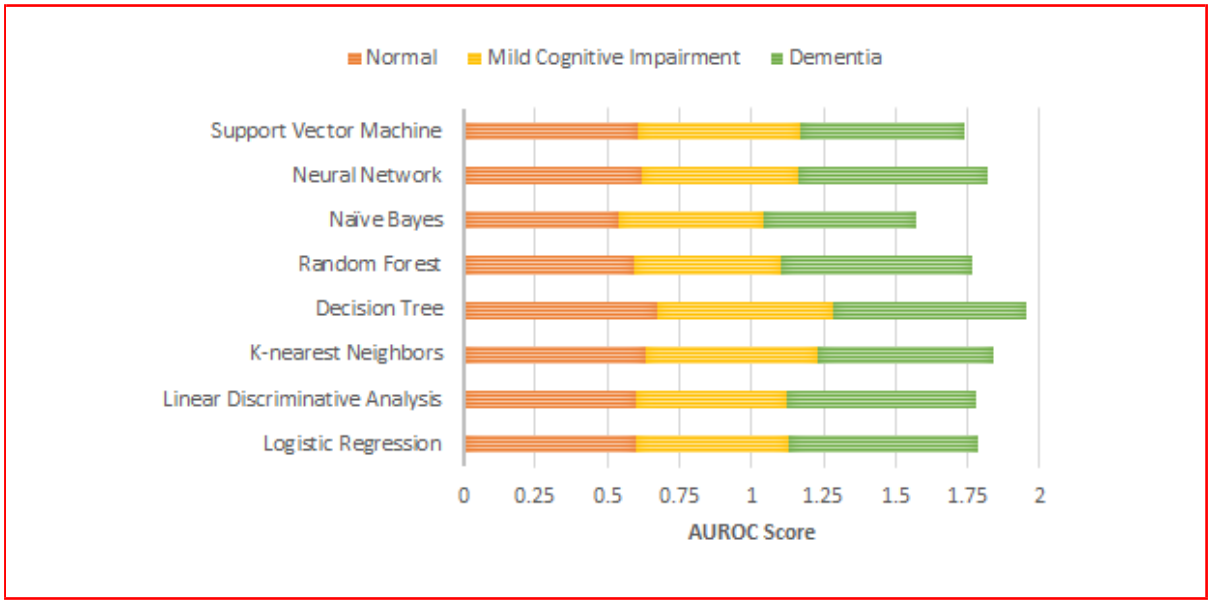

Figure 8: AUROC Score per class

4.8 Build a Classification Model using Tree-Based Algorithms

In this implementation, selected features to build classification model include subject's education, gender, age and a small list of biomarkers as discussed in different studies (Li et al., 2017), (Goyal et al., 2018), (Azvan et al., 2018) and analyzed in section 2.1.4. The biomarkers are cognitive tests such as Clinical Dementia Rating Sum of Boxes (CDRSB),

Mini-Mental State Exam (MMSE), MRI of the whole brain, hippocampus, middle temporal gyrus and entorhinal, PET measures of Fludeoxyglucose (FDG) and Florbetapir (AV45) and cerebrospinal fluid (CSF) measures of amyloid-beta, tau and phosphorylated tau level. New feature is created from age to reflect the age when the subject visited the clinic. Gender and education are converted into dummy variables. The missing values are imputed and then all the values are scaled. The data is then divided into training and test data sets. The model is trained on the training data and evaluated on test data.

4.8.1 Implementation, Evaluation and Results of Decision Tree using Different Numbers of Leaf Nodes

Decision tree is a non-linear, non-parametric algorithm which uses a tree-like graph in which each branch is an outcome of a conditional test and leaf node is a class label. The advantages include ease of understanding, identify relationships between two or more features and can handle both numeric and categorical features. However, the learners can create complex trees which overfit and are unstable because small variation in the data can result in different trees being created. The weaknesses are handled through methods such as bagging and boosting. It is implemented using scikit-learn library and the function used is DecisionTreeClassifier( ). The model is trained using 5, 50, 500, 5000, 50000 leaf nodes. The model resulted in average AUROC score of 0.845 using 5 leaf nodes, average AUROC score of 0.850 using 50 leaf nodes, average AUROC score of 0.78 using 500 leaf nodes, average AUROC score of 0.795 using 5000 leaf nodes and average AUROC score of 0.795 using 50,000 leaf nodes. The metrics show that the increase in number of leaf nodes results in poor performance for the model.

4.8.2 Implementation, Evaluation and Result of Random Forest

Random forest and gradient boosting trees are ensemble learning methods and combine outputs from multiple individual trees. Random forest uses a random sample of data to train tree independently while gradient boosting trees build new tree to correct the errors of the previous tree. Random forest is less likely to overfit than gradient boosting trees. Yet, the algorithm is slow to make predictions, biased towards features with more level and smaller groups are preferred if correlated features of the smaller groups. It is implemented using scikit-learn and the function is RandomForestClassifier(). The model resulted in predicting normal with AUROC score of 0.728 against dementia and MCI, MCI with AUROC score of 0.60 against normal and dementia and classify dementia with AUROC score of 0.877 against normal and MCI when the threshold is unfixed.

4.8.3 Implementation, Evaluation and Result of XGBoost

Gradient boosting trees can solve ranking problems because it is possible to write a gradient but take a long time to train as trees are built sequentially. Extreme Gradient Boosting (XGBoost) uses parallel computing to implement gradient boosting algorithm. It has regularization to reduce overfitting and built-in methods to handle missing values and cross-validation. It is implemented using XGBoost[25] library and function used to implement is XGBClassifier() with number of estimators set at 100.

Figure 9 is a normalized confusion matrix and shows the values of correctly predicted class i.e., 0.90 for normal (NL), 0.38 for MCI and 1.0 for dementia. The off-diagonal

[25] `https://xgboost.readthedocs.io/en/latest/`

elements are mistakenly confused with the other classes. Therefore, it is better in classifying clinical stages of normal and dementia than MCI. The value of 1.0 for dementia shows that the model overfits when the threshold is fixed at 0.5. Hence, the threshold should be selected carefully.

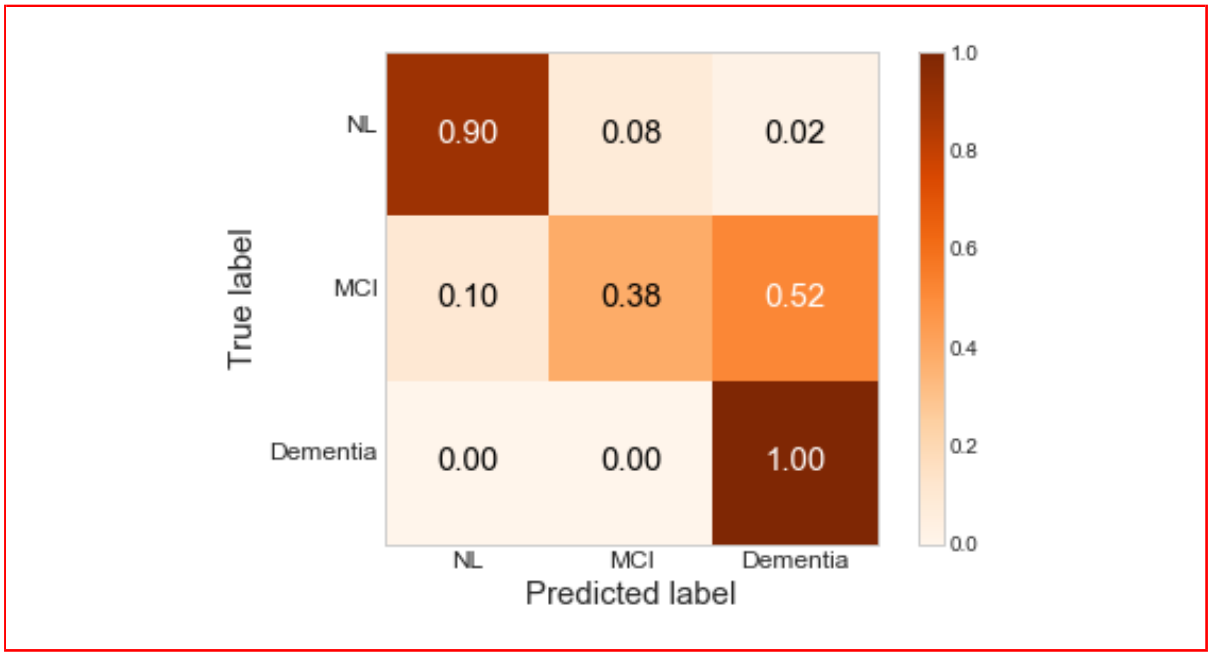

Figure 9: Normalized Confusion Matrix for XGBoost

The model resulted in predicting normal with AUROC score of 0.908 against dementia and MCI, MCI with AUROC score of 0.659 against normal and dementia and classify dementia with AUROC score of 0.853 against normal and MCI. Figure 10 is AUROC curve and shows that the classifier is better in classifying normal against the other two classes than when predicting dementia or MCI.

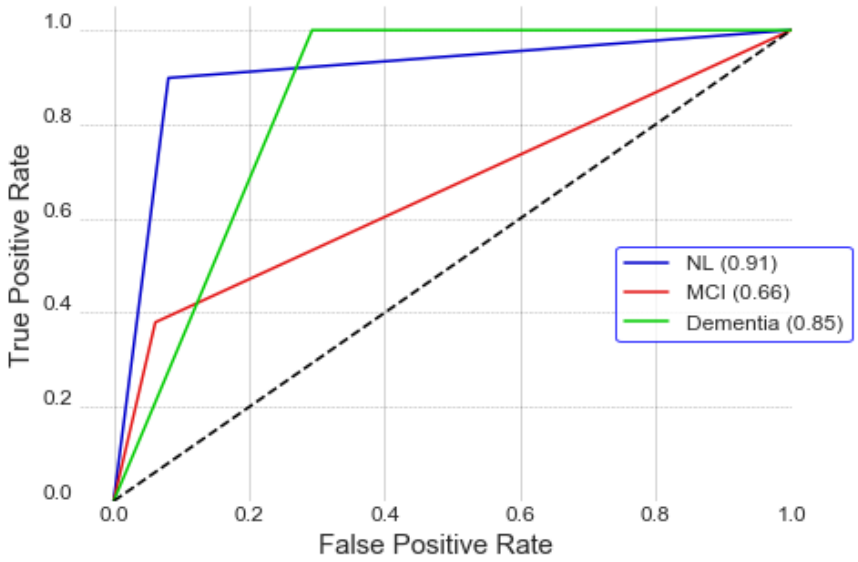

Figure 10: AUROC Curve for XGBoost

4.8.4 Comparison of Developed Models

Figure 11 shows AUROC score per class for the developed models. Decision tree with 50 leaf nodes shows the best performance but it tends to overfit. XGBoost is the best model to distinguish between a class and other classes with AUROC score 0.908 for normal, 0.659 for MCI and 0.853 for dementia. It also shows that all the algorithms have a higher AUROC score for normal and dementia than MCI.

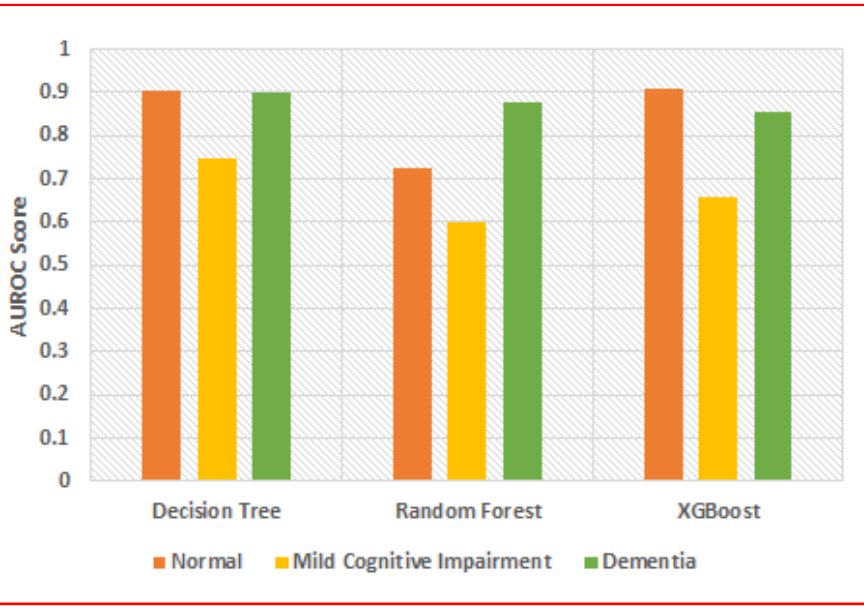

Figure 11: AUROC Score per class

4.8.5 Interpreting Machine Learning Model

Figure 12 shows the interpretation of the model prediction. It shows which features are important for a model by plotting the SHAP values of every feature for every sample. It takes the mean absolute value of the SHAP values for each feature. CDRSB score is the most important feature with a mean SHAP value of 0.35. It is used to accurately stage severity of dementia and MCI. MRI of whole brain is the next important feature followed by another cognitive test called MMSE with mean values of 0.12 and 0.11 respectively. Gender has no impact on the model. This is in contrast to the finding from one of the study that the accuracy of the prediction depends on the sex of the patient (Lee et al., 2016). Further, education has a small impact on the disease.

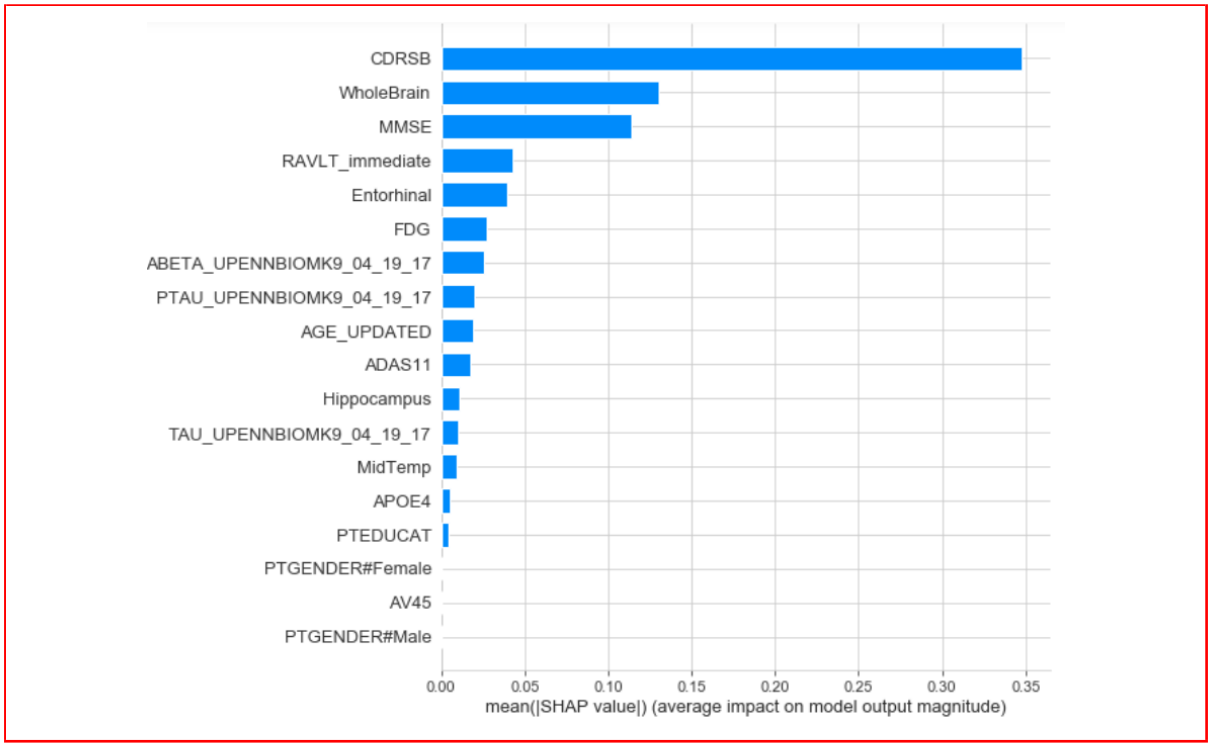

Figure 12: Feature Importance for XGBoost

4.9 Create a Web-based Application using XGBoost

The implementation uses six features which are important to classify the stages of the disease (section 4.8.5) and XGBoost to develop the web-based application. The features are CDRSB, Alzheimer's Disease Assessment Scale (ADAS11), Rey Auditory Verbal Learning Test (RAVLT)_immediate, MRI of whole brain and age of the subject at the time of visit to the clinic.

4.9.1 Implementation, Evaluation and Result of XGBoost

XGBoost library is used for implementation and function used to implement is XGBClassifier() with 100 estimators. It is trained on the training data and evaluated on test data. It resulted in predicting normal with AUROC score of 0.62 against dementia and MCI, MCI with AUROC score of 0.54 against normal and dementia and classify

dementia with AUROC score of 0.90 against normal and MCI. Therefore, it is not a good model for the selected features and require more training to improve the performance of the model. The project implements the application as a proof of concept.

4.9.2 Interpreting Machine Learning Model

In Figure 13, red means feature has a high impact and blue means feature has a low impact on the model. After plotting the sum of SHAP value magnitude over all samples for each feature, it is observed that CDRSB and ADAS11 push the prediction higher while MMSE, RAVLT_immediate and MRI of the whole brain push the predictions lower. CDRSB has a big impact on the model output with the value of 0.6. ADAS11 has a small impact on the model.

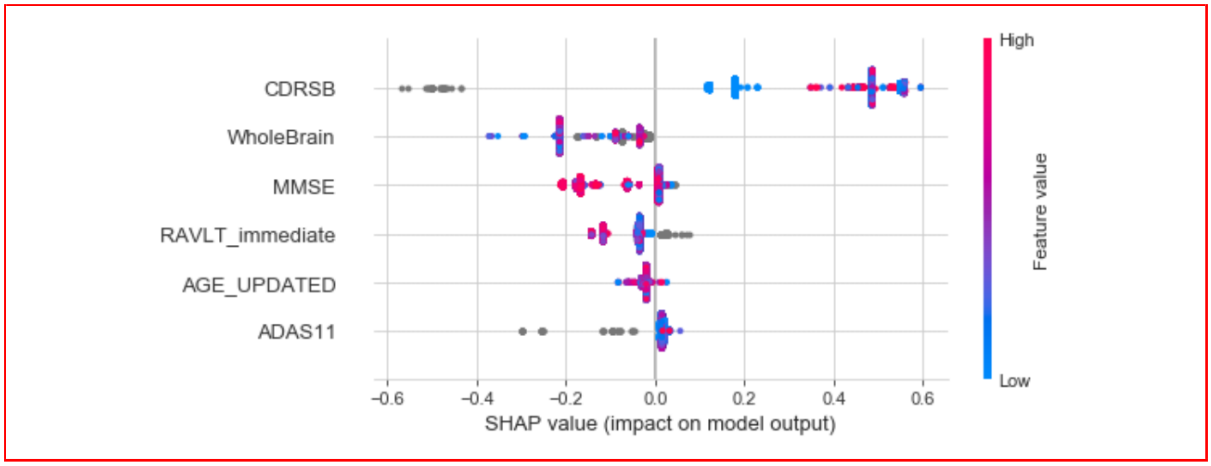

Figure 13: Feature Contribution for XGBoost (SHAP value)

Figure 14 takes the absolute mean value of the SHAP values for each feature. CDRSB is the most important feature followed by MRI of whole brain and MMSE. The impact of CDRSB is increased to a mean SHAP value to 0.42 when number of selected features is reduced than in implementation in section 4.8.3. ADAS11 and age at the time of visit to the clinic have small contribution to the model.

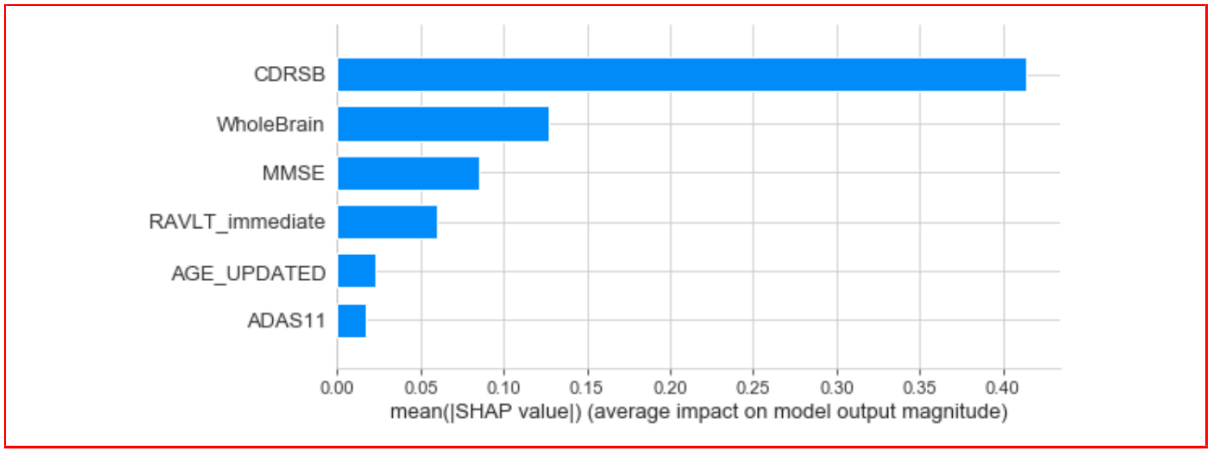

Figure 14: Feature Contribution for XGBoost (Absolute Mean Value)

4.9.3 Implementation, Design and Deployment of Web-based Application

The components of the web-based application are deployed on Heroku platform using Flask framework, Hypertext Markup Language (HTML) and Cascading Style Sheets (CSS). The process of deploying the application is as follows:

1. Download git[26] and install it to send local files to the online web server or cloud.
2. Sign up for a free account on Heroku website.
3. Download and install Heroku Toolbelt[27]. Heroku Toolbelt is a package that allows interacting with the Heroku cloud through computer command.
4. Make a git repository for the app by typing: "git init"
5. Start interacting with Heroku account by typing : "heroku login"
6. Create a custom Heroku app: "heroku create ad-model"
7. Generate the required Heroku files:
   - Create a Procfile by typing: "web: gunicorn app:app"
   - Create a requirements.txt file to tell Heroku which packages to install for web app by typing: "Flask", "Pandas", "gunicorn", "sklearn", "xgboost"
8. Add all local files to the online repository by typing: "git add ."
9. Commit your files by typing: "git commit m First commit"
10. Set the remote destination for pushing from git to Heroku : "heroku git:remote -a ad-model"
11. Push application to Heroku by typing: "git push heroku master"
12. Ensure at least one instance of the app is running by typing: "heroku ps:scale web=1"
13. Check out the deployed web-based application (Figure 15) by visiting the link for the responsive application[28].

    It requires the following inputs to make the prediction:

    - **CDRSB**: minimum value of 0, maximum value of 18 and step size of 0.5
    - **ADAS11**: minimum value of 0, maximum value of 70 and step size of 0.33
    - **MMSE**: minimum value of 0, maximum value of 30 and step size of 1
    - **RAVLT_immediate**: minimum value of 0, maximum value of 75 and step size of 1
    - **MRI of whole brain**: minimum value of 649091, maximum value of 1486040 and step size of 100

[26] `https://git-scm.com/downloads`
[27] `https://toolbelt.heroku.com/.`
[28] `https://ad-model.herokuapp.com/`

- **Age**: minimum value of 55 and maximum value of 75

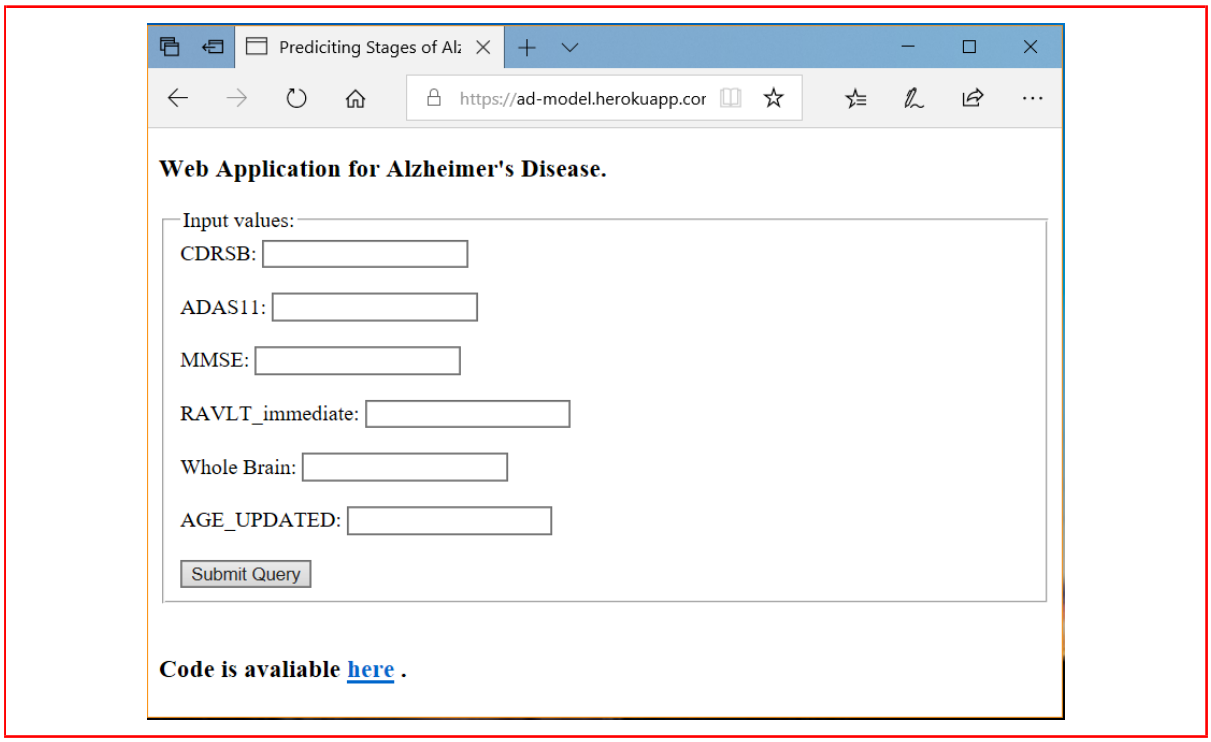

Figure 15: Web-based Application for Alzheimer's Disease Progression

4.10 Fine Tuning of an Ensemble of Classification Models Using Random Grid Search

Parameters for a model are learned during the training while hyperparameters are set to control the implementation of the model. Grid-search is a technique used to find the optimal hyperparameters for a model. Ensemble learning involves training multiple models and combining the diverse classifiers together to form a strong machine learning learner. It helps to improve robustness over a single learner and handles large volumes of data or not adequate data (Yao et al., 2018). The technique is applied by certain papers (Kruthika et al., 2019a), (Zhang and Sejdić, 2019). In this implementation, selected features are same as in the implementation in section 4.8. The goal of this implementation is to achieve a better performance by running grid search and use of ensemble of classifiers.

4.10.1 Run Grid Search to Find Most Acceptable Hyperparameters

Hyperparameter values need to be set before the learning process because the values are used to control the learning process and cannot be estimated from the data. A combination of values is used on a validation data set to find the optimal hyperparameters. Random grid search uses a grid of hyperparameters and random combination to train and score on the validation data set and not a test data set. This helps to generalize performance. Running random grid search (RandomizedSearchCV[29]) several times using cross-validation

[29]`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html`

helps to find the most acceptable parameters for the model. These hyperparameters are used for the model in section 4.10.3 and are stated and used for ensemble of classifiers in section 4.11.6

4.10.2 Generate Feature Importance of the Classifiers

Figure 16 shows the feature importance of different classifiers. It shows that in a given model the features which are important in explaining the target feature. MRI of entorhinal is the most important feature followed by MMSE for all the classifiers except Ada boost. RAVLT_immediate is the next important feature for Ada boost. However, the presence of continuous features or high-cardinality categorical features can result in a bias.

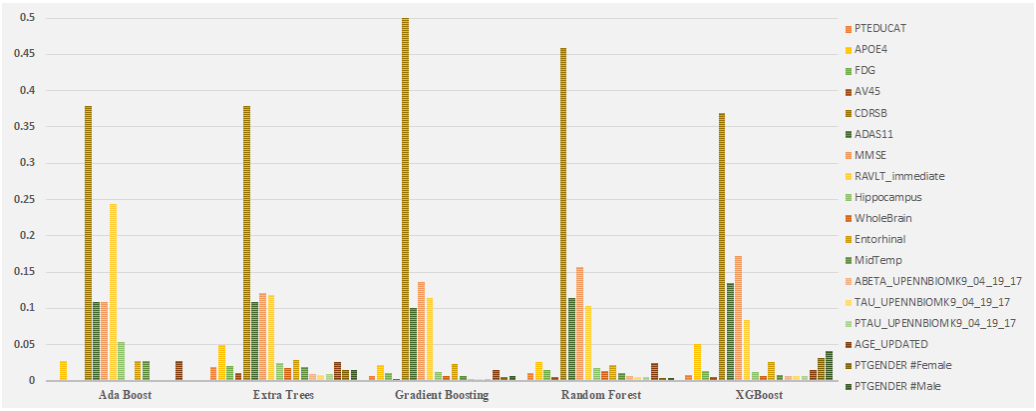

Figure 16: Feature Importance for Different Classifiers

4.10.3 Implementation, Evaluation and Result of Ensemble of Classifiers

Ensemble learning generally improves the performance of the models (Nanni et al., 2016). Random Forest[30] , extra trees classifier[31] , Ada boost classifier[32] , gradient boosting classifier[33] , and XGBoost[34] with optimized hyperparameters are combined using voting classifier[35]. Voting classifier combines the above models using soft voting which is:

$$\hat{y} = argmax_i \sum_{j=1}^{m} w_j p_{ij}$$

It is used to predict the class labels based on the predicted probabilities for well-calibrated classifier. w_j is the weight that can be assigned to the j^{th} classifier. It is implemented

[30] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[31] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

[32] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[33] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

[34] https://xgboost.readthedocs.io/en/latest/

[35] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html

using scikit-learn library and function used to implement is VotingClassifier() and weights for the model are 2, 3, 3, 1 and 3.

Figure 17 is a normalized confusion matrix for ensemble of classifiers. The values of the diagonal elements denote the degree of correctly predicted class i.e., 0.50 for normal (NL), 0.61 for MCI and 0.99 for dementia. The off-diagonal elements are mistakenly confused with the other classes. Therefore, the model is better for predicting dementia and MCI than normal with the threshold of the ensemble of classifiers fixed at 0.5.

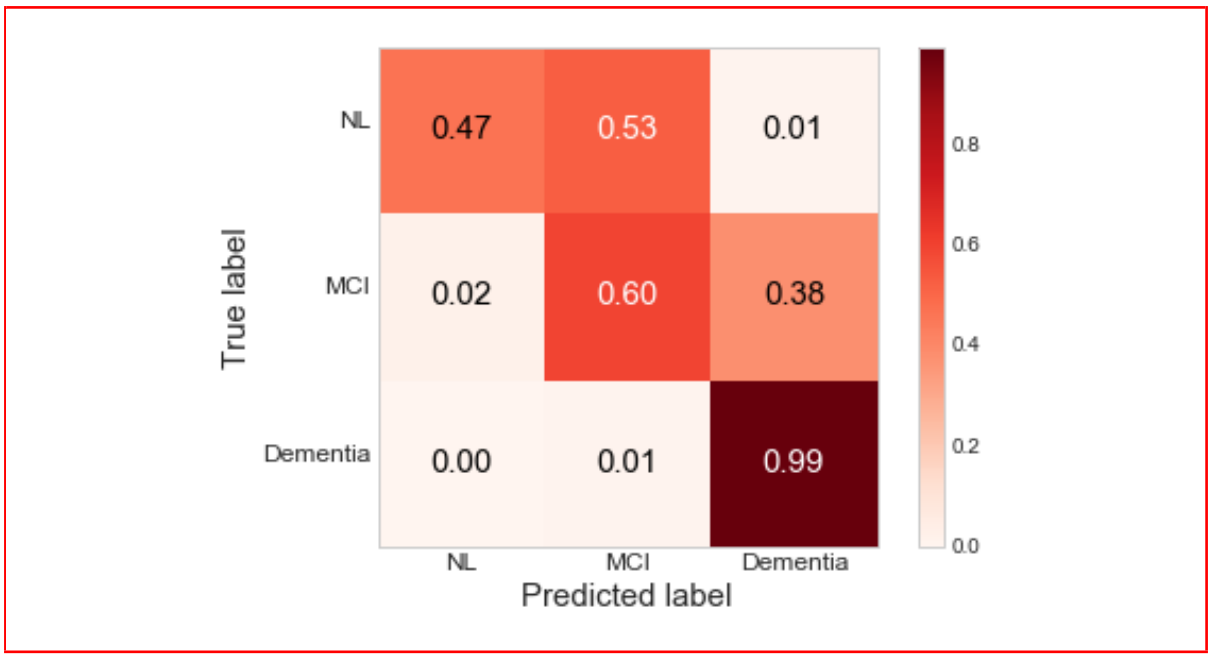

Figure 17: Normalized Confusion Matrix for Ensemble of Classifiers

The model resulted predicting normal with AUROC score of 0.739 against dementia and MCI, MCI with AUROC score of 0.613 against normal and dementia and classify dementia with AUROC score of 0.89 against normal and MCI. AUROC curve helps to measure the performance of the model without fixing the threshold. It plots a point for every possible threshold and is helpful to select the threshold of the model depending on the use case. Figure 18 shows that the ensemble of classifiers is better in predicting dementia than normal and MCI when the threshold is not fixed. Hence, the model is better than the implementation of XGboost in section 4.8.3.

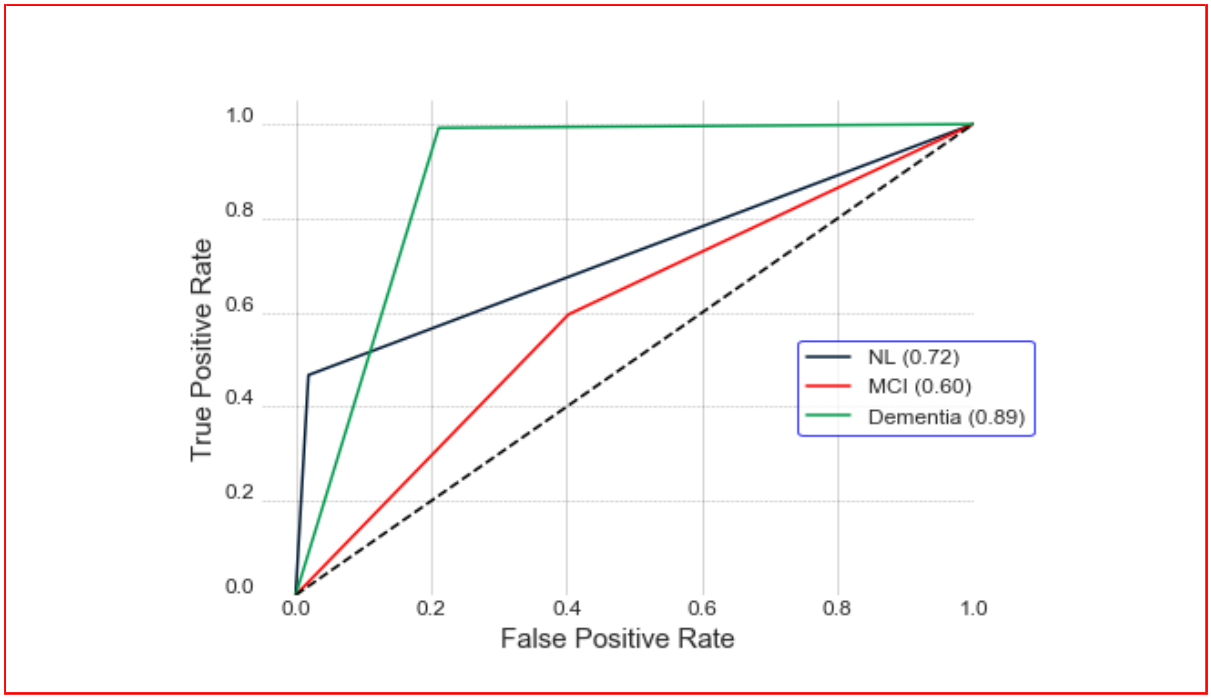

Figure 18: AUROC Curve for Ensemble of Classifiers

4.11 Use Feature Selection Techniques and Build an Ensemble of Classification Models

Feature selection is an automatic or manual process to select features which contribute to the prediction and remove irrelevant features that negatively impact the performance of the model. It helps to reduce overfitting and training time while improving performance. It is also an important technique if there are many features (Zhang and Sejdić, 2019).

4.11.1 Feature Selection using Variance Threshold, Random Forest and Univariate Selection

The implementation uses all the features from the data set where the rows were not empty are included. Certain features that do not provide information about the predicted class e.g., update stamp, site are removed. Categorical features such as gender, marital status and race are converted to dummy variables. Feature selection is done to have important predictive features from the data set. The project selects features using:

- Variance threshold[36] of over 0.90
- Random forest classifier to identify important features
- Univariate feature selection using chi-square test[37]

The number of features selected are 54 from a total of 1452 features available. Imputation is not required and scaling is done before dividing into training and test data sets.

[36] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html`

[37] `https://scikit-learn.org/stable/modules/feature_selection.html`

4.11.2 Use Principal Component Analysis to Understand the Variance Explained

Figure 19 shows the variance explained after feature selection using principal component analysis[38]. The first principal component explains the most variance (approximately 99% of the variance) among the new variables. This shows that the selected features contain most of the information in the data set and hence are good for modelling.

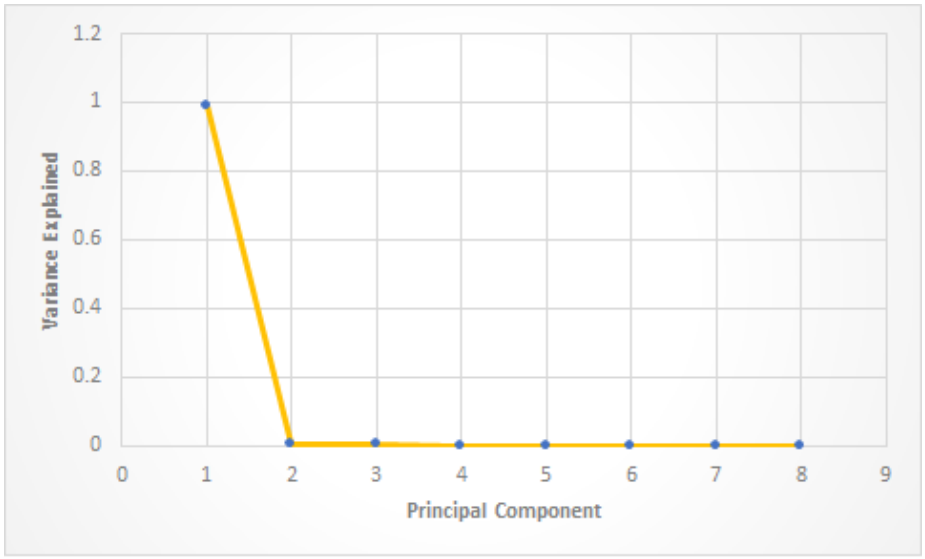

Figure 19: Principal Component Analysis

4.11.3 Implementation, Evaluation and Result of Decision Tree using 50 Leaf Nodes

It is implemented using scikit-learn library and the function is DecisionTreeClassifier()[39] with 50 leaf nodes. The model resulted in average AUROC score of 0.821. The performance of the model using more and different features is better than in section 4.8.1. Figure 20 is a decision tree in which the probability of an input belonging to each class is 0.30, 0.45, 0.23 if the value of baseline CDRSB is less than or equal to 0.25 and "gini" is equal to 0.642. The probability of an input belonging to each class is 0.908, 0.071 and 0.021 if CDRSB is true and baseline month is less than or equal to 23.9 and "gini" is equal to 0.17. However, only 29.6% of the data is classified. Similarly, the probability of an input to belong to each class is 0.057, 0.617 and 0.326 if value of CDRSB is less than or equal to 2.75 and "gini" is equal to 0.51. Only 70.4% of the data is classified.

[38] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[39] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

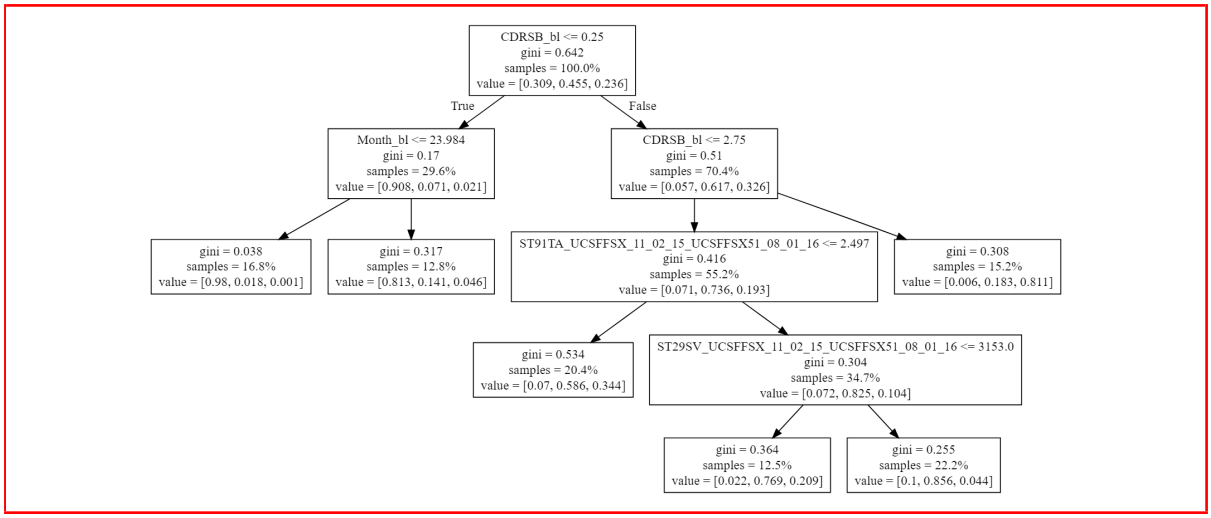

Figure 20: Decision Tree

4.11.4 Implementation, Evaluation and Result of Random Forest

Random forest uses a random sample of data to train tree independently and is less likely to overfit than gradient boosting trees. It is implemented using scikit-learn library. The function used to implement is RandomForestClassifier(). The model resulted in predicting normal with AUROC score of 0.898 against dementia and MCI, MCI with AUROC score of 0.779 against normal and dementia and classify dementia with AUROC score of 0.841 against normal and MCI when the threshold is unfixed. Hence, the performance of the model using more and different features is better than in section 4.8.2.

4.11.5 Implementation, Evaluation and Result of XGBoost

Gradient boosting is an ensemble method that is applied to train many individual trees sequentially with each tree improving over the previous tree. A combination of weak learners create a strong ensemble learner. Extreme Gradient Boosting (XGBoost) is designed for efficient multi-core parallel processing to ensure use of all the CPU cores. It is implemented using XGBoost library and function used to implement is XGBClassifier() with 100 estimators. The function has built-in methods to regularize, cross-validate and handle missing values. The model resulted in predicting normal with AUROC score of 0.901 against dementia and MCI, MCI with AUROC score of 0.685 against normal and dementia and classify dementia with AUROC score of 0.812 against normal and MCI when the threshold is unfixed. The result of the model is better when using a different set of features and more features than in section 4.8.3.

4.11.6 Implementation, Evaluation and Result of Ensemble of Classifiers

Random grid search (section 4.10.1) and ensemble learning using voting classifier is used to combine the result of multiple models to produce a single output. The hyper parameters for the classifiers are as follows:

- **Random Forest:** number of estimators - 49, minimum number of samples required to split an internal node - 25, number of features to consider when looking for the best split - auto and maximum depth of the tree - 22

- **Extra Tree:** number of trees in the forest - 60, minimum number of samples required to split an internal node - 8, number of features to consider when looking for the best split - square root and maximum depth of the tree - 37.
- **Ada boost:** number of estimators - 37 and learning rate - 0.2
- **Gradient Boosting:** subsample - 0.9, number of trees - 32, minimum number of samples - 0.01, minimum number of samples required to be at a leaf node - 10, number of features to consider when looking for the best split - square root, maximum depth of the individual estimators - 25, loss function to be optimized - deviance, learning rate to shrink the contribution of each tree by - 0.025, function to measure the quality of a split - friedman_mse
- **XGBoost:** fraction of observations to be randomly sampled for each tree - 0.6, silent mode - False, L2 regularization term on weights - 0.1, number of trees - 120, minimum sum of weights of all observations required in a child - 1.0, maximum depth of a tree - 6, learning rate - 0.01, minimum loss reduction required to make a split - 0.25, fraction of columns to be randomly sampled for each tree - 0.9, subsample ratio of columns for each split at each level - 0.4

Classifiers with optimized hyperparameters are combined using voting classifier. Voting classifier is implemented using scikit-learn library and function used to implement is VotingClassifier(). It uses "soft" voting and weights of the model are 2, 3, 3, 1 and 3.

Figure 21 is a normalized confusion matrix with the values of the diagonal elements denoting the degree of correctly predicted class i.e., 0.87 for normal, 0.59 for MCI and 0.92 for dementia when the threshold for the model is fixed at 0.5. The off-diagonal elements are mistakenly confused with the other classes e.g., 0.35 are classified as dementia when the elements are actually MCI. Therefore, the model is better in predicting normal and dementia than MCI.

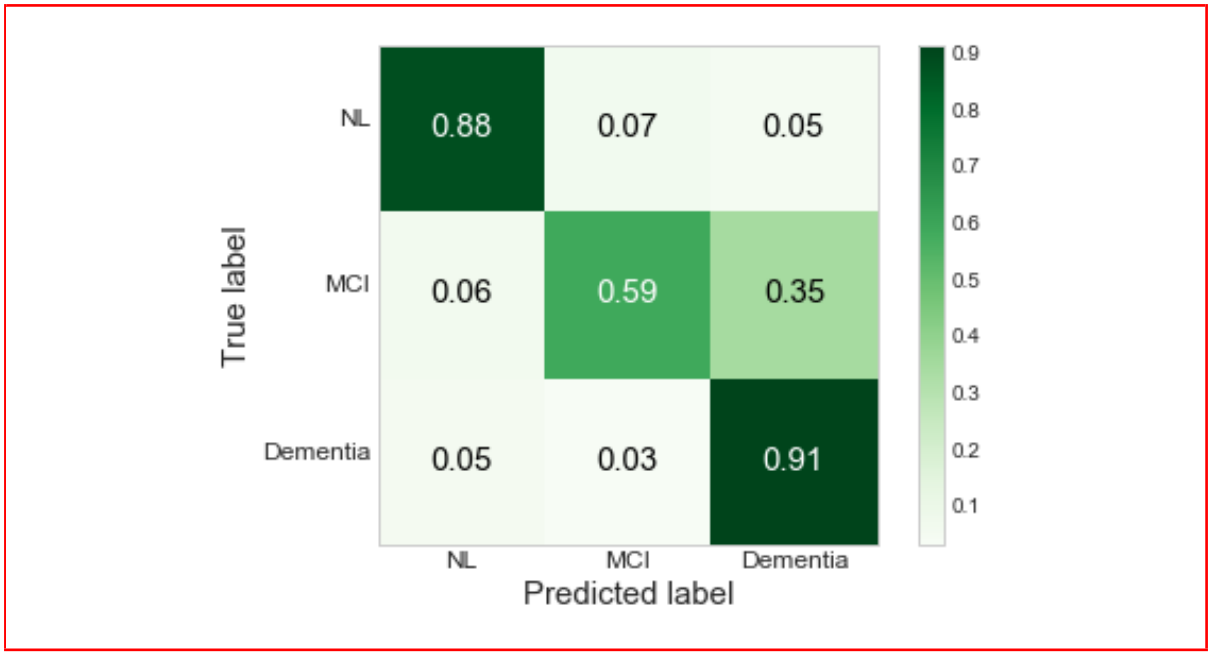

Figure 21: Normalized Confusion Matrix for Ensemble of Classifiers

The model resulted in predicting normal with AUROC score of 0.908 against dementia and MCI, MCI with AUROC score of 0.760 against normal and dementia and classify dementia with AUROC score of 0.846 against normal and MCI. Figure 22 is AUROC curve and plots a point for every possible threshold. It shows that ensemble of classifiers is better in predicting normal and dementia than MCI when the threshold is not fixed. Use of more and different sets of features has resulted in a better performing model than in section 4.10.3.

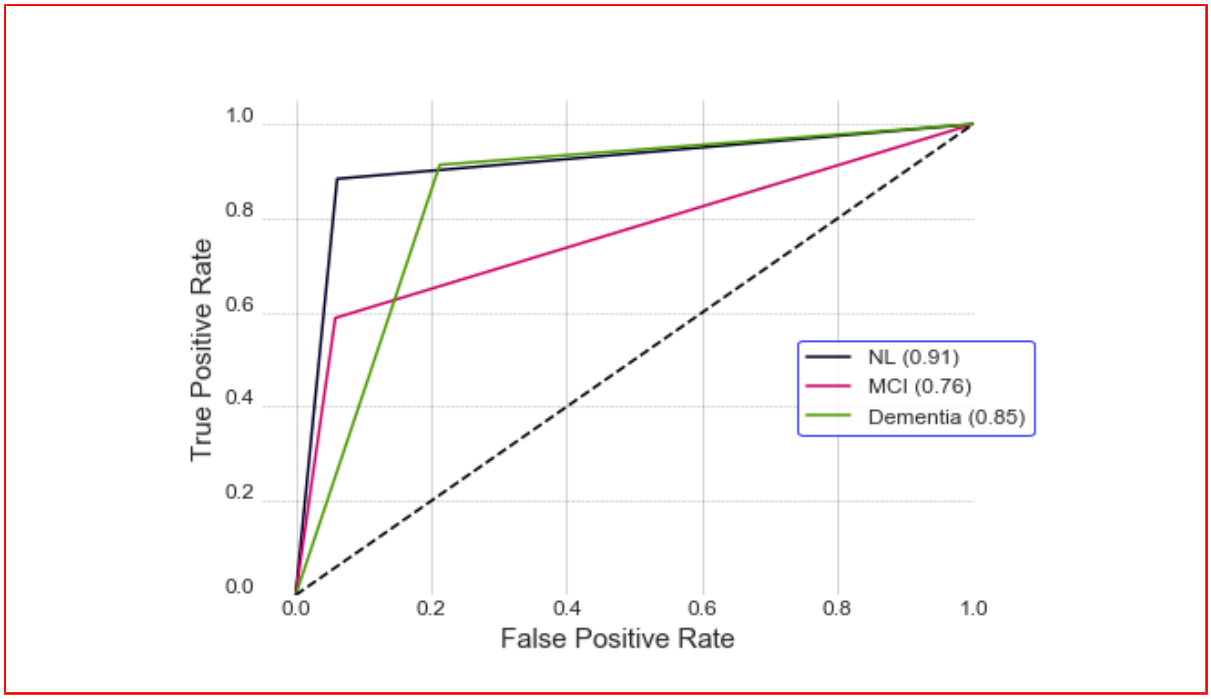

Figure 22: AUROC Curve for Ensemble of Classifiers

4.11.7 Comparison of Developed Models

Figure 23 shows AUROC score per class for each of the developed models. Both XGBoost and ensemble of classifiers are good in predicting normal clinical stage. However, the ensemble is better than the other models in distinguishing between a class and other classes with AUROC score of 0.908 for normal, 0.760 for MCI and 0.846 for dementia. Therefore, the combination of diverse classifiers and feature selection has resulted in a strong machine learning model.

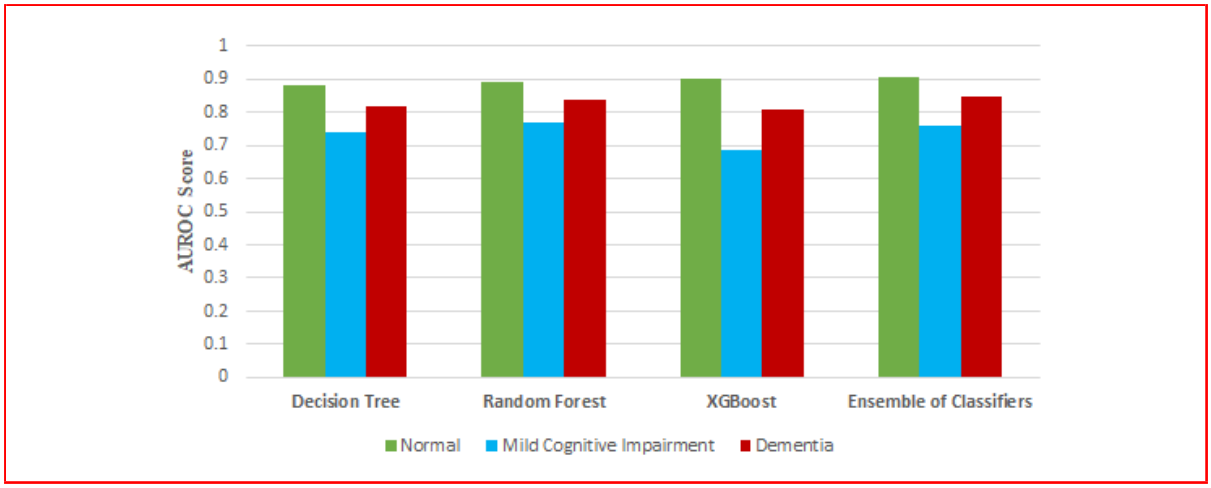

Figure 23: AUROC Score per class

4.11.8 Interpreting Machine Learning Model

It is not possible to explain the prediction made by voting classifiers using SHAP. Hence, XGBoost is implemented to gain insight and discover important features. Figure 24 explains most important features by plotting the sum of SHAP value magnitude over all samples for each feature. Red means the feature has a high impact and blue means the feature has a low impact on the model. Therefore, CDRSB from baseline has a positive impact and MMSE from baseline has a negative impact on the model. The importance of cognitive tests are in accordance with other papers such as (Goyal et al., 2018) and (Mehdipour Ghazi et al., 2019) but different than the findings of a paper (Lee et al., 2016). Volume of left hippocampus, cortical thickness of right entorhinal and months to nearest 6 months from baseline as a continuous factor are the next three important features. The feature whether or not the base image includes a timepoint has very limited impact on the model.

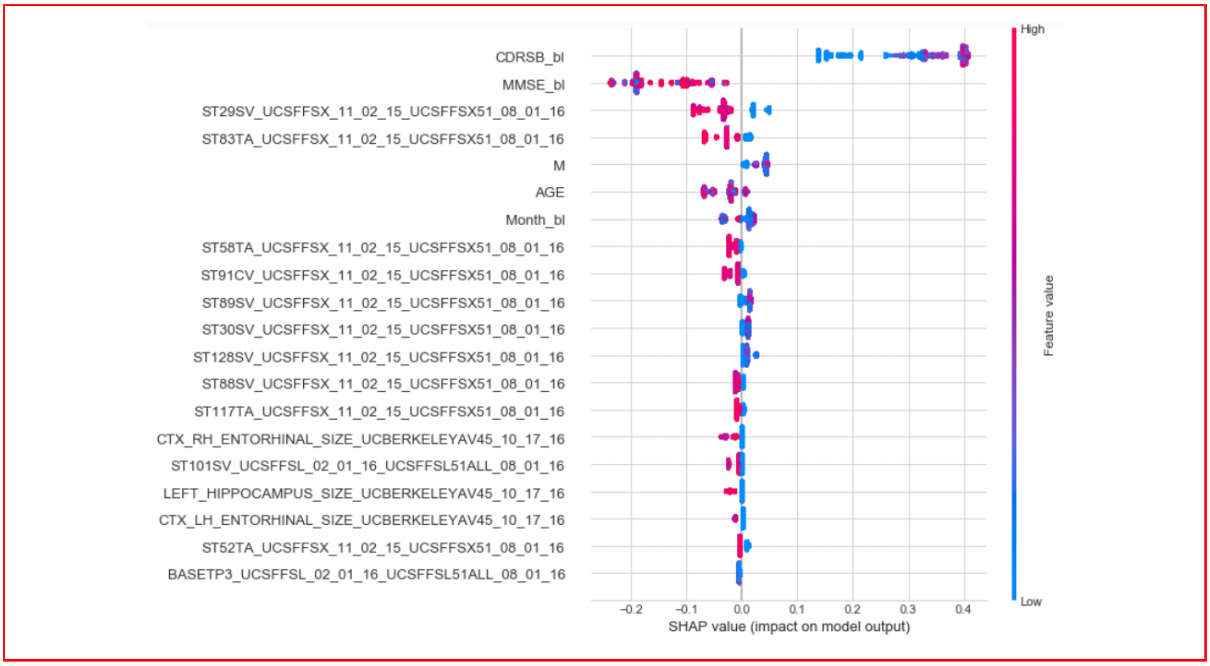

Figure 24: Feature Contribution for XGBoost (SHAP value)

Figure 25 further confirms that both CDRSB and MMSE at baseline are two most important features. Plotting mean absolute value of the SHAP values of every feature for every sample shows that CDRSB is twice important than MMSE. Furthermore, volume of left hippocampus, average cortical thickness of right entorhinal, months to nearest 6 months from baseline as a continuous factor, months from baseline and age at baseline are the next few significant features which have impact on the model order magnitude. Whether or not the base image includes a time point has very little significance.

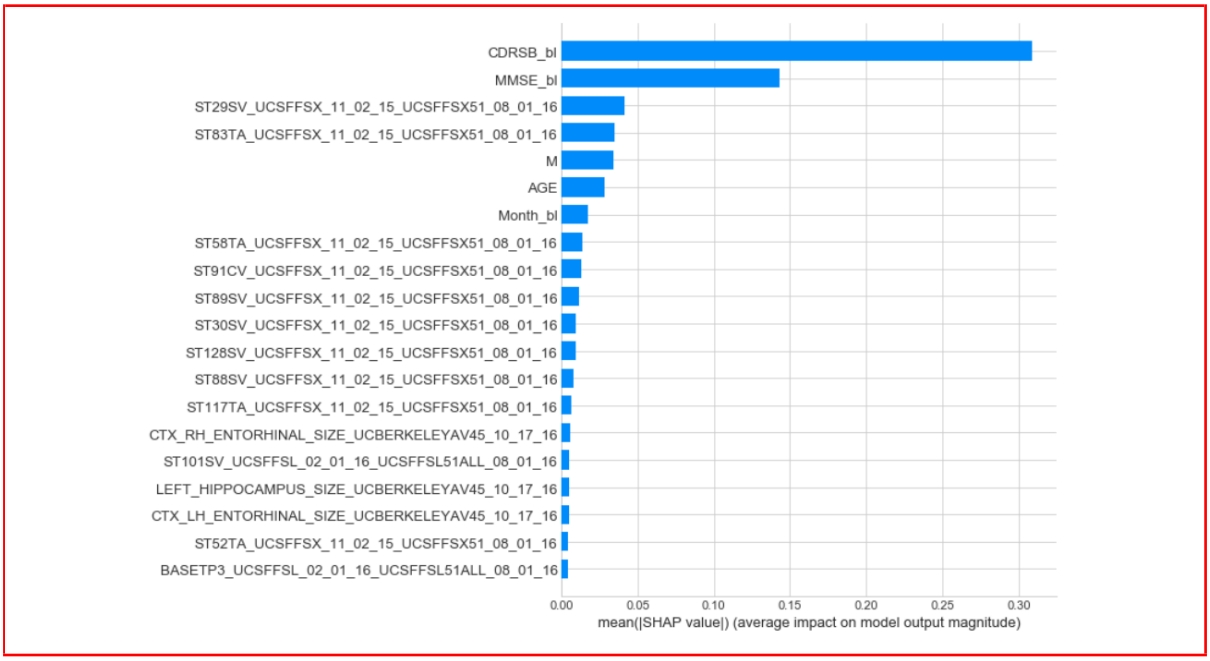

Figure 25: Feature Contribution for XGBoost (Absolute Mean Value)

4.11.9 Conclusion

The project has implemented different machine learning techniques on data from the domain of cognition behavior and radiology images after doing an exploratory analysis. The evaluation and result of the models are also discussed. A web-based application is developed and deployed to cloud.

5 Discussion

The project has fully answered research questions in section 1.2 after implementing different machine learning techniques and evaluating the performance of the models. Furthermore, all the research objectives (section 1.3) have been met. The developed model and user-friendly web application contribute to supporting an end-to-end pipeline to predict the stages of Alzheimer's disease. Further, the practical application, use of libraries, software contribute to the innovation in the research of the disease.

The project consolidates the findings from the literature review (Green et al., 2011), (Lee et al., 2016), (Goyal et al., 2018), (Venkatraghavan et al., 2019), (Wang et al., 2019) that radiology images and longitudinal clinical data can help to effectively and efficiently predict disease progression. The project successfully implements the most recent techniques such as building a machine learning classifier, explaining the output of the models, publishing the code to GitHub[40] and deploying a functional web-based application. The model also handles missing data to improve the performance of the model. It supports the various features from ADNI data set to provide a data-driven approach to model the disease progression. The project finds cognitive tests like CDRSB, MMSE, MRI of the whole brain and other factors such as age to be important features to predict the stages. The importance of cognitive tests is in accordance with other papers such as (Goyal et al., 2018) and (Mehdipour Ghazi et al., 2019) but different than the findings of a paper (Lee et al., 2016). The practical experience has resulted in the researcher gaining new skills in Python programming, data manipulation, visualization and machine learning techniques.

However, it needs to be stressed that the project aims to act as a supplementary tool to be used in conjunction with the skills and knowledge of the clinical staff. The models are simplified and find the major patterns in the data. The results from the model need to be verified by the physician. Supervised classification approach requires many labelled data thus making it dependent on the ground truth provided by the diagnosis from the clinicians. Hence, it does not consider any misdiagnosis. Further, a single sequence of features is used to predict the stage of the progression but sometimes the rate of progression is unpredictable. The project uses historical longitudinal data which have been pre-processed to classify the different stages. In future, unsupervised classification algorithms e.g., k-means clustering can be applied to improve reliability. Moreover, it uses the data set from a challenge. The model can also be utilized on a different data set to ensure the ability to transfer the techniques. Furthermore, the model classifies only three stages namely normal, MCI and dementia. The granularity of the disease can be further studied and modelled as future work.

Despite its limitations, the project showcases a framework that has a clear importance as exhibited by the impressive performance of the model. The proposed ensemble of classifiers includes feature selection component and results in good classification after testing on unseen data.

[40] `https://github.com/piushvaish/ad_ncirl`

5.1 Comparison of Developed Model with Existing State-of-the-art Models

Table 5 shows a comparison of the performance score from different studies and the developed model. It shows that the project achieves the average multiclass AUROC of 0.83 which is within the AUROC scores of the state-of-art classification models. The literature reviewed uses different models such as logistic regression and Long short-term memory (LSTM) while the project uses a new approach that was not discovered when reviewed.

Table 5: Comparison with Existing State-of-the-art Models

Machine Learning Algorithms	**Performance Score**	**Author**
Logistic Regression	AUROC Score = 0.84	(Moscoso et al., 2019)
LSTM + Linear Discriminant Analysis	AUROC Score = 0.90	(Mehdipour Ghazi et al., 2019)
Ensemble of Classifiers	AUROC Score = 0.83	This project

6 Conclusion and Future Work

The main aim of the project is to develop model and web-based application to predict the different stages of Alzheimer's disease. The gaps are identified through the literature review and then necessary steps are taken to address these gaps. Different machine learning techniques are used to build a multiclass classifier. Ensemble of classifiers is found to be the best model. A web-based application is also deployed to further the research and enable the user to predict the occurrence of the disease. It is important to mention that all the objectives mentioned in section 1.3 have been achieved.

6.1 Future Work

The proposed approach can be improved in the future in the following ways:

- The resulting predictions just indicate the three stages of Alzheimer's disease. This can be extended to more granular stage
- The predictions can be more probabilistic in nature to include the likelihood of diagnosis and present the next course of action for treatment.
- The results of the model can be logged and monitored to find the performance of the machine learning model over time.
- The web-based application is simple and provides a proof of concept. The features and model used to create can be improved further along with enhanced user design and experience.

7 Acknowledgement

I would like to thank my supervisor, Dr. Catherine Mulwa, for her advice, supervision and guidance throughout the entire process. Furthermore, I want to thank my parents

and brother for providing me with emotional support.

References

C. Aditya and M. S. Pande. Devising an interpretable calibrated scale to quantitatively assess the dementia stage of subjects with alzheimer's disease: A machine learning approach. *Informatics in Medicine Unlocked*, 6:28–35, jan 2017. ISSN 2352-9148. doi: 10.1016/J.IMU.2016.12.004.

V. Alves, R. Braga, E. Muratov, C. Andrade, V. M. Alves, R. C. Braga, E. Muratov, and C. H. Andrade. Development of Web and Mobile Applications for Chemical Toxicity Prediction. *Journal of the Brazilian Chemical Society*, 29(5):982–988, 2018. ISSN 01035053. doi: 10.21577/0103-5053.20180013.

Alzheimer's Association. 2016 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12(4):459–509, apr 2016. doi: 10.1016/j.jalz.2016.03.001.

R. Azvan, V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, S. Klein, and D. C. Alexander. TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer's Disease. Technical report, 2018.

R. J. Bateman, C. Xiong, T. L. Benzinger, A. M. Fagan, A. Goate, N. C. Fox, D. S. Marcus, N. J. Cairns, X. Xie, T. M. Blazey, D. M. Holtzman, A. Santacruz, V. Buckles, A. Oliver, K. Moulder, P. S. Aisen, B. Ghetti, W. E. Klunk, E. McDade, R. N. Martins, C. L. Masters, R. Mayeux, J. M. Ringman, M. N. Rossor, P. R. Schofield, R. A. Sperling, S. Salloway, and J. C. Morris. Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease. *New England Journal of Medicine*, 367(9):795–804, aug 2012. ISSN 0028-4793. doi: 10.1056/NEJMoa1202753.

M. Bilgel and B. M. Jedynak. Predicting time to dementia using a quantitative template of disease progression. 2018. doi: 10.1101/458273.

CADDementia. CADDementia - Evaluation, 2014. URL `https://caddementia.grand-challenge.org/Evaluation/`.

E. Colantuoni, G. Surplus, A. Hackman, H. M. Arrighi, and R. Brookmeyer. Web-based application to project the burden of Alzheimer's disease. *Alzheimer's & Dementia*, 6 (5):425–428, sep 2010. ISSN 1552-5260. doi: 10.1016/J.JALZ.2010.01.014.

R. Cui and M. Liu. RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, 73:1–10, apr 2019. ISSN 0895-6111. doi: 10.1016/J.COMPMEDIMAG.2019.01.005.

R. S. Doody, V. Pavlik, P. Massman, S. Rountree, E. Darby, and W. Chan. Predicting progression of Alzheimer's disease. *Alzheimer's Research & Therapy*, 2(1):2, 2010. ISSN 1758-9193. doi: 10.1186/alzrt25.

T. Fawcett. doi:10.1016/j.patrec.2005.10.010. 2005. doi: 10.1016/j.patrec.2005.10.010.

J.-B. Fiot, H. Raguet, L. Risser, L. D. Cohen, J. Fripp, and F.-X. Vialard. Longitudinal deformation models, spatial regularizations and learning strategies to quantify Alzheimer's disease progression. *NeuroImage: Clinical*, 4:718–729, jan 2014. ISSN 2213-1582. doi: 10.1016/J.NICL.2014.02.002.

C. K. Fisher, A. M. Smith, and J. R. Walsh. Using deep learning for comprehensive, personalized forecasting of Alzheimer's Disease progression. Technical report, 2018.

D. Goyal, D. Tjandra, R. Q. Migrino, B. Giordani, Z. Syed, and J. Wiens. Characterizing heterogeneity in the progression of Alzheimer's disease using longitudinal clinical and neuroimaging biomarkers. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:629–637, jan 2018. ISSN 2352-8729. doi: 10.1016/J.DADM.2018.06.007.

C. Green, J. Shearer, C. W. Ritchie, and J. P. Zajicek. Model-Based Economic Evaluation in Alzheimer's Disease: A Review of the Methods Available to Model Alzheimer's Disease Progression. *Value in Health*, 14(5):621–630, jul 2011. ISSN 1098-3015. doi: 10.1016/J.JVAL.2010.12.008.

C. K. Fisher, A. M. Smith, J. R. Walsh, and the Coalition Against Major Diseases. *Using deep learning for comprehensive, personalized forecasting of Alzheimer's Disease progression*. jul 2018.

K. Kruthika, Rajeswari, and H. Maheshappa. CBIR system using Capsule Networks and 3D CNN for Alzheimer's disease diagnosis. *Informatics in Medicine Unlocked*, 14:59–68, jan 2019a. ISSN 2352-9148. doi: 10.1016/J.IMU.2018.12.001.

K. Kruthika, Rajeswari, and H. Maheshappa. Multistage classifier-based approach for Alzheimer's disease prediction and retrieval. *Informatics in Medicine Unlocked*, 14: 34–42, jan 2019b. ISSN 2352-9148. doi: 10.1016/J.IMU.2018.12.003.

S. Lahmiri and A. Shmuel. Performance of machine learning methods applied to structural MRI and ADAS cognitive scores in diagnosing Alzheimer's disease. *Biomedical Signal Processing and Control*, nov 2018. ISSN 1746-8094. doi: 10.1016/J.BSPC.2018.08.009.

S. H. Lee, A. H. Bachman, D. Yu, J. Lim, and B. A. Ardekani. Predicting progression from mild cognitive impairment to Alzheimer's disease using longitudinal callosal atrophy. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 2:68–74, jan 2016. ISSN 2352-8729. doi: 10.1016/J.DADM.2016.01.003.

Lemaitre G., Nogueira F., Oliveira D., and Aridas C. imblearn.over_sampling.SMOTE imbalanced-learn 0.5.0 documentation, 2017.

K. Li, W. Chan, R. S. Doody, J. Quinn, S. Luo, and t. A. D. N. Alzheimer's Disease Neuroimaging Initiative. Prediction of Conversion to Alzheimer's Disease with Longitudinal Measures and Time-To-Event Data. *Journal of Alzheimer's disease : JAD*, 58(2):361–371, 2017. ISSN 1875-8908. doi: 10.3233/JAD-161201.

S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions, 2017.

J. C. Masdeu, J. L. Zubieta, and J. Arbizu. Neuroimaging as a marker of the onset and progression of Alzheimer's disease. *Journal of the Neurological Sciences*, 236(1-2):55–64, sep 2005. ISSN 0022-510X. doi: 10.1016/J.JNS.2005.05.001.

M. Mehdipour Ghazi, M. Nielsen, A. Pai, M. J. Cardoso, M. Modat, S. Ourselin, and L. Sørensen. Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling. *Medical Image Analysis*, 53:39–46, apr 2019. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2019.01.004.

M. Memedi, J. Westin, D. Nyholm, M. Dougherty, and T. Groth. A web application for follow-up of results from a mobile device test battery for Parkinson's disease patients. *Computer Methods and Programs in Biomedicine*, 104(2):219–226, nov 2011. ISSN 01692607. doi: 10.1016/j.cmpb.2011.07.017.

A. J. Mishizen-Eberz, R. A. Rissman, T. L. Carter, M. D. Ikonomovic, B. B. Wolfe, and D. M. Armstrong. Biochemical and molecular studies of NMDA receptor subunits NR1/2A/2B in hippocampal subregions throughout progression of Alzheimer's disease pathology. *Neurobiology of Disease*, 15(1):80–92, feb 2004. ISSN 0969-9961. doi: 10.1016/J.NBD.2003.09.016.

A. Moscoso, J. Silva-Rodríguez, J. M. Aldrey, J. Cortés, A. Fernández-Ferreiro, N. Gómez-Lado, Á. Ruibal, and P. Aguiar. Prediction of Alzheimer's disease dementia with MRI beyond the short-term: Implications for the design of predictive models. *NeuroImage: Clinical*, page 101837, apr 2019. ISSN 2213-1582. doi: 10.1016/J.NICL.2019.101837.

L. Nanni, C. Salvatore, A. Cerasa, and I. Castiglioni. Combining multiple approaches for the early diagnosis of Alzheimer's Disease. *Pattern Recognition Letters*, 84:259–266, dec 2016. ISSN 0167-8655. doi: 10.1016/J.PATREC.2016.10.010.

New York State Coordinating Council. 2017 Report of the New York State Coordinating Council for Services Related to Alzheimer's Disease and Other Dementias to Governor Andrew M. Cuomo and the New York State Legislature. Technical report, 2017.

N. O'Kelly. Use of Machine Learning Technology in the Diagnosis of Alzheimer's Disease Declaration of Authorship. Technical report, 2016.

S. Patel, B.-R. Bor-rong Chen, T. Buckley, R. Rednic, D. McClure, D. Tarsy, L. Shih, J. Dy, M. Welsh, and P. Bonato. Home monitoring of patients with Parkinson's disease via wearable technology and a web-based application. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, volume 2010, pages 4411–4414. IEEE, aug 2010. ISBN 978-1-4244-4123-5. doi: 10.1109/IEMBS.2010.5627124.

T. Pereira, L. Lemos, S. Cardoso, D. Silva, A. Rodrigues, I. Santana, A. de Mendonça, M. Guerreiro, and S. C. Madeira. Predicting progression of mild cognitive impairment to dementia using neuropsychological data: a supervised learning approach using time windows. *BMC Medical Informatics and Decision Making*, 17(1):110, dec 2017. ISSN 1472-6947. doi: 10.1186/s12911-017-0497-2.

K. Przednowek, K. Wiktorowicz, T. Krzeszowski, and J. Iskra. A web-oriented expert system for planning hurdles race training programmes. *Neural Computing and Applications*, pages 1–17, may 2018. ISSN 0941-0643. doi: 10.1007/s00521-018-3559-1.

A. Schmidt-Richberg, C. Ledig, R. Guerrero, H. Molina-Abril, A. Frangi, D. Rueckert, and o. b. o. t. A. D. N. Alzheimer's Disease Neuroimaging Initiative. Learning Biomarker Models for Progression Estimation of Alzheimer's Disease. *PloS one*, 11(4):e0153040, 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0153040.

J. Skinner, J. O. Carvalho, G. G. Potter, A. Thames, E. Zelinski, P. K. Crane, L. E. Gibbons, and Alzheimer's Disease Neuroimaging Initiative. The Alzheimer's Disease Assessment Scale-Cognitive-Plus (ADAS-Cog-Plus): an expansion of the ADAS-Cog to improve responsiveness in MCI. *Brain imaging and behavior*, 6(4):489–501, dec 2012. ISSN 1931-7565. doi: 10.1007/s11682-012-9166-3.

V. Venkatraghavan, E. E. Bron, W. J. Niessen, and S. Klein. Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage*, 186:518–532, feb 2019. ISSN 1053-8119. doi: 10.1016/J.NEUROIMAGE.2018.11.024.

S. Vieira, W. H. Pinaya, and A. Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, 74:58–75, mar 2017. ISSN 0149-7634. doi: 10.1016/J.NEUBIOREV.2017.01.002.

M. Wang, D. Zhang, D. Shen, and M. Liu. Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data. *Medical Image Analysis*, 53:111–122, apr 2019. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2019.01.007.

T. Wang, R. G. Qiu, and M. Yu. Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks. *Scientific Reports*, 8(1):9161, dec 2018. doi: 10.1038/s41598-018-27337-w.

E. Yang, M. Farnum, V. Lobanov, T. Schultz, N. Raghavan, M. N. Samtani, G. Novak, V. Narayan, and A. DiBernardo. Quantifying the Pathophysiological Timeline of Alzheimer's Disease. *Journal of Alzheimer's Disease*, 26(4):745–753, oct 2011. ISSN 18758908. doi: 10.3233/JAD-2011-110551.

D. Yao, V. D. Calhoun, Z. Fu, Y. Du, and J. Sui. An ensemble learning system for a 4-way classification of Alzheimer's disease and mild cognitive impairment. *Journal of Neuroscience Methods*, 302:75–81, may 2018. ISSN 0165-0270. doi: 10.1016/J.JNEUMETH.2018.03.008.

A. L. Young, N. P. Oxtoby, P. Daga, D. M. Cash, N. C. Fox, S. Ourselin, J. M. Schott, and D. C. Alexander. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*, 137(9):2564–2577, sep 2014a. ISSN 1460-2156. doi: 10.1093/brain/awu176.

A. L. Young, N. P. Oxtoby, P. Daga, D. M. Cash, N. C. Fox, S. Ourselin, J. M. Schott, and D. C. Alexander. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*, 137(9):2564–2577, sep 2014b. ISSN 1460-2156. doi: 10.1093/brain/awu176.

R. Zhang, G. Simon, and F. Yu. Advancing Alzheimer's research: A review of big data promises. *International Journal of Medical Informatics*, 106:48–56, oct 2017. ISSN 1386-5056. doi: 10.1016/J.IJMEDINF.2017.07.002.

Z. Zhang and E. Sejdić. Radiological images and machine learning: Trends, perspectives, and prospects. *Computers in Biology and Medicine*, 108:354–370, may 2019. ISSN 0010-4825. doi: 10.1016/J.COMPBIOMED.2019.02.017.

J. Zhou, E. Gennatas, J. Kramer, B. Miller, and W. Seeley. Predicting Regional Neurodegeneration from the Healthy Brain Functional Connectome. *Neuron*, 73(6): 1216–1227, mar 2012. ISSN 08966273. doi: 10.1016/j.neuron.2012.03.004.