MACHINE LEARNING CLASSIFICATION OF ALZHEIMER'S PATIENTS USING SMRI VOLUMETRIC DATA

A Thesis

Submitted By

Vishal Raj

For the award of the degree of

MASTER OF TECHNOLOGY, CLINICAL ENGINEERING

Jointly offered by

Indian Institute of Technology, Madras

Christian Medical College, Vellore

Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum

Is evaluated and approved by

Mr. Ranjith G

(Guide)

Dr. Bejoy Thomas

(Examiner)

PROFESSOR AND HEAD
DEPT. OF IMAGING SCIENCES &
INTERVENTIONAL RADIOLOGY
SREE CHITRA TIRUNAL INSTITUTE
FOR MEDICAL SCIENCES & TECHNOLOGY
THIRUVANANTHAPURAM 695 011, INDIA

June 2021

CERTIFICATE

This is to certify that the thesis titled '**Machine Learning Classification of Alzheimer's Patients Using SMRI Volumetric Data'** being submitted by Vishal Raj to SCTIMST Trivandrum, for the award of degree of **Master of Technology in Clinical Engineering** jointly offered by IIT Madras, CMC Vellore and SCTIMST Trivandrum,is a bonafide record of research work done by him under our supervision. The contents of this thesis in full or in parts have not been submitted to any other Institute or University for the award of any degree or diploma.

The research had been carried out at Sree Chitra Institute of Medical Sciences and Technology, Trivandrum.

Mr. Ranjith G

Guide

Engineer 'E',

Bio-Medical Technology Wing,

SCTIMST, Trivandrum

ACKNOWLEDGEMENT

This research project is part of my course of Master of Technology degree (Specialization in Clinical Engineering) at Indian Institute of Technology Madras with collaboration with SCTIMST Trivandrum and CMC Vellore. During the course of this project I have learnt a lot in the area of Clinical Engineering. The journey was evolutionary and enriching. The present research work was interesting and for the successful completion it required assistance and support of several people. There are many people for whom I cannot express my gratitude in words but want to acknowledge their contribution in making this research successful. Firstly, I would thank my guide/mentor, Mr. Ranjith G Sir, for his continuous supervision and valuable feedbacks which helped me to constantly engage with my objectives of the project. Specially I would like to thank Muskan Khetan, my classmate and also one of my good friend, she helped me a lot to complete this project and lastly I would also like to thank my friends and family for their support in understanding and respecting my efforts towards the project. Last but not the least I would like to thank the Sree Chitra Tirunal Institute for Medical Sciences and Technology Trivandrum and Indian Institute of Technology Madras to provide me this opportunity to work on the present research.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

MCI	MILD COGNITIVE IMPAIRMENT
MRI	MAGNETIC RESONANCE IMAGING
CT	COMPUTED TOMOGRAPHY
RF	RANDOM FOREST
DT	DECISION TREE
SVM	SUPPORT VECTOR MACHINE
LGBM	LIGHT GRADIENT BOOSTING
XGB	XG BOOSTING
GB	GRADIENT BOOSTING
ROC	RECEIVER OPERATOR CHARACTERISTIC
GM	GRAY VOLUME
CSF	CEREBROSPINAL FLUID
PET	POSITRON EMISSION TOMOGRAPHY
SMRI	STRUCTURAL MAGNETIC RESONANCE IMAGING
AUC	AREA UNDER THE CURVE
FPR	FALSE POSITIVE RATE
FNR	FALSE NEGATIVE RATE
TNR	TRUE NEGATIVE RATE
TPR	TRUE POSITIVE RATE
EDA	EXPLORATORY DATA ANALYSIS
CSV	COMMON SEPARATED VALUE

ABSTRACT

Alzheimer's disease is growing concern especially among the aged population and it is also expected to increase in coming years. Unfortunately, there is no cure for this disease and the treatment strategy is the management of the symptoms. Even though the disease cannot be reversed, early detection can lead to better management of the symptoms .. Therefore, there is an urge for development of advance techniques to detect the disease as early as possible. One of the characteristics of this disease is that it brings certain structural infirmities in the brain structure. These biological markers are detectable in Magnetic Resonance Imaging. Here, the researcher attempts to use image processing algothms for segmentation of ROIs coupled with machine learning algorithms to classify the Alzheimer's disease patients into AD,MCI and normals using structural MRI data. The present research involves investigation of structural MRI of 289 subjects. The calculation of grey matter and CSF volume was conducted for 142 Region of Interest using MATLAB with the help of SPM12 and CAT12 toolbox. These were used as inputs for the machine learning algorithms The machine learning algorithms used include: Decision Tree, Random Forest, Gradient Boosting, XG Boosting, Light XGB. To select the best features out of the data set available a machine algorithm, i.e., SHAP has been used.

The accuracy we got for different machine learning algorithms to classify the multi class patients can be mentioned as , for Decision tree (91.69), for Random Forest (92.35), For logistic regression (94.11), for XG boosting(91.69), Gradient Boosting (91.38), Light GBM (93.64)

Features extracted from certain anatomical regions like caudate, occipital fusiform gyrus, frontal gyrus , hippocampus , supramarginal gyrus , middle frontal gyrus were found to be most discriminative as regards the classification task.

Keywords : Alzheimer's Disease, Machine Learning, Decision Tree, Random Forest, Gradient Boosting, XG Boosting, Light XGB

CHAPTER 1
INTRODUCTION

1. UNDERSTANDING ALZHEIMER

Alzheimer's disease is a growing medical problem in many people these days. It is, in fact, a kind of dementia. This illness originates in the brain and is a neurodegenerative illness. This disease deteriorates the brain capacity as it diminishes the nerve cells. In case where the nerve cell diminishes it leads to lowering of responses to the connecting cells in the brain. The situation increases results in compound interferences, the transmission of impulse is moderate, lastly tissues in the brain start to deteriorate.Alzheimer's disease is distinguished by the presence of plaques and tangles.. The initial part of brain which is affected is the Hippocampus. Hippocampus is dedicated in memory creation and it acts as a relay frame between body and brain. In patients with Alzheimeirs hippocampus shrinks usually between 2.2 to 5.9 % every year. The contraction of the hippocampus is primarily because of cell loss and diminishing of synapses. The damage to synapses leads to loss of capability of neurons to communicate among themselves through signaling. Therefore, in certain stages problems in episodic and short-term memory along with cessation of neural transmission is observed. Alzheimer's disease damages the neurotransmitters, neurons, and brain cells. The demolition of these results in cluster of protein which creates and surround the cells of brain. The formed cluster are called as 'plaques' and 'bundles'. The existence of these 'plaques' and 'bundles' begin to affect many more connections among the brain cells and this deteriorate the patient situation.

1.3 Comparisonbetween Normal and Alzheimers's Disease brain.

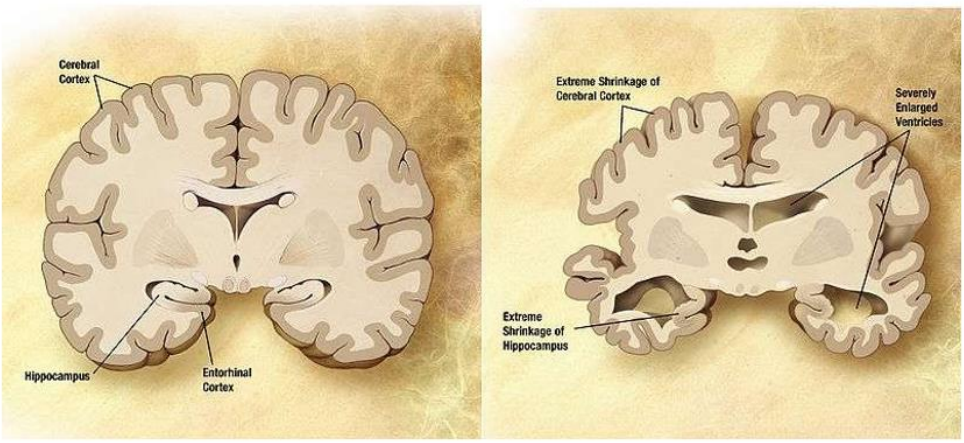


Fig 1: Comparison of diagram of normal brain and diagram of a brain od a person with Alzheimers's Disease

(Source: https://www.writework.com/essay/alzheimer-s-disease)

It is cerebral cortex which is exterior part of the brain and mainly it is dedicated for mind working like thoughts innovations. Area between the folds known as sulci is enlarged.

The initial diagnosis involves two stages: Examination of the past medical reports and questionnaire to find out past medical issues. Physical examination is next step and it involves checkup of different body organs. Then the Neuropsychological testing is conducted. Doctors utilize different kinds of tools to assess memory, problem-solving, attention and abstract thinking. The technique of Brain imaging scan is used like MRI and CT imaging. This scanning method is utilized to detect the presence of tumors and clotting as these are also indication of the disease. Symptoms of Alzheimer are like problem in doing usual tasks, for example cooking, opening or closing of car door or applying any general tools, Memory loss that decrease the job skills. If such forgetfulness is there occasionally it is normal but if it is frequent than it creates an issue like no proper concentration on work, Issues utilizing language might be an indication of Alzheimer's sickness.

2. MILD COGNITIVE IMPAIRMENT (MCI)

It is the phase amidst the age when it is expected that the cognitive aspect would decrease and diminishing of dementia. It affects the thinking and judgment making ability of the patient. In case anyone has the MCI, their working of mind must have fallen. Those who lives such patients would have recognized the alteration in the patient behaviour. However, many a times these shifts are not that evident and do not turn to be more severe in nature. MCI push the chances of forming and evolving of dementia due to Alzheimer. There are changes of recovery of patients with MCI.Your cerebrum changes as you become more elder. In most cases individuals notice step by step expanding tendency to forget as they are becoming aged. With age the people are unable to remember any word or phrase quickly. In any case, predictable or expanding worry about your psychological exhibition may recommend MCI. Psychological issues may go past what’s generally anticipated and demonstrate conceivable MCI in the event that you experience any or the entirety of the below mentioned:

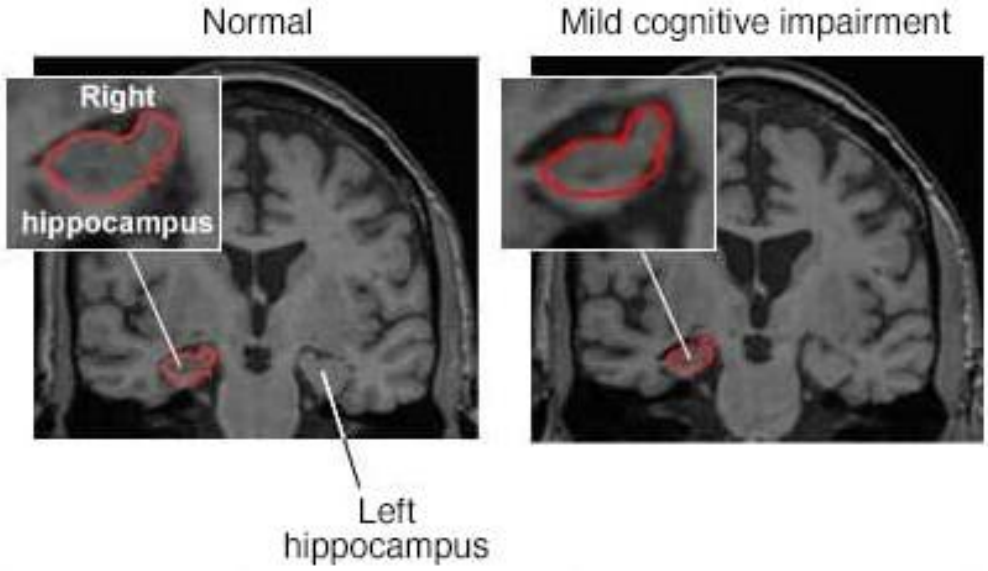


Fig 2 Comparison of Normal and MCI brain

(Source: https://www.mayoclinic.org/)

3. MAGNETIC RESONANCE IMAGING (MRI)

MRI also known as nuclear magnetic resonance imaging, is a scanning technique for creating detailed images of the human body. MRI is a clinical imaging innovation that makes definite pictures of your organs and tissues utilizing an attractive field and PC produced radio waves. Huge, tube-moulded magnets are utilized in most of MRI gear. The attractive field in a MRI machine realigns water atoms in your body for a brief time frame. These adjusted molecules give little signals that are utilized to build cross-sectional MRI pictures, like cuts in a portion of bread, utilizing radio waves. The MRI scanner can likewise make three-dimensional pictures that might be seen from different points of view.

3.1 Structural MRI Imaging:

The shape, size, and integrity of grey and white matter structures in the brain may be described subjectively and quantitatively using structural MRI. The MRI signal varies between tissue types since grey matter has more cell bodies (e.g., neurons and glial cells) than white matter, which is for the most part comprised of long-range nerve filaments (myelinated axons) and assisting glial cells. Morphometric strategies are utilized to decide the volume and state of dark matter areas like the subcortical cores and the hippocampus, just as the volume, thickness, and surface space of the cerebral neocortex. Macrostructural white matter trustworthiness can be estimated utilizing volumes of typical and strange white matter notwithstanding microstructural dissemination weighted MRI, giving signs of irritation, edema and supplementing microstructural dispersion weighted MRI to give a total image of white matter.

3.1.1 Different types of sequences

There are a variety of pulse sequences available, each focusing on a distinct feature of normal and diseased brain tissue. Anatomical pictures can emphasize contrast among grey and white matter or between mind tissue and cerebrospinal liquid by changing succession boundaries, for example, reiteration time (TR) and reverberation time (TE). Arrangements vary as far as the data they supply and, obviously, the time it takes to acquire it. Diverse picture preparing frameworks oftentimes require such successions, and may even supporter uniquely tuned arrangements for the best outcomes.

3.1.2 T1-weighted image.

T1-weighted image (also referred to as T1WI or the "spin-lattice" relaxation time) is one of the basic pulse sequences in MRI and demonstrates differences in the T1 relaxation times of tissues. A T1WI depends upon the longitudinal unwinding of a tissue's net polarization vector (NMV). Essentially, turns adjusted in an outside field (B0) are placed into the cross over plane by a radiofrequency (RF) beat. They then, at that point slide back toward the first balance of B0. Not all tissues return back to harmony in a similar measure of time, and a tissue's T1 mirrors the measure of time taken for its protons' twists to realign with the principle attractive field (B0). T1 weighting will in general have short TE and TR times.

Fat rapidly realigns its longitudinal polarization with B0, and it hence shows up brilliant on a

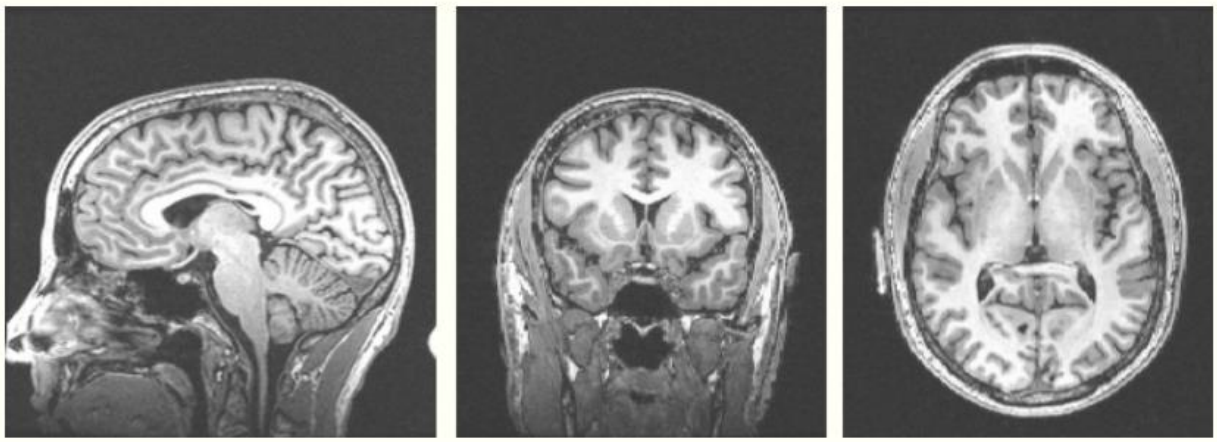


Fig 3 Gray matter(dark gray) and white matter (lighter gray) and CSF(void of signal)

(Source: http://fmri.ucsd.edu/Howto/3T/structure.html)

While CSF is devoid of signal, it provides strong contrast between grey matter (dark grey) and white matter (lighter grey) tissues (black). Water, like CSF, and thick bone, just as air, appear to be dark. Fat in the myelinated white matter, like lipids, shows up brilliant. The best difference is between the neocortex and white matter. The differentiation between some subcortical dark

matter cores and white matter is acceptable, yet not exactly as great as the difference between the cortex and white matter.

In cases of Caudate and putamen nuclei which have more white matter filaments and vascular foundation than other dim matter regions, bringing about expanded brilliance. Neurotic exercises that increment water content in tissues, like demyelination or irritation, diminish the sign on T1; white matter sickness usually shows up as more obscure patches in the lighter dark hued white matter. Broad white matter illness; T1 has moderate white matter infection. T2-weighted pictures are more touchy to humble white matter alterations because of a superior appraisal of water content.

3.1.3 Gray matter

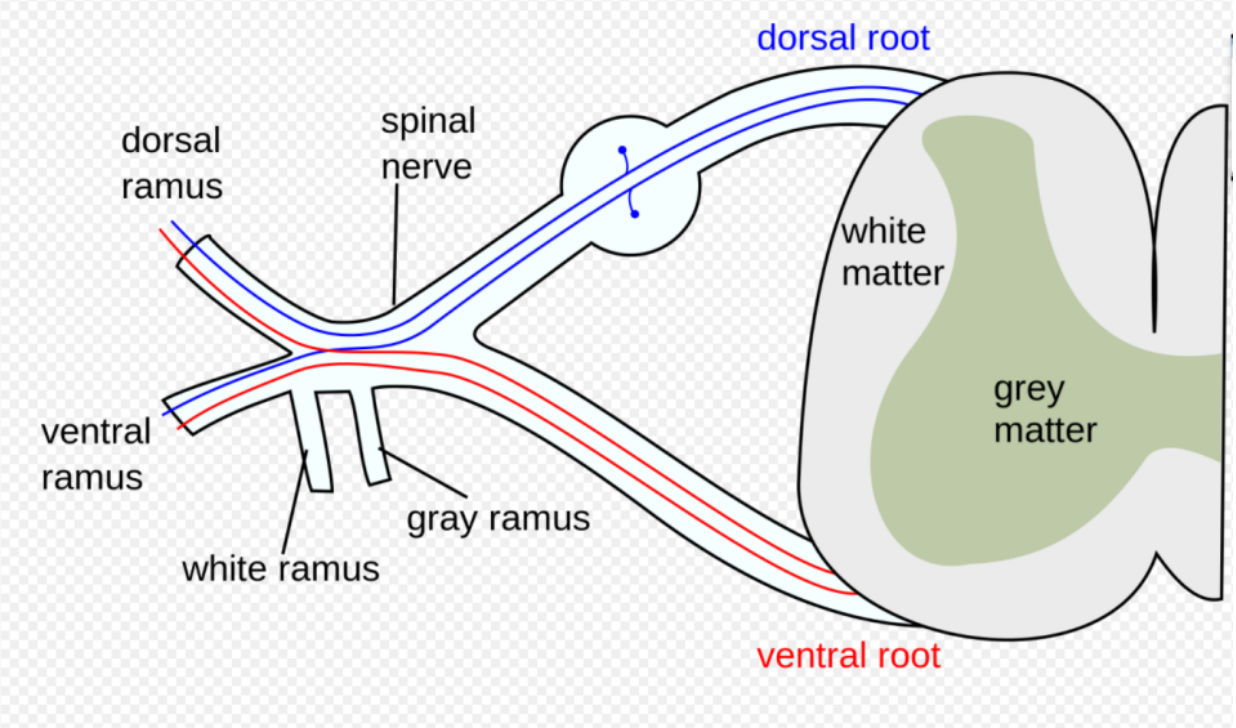


Fig 4 Gray matter labelled at centre right

(Source: https://en.wikipedia.org/wiki/Grey_matter)

Cell bodies ("grey matter") which stretch from the cell bodies make up brain tissue. The volume of grey matter in a certain brain area appears to be linked to a variety of abilities and capabilities.Both genetic and environmental variables, such as experience, influence cell density.The white matter controls our ability to digest information quickly.

Gray matter and white matter are two forms of brain tissue. These names are derived from the way they seem to the naked eye. The cell bodies of nerve cells make up grey matter. White matter is made up of long filaments that stretch from cell bodies, serving as the neural network's

“telephone lines,” conveying electrical signals that transport messages between neurons.The density of brain cells in a given location is measured by the volume of grey matter tissue.

3.1.4 CSF

In the human body there is a transparent body fluid present in the tissue around the vertebrates' brain and spinal cord, this fluid is termed as Cerebrospinal fluid (CSF). CSF is formed in choroid plexuses of the brain's ventricles. CSF is formed from blood plasma and is quite identical to it, with the exception that CSF is essentially protein-free and has slightly different electrolyte values than plasma. CSF has a greater chloride content than plasma and an equivalent sodium level due to the manner it is produced.

4. APPLIED SOFTWARE

4.1 MATLAB

The applied software is MATLAB. It is a programming language created by the MathWorks. Beginning of this was as a matrix language which has a unique feature that it be used for interactive sessions as well as batch work.

4.2 Statistical Parametric Mapping (SPM)

SPM implies building and assessment of spatially broadened factual cycles. This is used to check theories in regard to information imaging. Such developments has imbedded in a product which is named as SPM. The motivation behind fostering this was to research the psyche imaging information which are in arrangements. Present delivery is produced for checking of EEG, PET, MEG and fMRI.

4.3 CAT

This tool stash is an augmentation to SPM12 to give computational anatomy. This covers diverse morphometric methods such as voxel-based morphometry (VBM), surface-based morphometry (SBM) and region- or label-based morphometry (RBM).

4.4 VBM

VBM gives the voxel-wise assessment of the neighbourhood sum or volume of a particular tissue compartment. VBM is frequently applied to explore the neighbourhood conveyance of grey matter, however, can likewise be utilized to analyse white matter. Be that as it may, the affectability for discovering impacts in white matter is somewhat low and there exist more proper techniques for that reason.

PROBLEM STATEMENT

The objective of the study is to explore the use of machine learning algorithms to classify patients into AD, MCI and normal. The study aims to extract signature features from

volumetric MRI images using image processing algorithms .which can be used to discriminate between the classes. These features would then be used as inputs to different machine learning algorithms. The study also aims to rank the features using suitable ranking schemes in-order to identify features with the highest discriminative value between the classes These features based on their location in the brain would also be indicative of the location of the pathology for a large cohort of patients. The performance of different machine learning algorithms like Decision Tree, Random Forest, Gradient Boosting, XG Boosting, Light GBMwould be compared.

CHAPTER 2

LITERATURE REVIEW

This chapter's goal is to gather and then debate studies on the development of tools needed for objective Alzheimer's disease diagnosis. It also provides a summary of the concept that supports Machine Learning approaches.

1. ALZHEIMER'S DISEASE

Alzheimer's disease (AD) is a deadly neurological illness that causes issues with behaviour, memory, and thinking. It is the most frequent kind of dementia. It is one of the most expensive illnesses in affluent countries. According to the same survey, over two-thirds of dementia patients live in low- and middle-income nations. This will be complicated for a range of factors, including that of the fact that elderly patients in these nations rely heavily on informal care [7]. While some of its symptoms may resemble those of advanced age, it is vital to remember that Alzheimer's disease (and dementia in general) is not a natural component of the ageing process. Current therapies, on the other hand, can temporarily slow the progression of dementia symptoms if the condition is detected early. While better therapies and, eventually, prevention or even a cure, are critical. Except for a few examples with detectable genetic variations, the exact causation of Alzheimer's disease remains unclear. [8]The affectability/particularity of proposed imaging-based markers in determining distinct individuals as AD or typical has bit by bit improved. Such drives have utilized factual AI ways to deal with infer AD-related markers utilizing cerebrum examines as information[9]

However, current research suggests that it is linked to neurotic plaques and neurofibrillary tangles in the brain. While Amyloid beta, the protein that makes up neuritic plaques, is well established to have a role in the disease's progression, it is still debatable if it is a causal cause, as many think. It is, nevertheless, widely regarded as a disease indicator. [10] Advances in research have occurred in recent years, most notably the finding of biomarkers (particularly brain imaging methods) that allow for the detection and observation of AD-related processes months, years, and even decades before clinical symptoms manifest. Early biomarkers, which often measure amyloid accumulation in the brain, and later biomarkers, which often measure neurodegeneration, can be distinguished. Brain imaging scans are frequently performed to rule out other possible explanations of symptoms, but they can also reveal whether or not Alzheimer's disease is present. Both neurofibrillary tangles and neuritic plaques, on the other hand, appear to have a role in the onset and progression of Alzheimer's disease.(Biomarker paper)While the specific causation of Alzheimer's disease is unknown, many risk factors are obviously linked to the illness's progression. Alzheimer's disease is firmly linked to advancing

age. Every five years beyond the age of 65, the risk of Alzheimer's disease doubles, reaching nearly 50% by the age of 85.Close relatives who have been diagnosed with Alzheimer's disease are more prone to get the condition themselves. The chance of contracting the sickness increases if more than one member of the family is affected. When illnesses tend to run in families, both inherited and environmental factors may play a role. [11] Despite the fact that much research has been done on Alzheimer's disease, there is still a need for an early (non-invasive) diagnostic tool.

2. MILD COGNITIVE IMPAIRMENT

Before slipping into Alzheimer's disease, most individuals go through a period known as mild cognitive impairment (MCI). MCI is defined by a slight impairment in memory and executive function that is not severe enough to impede with everyday activities but noticeable to the individual and others.(Classification of Alzheimer's Patients Using Structural MRI Data Yoon-Suk Han, Wyatt Hon). A physician's clinical diagnosis is currently the sole way for a patient to be diagnosed with MCI or AD.

3. BIOCHEMISTRY AND BIOMARKERS

The build up of abnormally folded amyloid beta (-amyloid) protein in the brains of Alzheimer's patients has been recognised as a protein misfolding disorder. [60, 9]. Plaques are formed when sticky amyloid particles cluster together (these are also called neuritic plaques). These plaques appear to inhibit signalling (i.e. communication) between cells, triggering immunological responses that result in the impaired neurons' programmed cell death.Amyloid plaques are "a characteristic sign of a pathological diagnosis of AD" [12], and biomarkers that can detect andquantify the amyloid build-up in the brain reflect this. In cerebrospinal fluid (CSF) and plasma, the protein may be tested directly. [12] Due to aberrant aggregation of the tau protein, a microtubule7-associated protein produced in neurons that typically works to support microtubules in the cell cytoskeleton [14] , AD is often referred to as a tauopathy. Normally, tau straightens microtubules, allowing molecules to travel easily.The protein in Alzheimer's disease, on the other hand, collapses into twisted strands (i.e. tangles), which cause the tubes to dissolve, preventing nutrition from reaching nerve cells and resulting in cell death. [14] The build-up of -amyloid plaques and neurofibrillary tangles is thought to cause to the loss of neurons and synapses (brain structural breakdown), resulting in "memory impairment and other cognitive difficulties"[16]. Later it was a part of research that on investigating the links between cognitively normal ageing, extremely mild Alzheimer's disease, and the presence of neocortical senile plaques. Their findings indicate that neurotic plaques may reflect presymptomatic or

undetected early symptomatic Alzheimer's disease, rather than being a natural component of ageing. [17] The only significant difference between the normal control group and those with mild Alzheimer's disease was the frequency of neuritic plaques.

As a result, the sensitivity of neurofibrillary tangles as a marker of Alzheimer's disease was shown to be lower than that of neuritis plaques. [18)]Changes in the brain (i.e. structural abnormalities) create the symptoms that arise throughout the progression of Alzheimer's disease, which is a sensitive aspect of the illness. Early on in the course of Alzheimer's disease, the hippocampi of people with the condition atrophies, a characteristic that may be consistently recognised by volumetric (i.e. structural) MRI for diagnostic reasons. [19]CSF measurements can be utilised to predict AD, according to Hansson et al. (20).Recent investigations have revealed that imaging parameters based on brain scans are more consistent and sensitive measures of AD diagnosis and development than cognitive assessment, according to Ye et al. [21] In their 2011 study. They also described how neuroimaging techniques such as structural magnetic resonance imaging (sMRI) were used to evaluate particular structures such the hippocampus, entorhinal cortex, and amygdala, as well as any aberrant volumetric changes associated with Alzheimer's Disease. [22] Cortical atrophy in the temporal, frontal, and parietal regions is a common symptom of the condition, which is caused by degeneration of the cerebral cortex and hippocampus.[23]

Long before the first indications of memory loss, microscopic changes in the brain occur. A recent study discovered the existence of medio temporal lesions up to 5.6 years before the onset of dementia. Alzheimer's disease is diagnosed clinically. Biomarker magnitudes approach abnormal levels in a regular sequence as Alzheimer's disease advances. Some of the biomarkers can be listed as, Amyloid beta imaging has been identified in the CSF, as well as PET amyloid imaging. Increased levels of tau species in the CSF indicate neurodegeneration. MRI measurements of brain atrophy and cell loss (most notably in the hippocampus, caudate nucleus, and medial temporal lobe), Cognitive testing measures memory loss.[25]

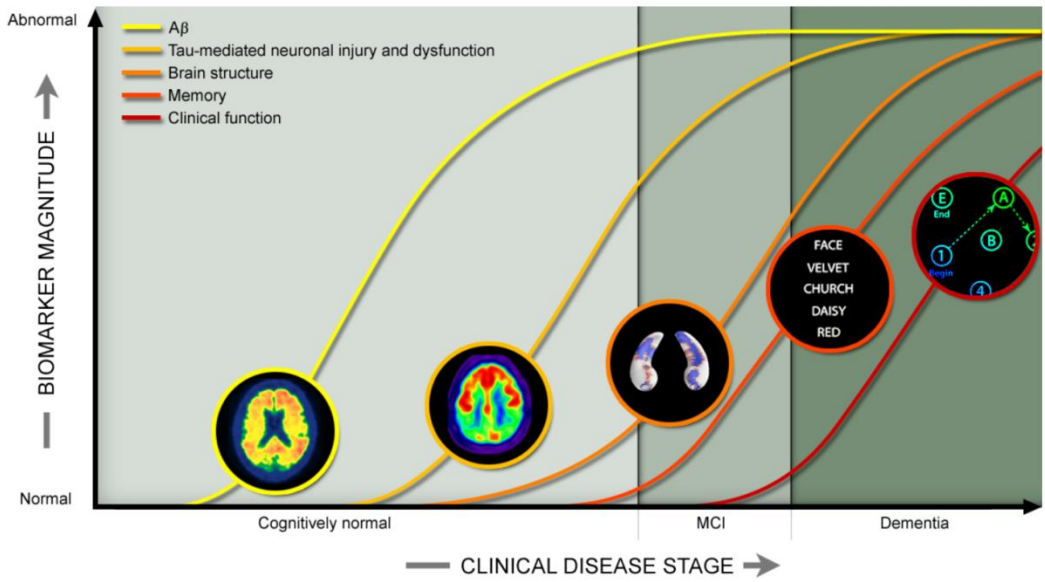


Fig 5: Alzheimer's biomarkers as the illness progresses

(Source : http://adni.loni.usc.edu/study-design/)

The hippocampus (which is crucial for memory formation) is where the destructive coupling of plaques and tangles begins. The earliest indication of Alzheimer's disease is frequently short-term memory loss. The proteins then spread to other areas of the brain, causing distinct alterations that signify different stages of illness. At the front of the brain, the ability to process rational reasoning is destroyed. Ingesting. Then there was a loss of emotion regulation, which resulted in irregular mood swings. Then it causes paranoia and hallucinations (at the top of the brain). The proteins then work together at the back of the brain to delete the mind's deepest memories. Finally, the control centres (which regulate heart rate and respiration) become overwhelmed, culminating in death. The situation is still unclear. Disease has a tremendously destructive tendency. [25]

4. MRI

In contrast to traditional X-ray and CT scans, magnetic resonance imaging (MRI) does not use ionising radiation. The imaging method may produce three-dimensional pictures in any depth and orientation. It provides non-invasive pictures at high resolution that are significantly better than other imaging technologies, and it is the diagnostic tool of choice for soft tissue. Although it is not always necessary, MRI patients may be given a contrast agent to help abnormalities (such as tumours) stand out more clearly. The imaging method simply maps the location of water molecules, which occur in various concentrations in various types of tissue. [26]

A positively charged proton spins around an axis in the nucleus of every hydrogen atom in a person's body's water. As a result, it produces its own magnetic field. When these atoms are exposed to a stronger magnetic field, the protons' axes realign to face the stronger field: about half of them face the field, while the other half face the other way. In the low-energy configuration (north-south north-south), a few more atoms align with the magnetic field than in the higher-energy configuration (north-north south-south). In a 1.5 tesla (the field strength most often used in clinical practise.) magnet driven MRI, "a few" represents around five out of one million protons [27] The MRI scanner will employ the few "leftover" protons that aren't cancelled out by a proton lined up in the other way. When these hydrogen nuclei are exposed to a stronger magnetic field, they "spin counter clockwise around the magnetic field direction like gyroscopes"[28], a phenomenon known as Larmor precession. Along with the patient, a radio frequency coil is installed inside the machine, which transmits radio frequency pulses (RF pulses) to the targeted spot. The pulses are carefully timed and modified to a certain range of frequencies (using the Larmor frequency) at which Hydrogen protons spin counter clockwise.[28]

As described by the Larmor frequency, the frequency of Larmor precession is proportional to the applied magnetic field strength.

$\omega_0 = \gamma B_0$

Where γ is gyromagnetic ratio and B_0 is the strength of the applied magnetic field. For hydrogen γ = 42.6MHz/Tesla [30]

The energy of the RF pulses is absorbed by the "leftover" protons, causing them to flip on their axes (still in line with the magnetic field, but now in the other direction: the high-energy configuration). In other words, once the RF pulses are switched on, unmatched protons flip.When the RF pulse ceases, the absorbed energy is released and the protons return to their original alignment. They do so by sending a signal back to the coil.

4.1 SMRI

Structural magnetic resonance imaging (sMRI) is valuable in deciding the anatomical changes related with Alzheimer’s illness, particularly in its beginning phases. Conventional strategies depend on space specialists' abilities to extricate hand-picked factors like grey matter substructures. [31]

Clinical neuroimaging is becoming more common in the diagnosis of individuals who appear to a memory clinic. Methods of structural neuroimaging, such as computed tomography (CT) and, more sensitively, magnetic resonance imaging (MRI), are well adapted to exclude potentially (surgically) curable dementia causes. More significantly, structural brain imaging can reveal specific patterns of atrophy, vascular disease, or inflammatory abnormalities, all of which can help to confirm a dementia diagnosis. [32]

In contrast to utilising functional magnetic resonance imaging (fMRI) to study brain activity, structural magnetic resonance imaging (MRI) is a non-invasive approach for assessing the architecture and disease of the brain. This generates pictures that may be utilised for both clinical radiological reporting and in-depth study. [33]

Because of the absence of radiation, superior grey matter/white matter contrast, and the flexibility to control tissue contrast with varied pulse sequences, structural MRI is the imaging technique of choice. [34] The MRI procedure should include pulse sequences that measure local and global cortical atrophy, as well as medial temporal lobe atrophy (especially the hippocampus);vascular alterations in white matter (small and big vessel disease); vascular changes in deep grey matter structures (especially thalamus infarction); cerebral microbleeds (MBs) and microhaemorrhages (including post-traumatic alterations).

Some MRI scan sequences are volumetric, which means that measurements of certain brain regions may be used to compute tissue volumes. In addition, these scans may be recreated in any plane. During childhood and adolescence, the quantities of regional grey and white matter (which make up the brain) alter dramatically, and they may alter again in old life. [35] In people with ASD, structural MRI investigations have consistently found abnormalities in cortical grey and white matter volume. DTI was used to reveal white matter anomalies at a microstructural level. [36]

4.1.1 Use of MRI to support the diagnosis of neurological disease associated with dementia.

The histological characteristics of amyloid- and hyperphosphorylated tau buildup define AD, which is the most frequent cause of dementia. [37] The importance of neuroimaging in the diagnosis of Alzheimer's disease is growing. This is reflected in new proposed diagnostic criteria, which include a structural imaging marker (medial temporal lobe atrophy) in addition to cerebrospinal fluid (CSF) (amyloid-42, tau, or phospho-tau) and positron emission tomography (PET) findings (temporo-parietal hypometabolism on FDG-PET, amyloid

imaging). [38] According to the phases of neuropathology. The medial temporal lobe, particularly the hippocampus, shows the first signs of neurodegeneration in terms of shrinkage in senile AD patients, which may be observed best on oblique coronal T1-weighted imaging. Although MTA is a hallmark of Alzheimer's disease, a normal medial temporal lobe volume does not rule out the disease. Furthermore, MTA is a frequent characteristic of other neurodegenerative illnesses, hence it is not a reliable diagnostic for excluding other neurodegenerative diseases linked to dementia.Aside from the diagnostic signature MTA, cortical atrophy, particularly in the parietal lobes, is a frequent radiological sign that may help distinguish Alzheimer's disease from other dementia-related neurodegenerative disorders. [39] Early-onset (pre-senile) AD may appear with an unique atrophy pattern including the parietal cortex, precuneus, and posterior cingulum and sparing the medial temporal lobe, as opposed to traditional senile AD signs.As the disease progresses, the medial temporal lobe gets increasingly afflicted. [40]The degree of cortical and hippocampal shrinkage determined by visual rating has a substantial predictive value for subsequent cognitive decline and the development of AD in individuals with moderate cognitive impairment (MCI) who present with cognitive impairment but do not meet the diagnostic criteria for dementia.(Likeman et al., 2005; Bouwman et al., 2007; Davatzikos et al.; 2010).The degree of cortical and hippocampal shrinkage determined by visual rating has a substantial predictive value for subsequent cognitive decline and the development of AD in individuals with moderate cognitive impairment (MCI) who present with cognitive impairment but do not meet the diagnostic criteria for dementia.[41]

5. ORGANISATION

This section covers organisations that strive to improve Alzheimer's disease research and knowledge, including several that create and distribute clinical data. without whom the dataset would not be available for research purposes, work mentioned in this thesis was also impossible to carry out.

5.1 The Alzheimer’s Disease Neuroimaging Initiative

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a long-term research project with the goal of "developing clinical, imaging, genetic, and biochemical biomarkers for the early diagnosis and monitoring of Alzheimer's disease" The ADNI1 trial began in 2004 with 400 participants with MCI, 200 participants with early AD, and 200 older control participants.The global effort supplies researchers with trustworthy clinical data on Alzheimer's disease. They hope to use biomarkers to better understand the pathology of Alzheimer's disease, potentially allowing for earlier diagnosis; to provide clinical trial data to support new research methods

pertaining to intervention, prevention, and treatment of the disease; and to provide clinical trial data to support new research methods pertaining to intervention, prevention, and treatment of the disease. The goal of ADNI1 was to develop more reliable diagnostic tools for detecting Alzheimer's disease in its early stages and to identify the pathology using biological markers. The programme was effective in developing early stage detective tools for Alzheimer's (including CSF biomarkers, -amyloid 42, and tau), as well as standardised clinical diagnostic procedures (including MRI, PET and CSF biomarkers). [42] ADNI is a significant step forward in the development of new diagnostic tools, as well as the potential creation of effective therapeutics to slow the course and, ultimately, prevent Alzheimer's disease.

6. CLASSIFICATION OF DISEASE

The most frequent data mining approach is classification, which involves separating data into non-overlapping pieces. Forecasting a discrete value can be considered as classification. Any categorization method demands some understanding of the data. As a result, a training set is employed to identify certain parameters. Training data necessitates sample input data, domain expertise, and data categorization. [44] We employed five different Machine Learning techniques in this study, including Decision Tree [44], Gradient Boosting , XG Boosting, Light GBM, Random Forest, and kNN, to classify various phases of Alzheimer's disease.

6.1 Machine learning

In general, a machine learning system consists of three components: training data, a model, and a training (i.e. learning) algorithm. The learning is then stated as fitting the model's parameters to the provided training data. However, this is not always the case, as in the case of decision trees (which are more of a generative approach). To assess the trained model's true performance, it is next applied to test data, which is a subset of the previously concealed data. The learning process's objective is for the system to be able to generalise from its experience (with training data) and perform well on previously encountered instances of data. Machine learning is classified into three learning paradigms: supervised and unsupervised learning, as well as reinforcement learning. Supervised learning works using labelled example data, in which the inputs are linked to the intended output values. This is analogous to the psychological idea of concept learning. Unsupervised learning, on the other hand, uses unlabeled learning instances (i.e., no known output). As a result, there is no error or reward signal to assess a potential solution with. Reinforcement learning is the process of striving to achieve a certain goal by executing an action in a dynamic environment in order to maximise a reward, without being expressly notified if the learner is getting close to its aim.In supervised and unsupervised

learning, there is a large range of machine learning models, all of which make distinct prior assumptions about the possible input-output mappings or data distribution. [45]

When deciding which algorithm to use for a given problem (and a given target function), consider the relationship between the size of the hypotheses space, its completeness, the number of training examples available, the learner's prior knowledge, and the confidence one can have that a hypothesis that is consistent with training data will correctly generalise to unseen examples.

Machine Learning and Knowledge Discovery from Databases (KDD) are increasingly being used in health care to build models, create practise recommendations, or enhance existing standards for improved medical decision making. They vary from traditional techniques in that they generate domain models from data, such as decision trees, decision rules, graphs, and so on. Machine learning approaches aim to learn a description that best differentiates the various phases of Alzheimer's disease. (46) Each of the first visits is represented as input by the characteristics indicated above. Each of the learning approaches produces a representation of the various stages of Alzheimer's disease that may be used to categorise individuals with an unknown Alzheimer's stage. To characterise the probable outcomes of the diagnosis, each learning approach employs a distinct search methodology and concept representation. We used each of the five ML algorithms in the proposed system to classify different phases of Alzheimer's disease. [47]

6.1.1Decision Tree

The decision tree method is a predictive modelling tool that is extensively used in data mining, statistics, and machine learning applications for categorization. It classifies the dataset by computing the information gain values for all of the dataset's properties.A class label is represented as a leaf node in a tree, while the branches to these leaf nodes represent the combination of input variables that lead to those class labels (Shahbaz et al., 2013) [49]

6.1.2Random Forest

Random Forest is a classification system that consists of a large number of decision trees. It develops the individual tree using feature randomness and catching, and then creates an uncorrelated forest of trees whose prediction is more accurate than the individual tree. (2019, Prabu.S.) [50]

In this paper , researcher have used an approach based on ANN to detect the Alzheimer’s condition from MRI scans. This research generally uses GLCM approach feature extraction of

features. They took hippocampus region of MRI scan as first region to get affected from this disease.[29]

In this paper, researcher has applied machine learning techniques on structural MRI data from ADNI website to categories patient into three categories that is normal, MCI, and AD. He applied Logistic regression model for classification and got accuracy as 76.34 % for binomial model and 59.8% for multinomial model.[13]

In this paper , research used the data from ADNI website and used around 2731 scans of 657 patients. Here they used CNN architecture performances on the dataset and got more than 95 % accuracy.[15]

This study shows numerous models for classifying different phases of Alzheimer's disease using machine learning approaches such as Neural Networks, Multilayer Perceptron, Bagging, Decision Tree, CANFIS, and Genetic algorithms. The formalised paraphrase. The classification accuracy of CANFIS was determined to be 99.55 percent, which was reported to be higher than the industry average, various techniques of categorisation. [48]

CHAPTER 3
METHODOLOGY

1. DATA

1.1 Data Collection

The ADNI is a key source for public data of AD research. ADNI is a multi-center, longitudinal research institute that gather data from Alzheimer's patients across the United States. The biomarkers present in the patients from over 1,000 participants who participated, are included in the collection, which generally includes personal history, cognitive evaluations, genomic sequencing, MRI, and PET scans.

We have worked primarily with T1-W1 structural MRI scans of around 93 healthy older person, 159 mild cognitive impairment patients, and 37 Alzheimer's disease patients (total of 289 subjects).

1.2 Dataset Description

T1 weighted MRI data will be used.

The dataset comprises of T1 weighted MRI data from around 289 patients ranging in age from 60 to 96. Each subject was scanned at least once.

Throughout the trial, 93 of the individuals were labelled as 'Nondemented.'.

When they first visited, 159 of the individuals were classified as 'MCI,' and they remained such throughout the research.

37 participants were classified as 'AD' at first visit and were classified as 'AD' while later visit also.

1.3 Data Processing

We translated the structural T1 weighted MRI scans from the ADNI website from DICOM to niftii for each of the patient.

We retrieved the structural T1W1 MRI images collected during the time of study inclusion for each of the patient. We next spatially normalised the brain images to the MNI space using SPM and CAT12 tools in MATLAB to account for changes of the brain size and shape across the different subject group. Then, using MATLAB, SPM12, and CAT12, we completed the following steps:

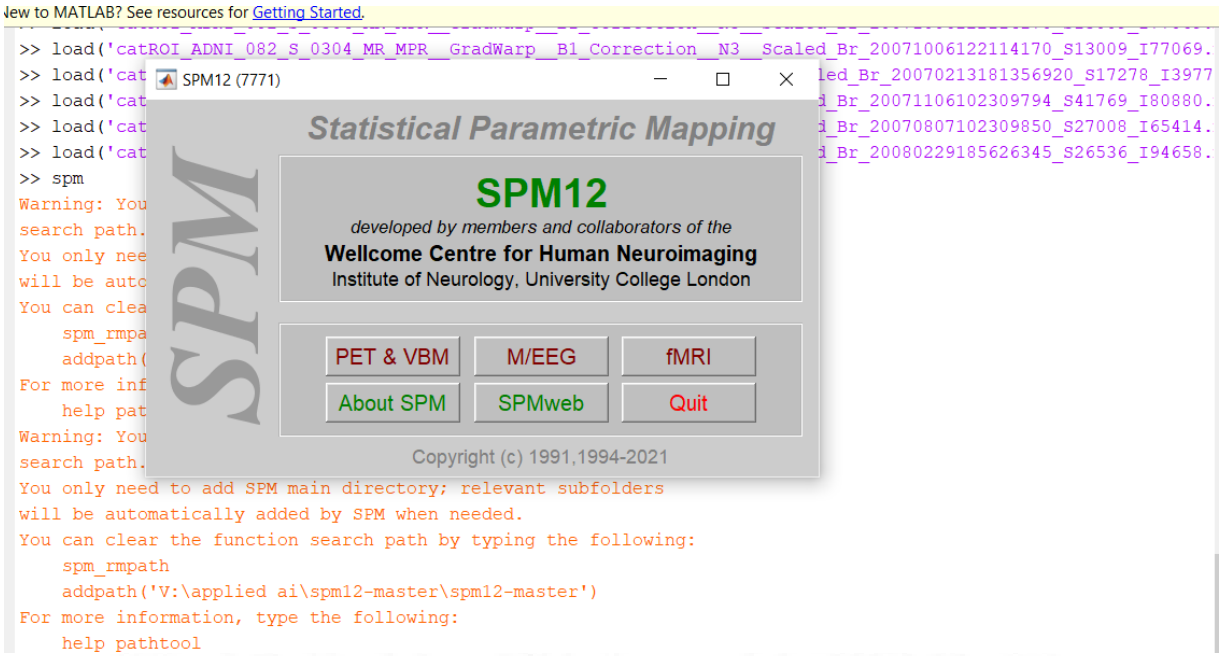


Fig 6: SPM12 toolbox

Choose “cat12” from the SPM menu. The drop-down menu located between the “Display” and “Help” buttons. This will launch the CAT Toolbox in a new display window.

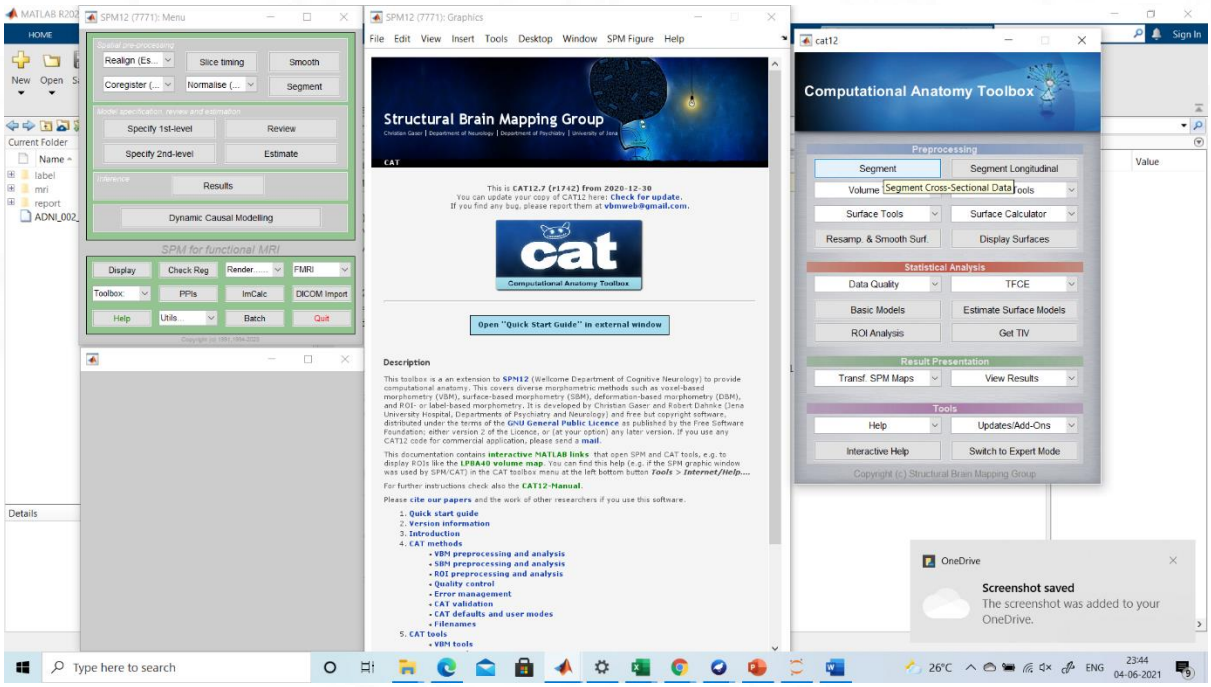


Fig 7: CAT12 toolbox

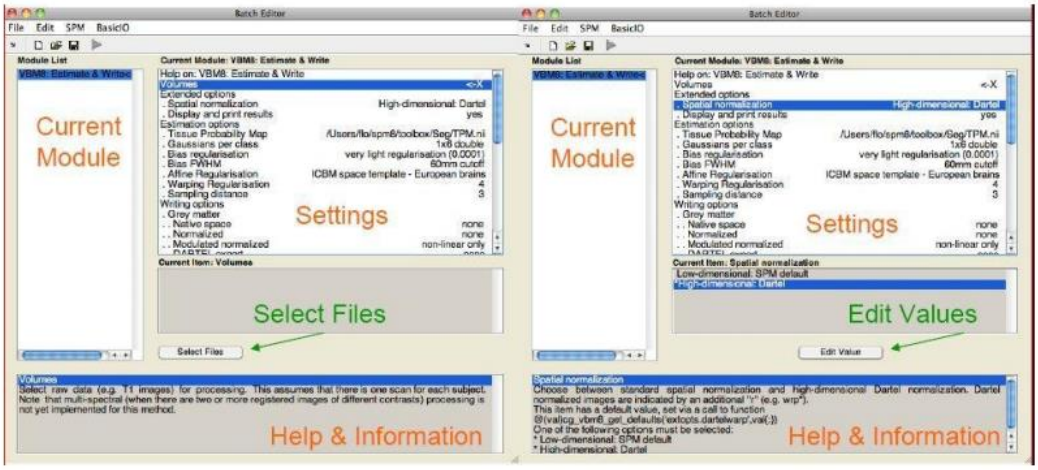


Fig 8: The Batch Editor.

Figure on the left side isto select the desired file from “Select Files” icon, and on theright side it shows “Edit Value” from where we can edit things.

1.4 BASIC VBM ANALYSIS (OVERVIEW)

The CAT Toolbox generally comes with different types of modules, which can be used for different types of analysis. Usually, a VBM analysis includes these different steps:

1.4.1 Pre-processing

1. T1 weighted images are standardised to a template MNI space and classified as GM, WM, or CSF. The pre-processing parameters can be changed using the given icon named as "Segment Data" module.

2. . A quality check is strongly advised when the pre-processing is completed. This is possible by using the modules "Display slices" and "Check sample." Both choices may be found under "Check Data Quality" in the CAT12 box. Additionally, during pre-processing, quality description are calculated and recorded in xml-files for each and every data set. These quality criteria are also presented on the report given onpdf page and may be utilised in the "Check sample" module.

3. Image data must be flattened before adding GM pictures into a statistical model. This stage, it should be noted, is not applied in the CAT Toolbox, nevertheless is accomplished using the standard SPM module named as "Smooth."

1.4.2 MNI Normalization:

Concept of Co-registration

Normalization is the process of coregistering a subject's (primarily anatomical) picture to a standard template in order to solve the issue of brain shape variations between people.

1.4.3 Denoising:

MRI denoising is a traditional preprocessing procedure that seeks to reduce the noise that is naturally present in MR images. Denoising is simply the act of downsampling high-resolution MRI data and then discarding the high-resolution information using low pass filtering.

1.4.4 Segmentation:

Image segmentation is a vital step in many clinical applications and is one of the most significant jobs in medical image processing. Image segmentation is extensively utilised in brain MRI analysis for measuring and visualising anatomical features, assessing brain changes, identifying diseased areas, and surgical planning and image-guided procedures. In the last few decades, various segmentation techniques of different accuracy and degree of complexity have been developed and reported in the literature.Picture segmentation is the process of dividing an image into semantically relevant, homogenous, and nonoverlapping sections with comparable qualities such as intensity, depth, colour, or texture. The segmented picture is either an image of labels identifying each homogenous region or a set of contours describing the region borders.

Because segmentation implies classification, and a classifier implicitly segments an image, the challenges of segmentation and classification are inextricably intertwined. Image components in brain MRI are commonly divided into three major tissue types: white matter (WM), grey matter (GM), and cerebrospinal fluid (CSF).

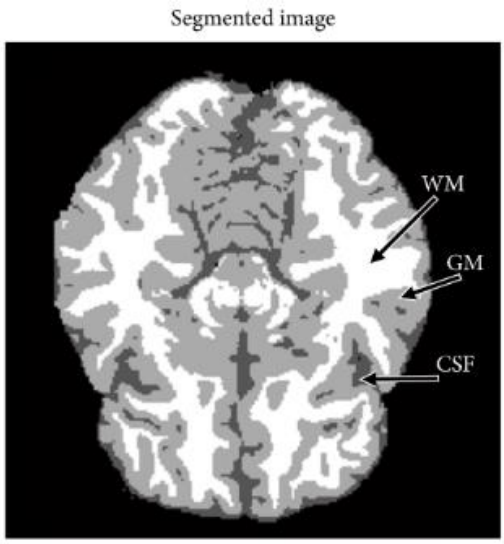


Fig 9 : Segmented Image

1.4.5MRI smoothing:

The primary disadvantage of smoothing data is the loss of geographical specificity. Smoothing, in other words, spreads out each subject's signal in space, so we can't be as definite about its

position. This may or may not be an issue depending on the spatial extent of activation you are interested in. Despite this loss of spatial specificity, there are various reasons why smoothing MRI data is beneficial. These may be divided into two categories: statistical reasons (smoothing aids in detecting activation) and inferential reasons (smoothing influences how you interpret your results).Although the recorded MRI signal has intrinsic spatial correlation, spatially smoothing the data induces an even larger degree of correlation. Each voxel is a weighted average of its own and its neighbours' pre-smoothed values after smoothing.

1.5 Overview regarding CAT12 processing steps:

CAT12 most important processing step:

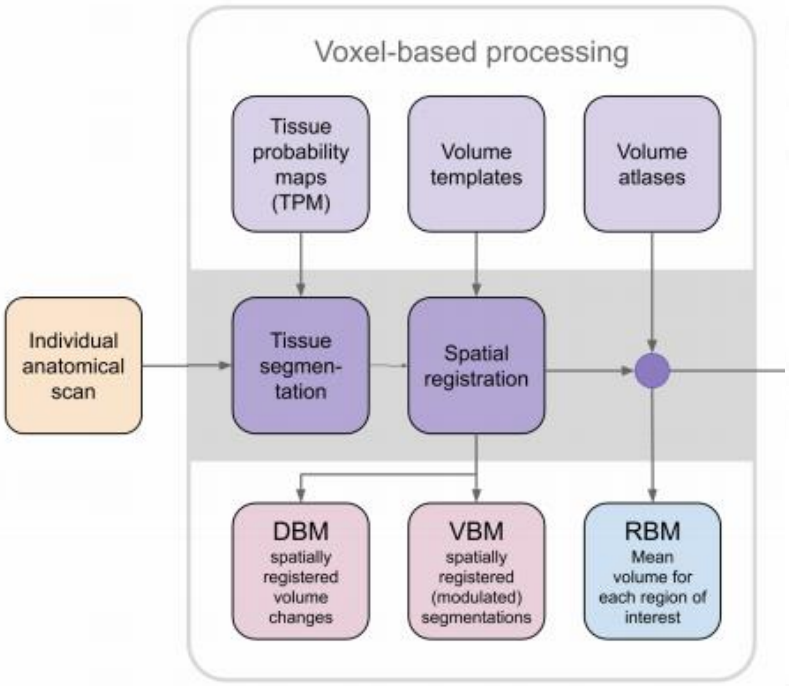


Fig 10: Overview of CAT's major processing steps

(Source:http://www.neuro.uni-jena.de/cat12/CAT12-Manual.pdf)

As shown in Figure , CAT's processing workflow comprises of two main steps: voxel-based processing and surface-based processing. The former is a prerequisite for the latter, but not the other way round. That is, while voxel-based processing is always required for surface-based analyses, users not interested in surface-based analyses can simply omit this second step to save processing time. The 'Voxel-based processing' step can be thought of as one module for tissue segmentation and another one for spatial registration. An optional third module allows for the generation of ROIs and the calculation of ROI-based measures. An optional third module allows for the generation of surfaced-based ROIs and the calculation of ROI-based

measures. Voxel-based processing: While the final tissue segmentation in CAT is independent of tissue priors, the segmentation procedure is initialized using Tissue Probability Maps (TPMs). The standard TPMs (as provided in SPM) suffice for the vast majority of applications, and customized TPMs are only recommended for data obtained in young children. Please note that these TPMs should contain 6 classes: GM/WM/CSF and 3 background classes. For spatial registration, CAT uses DARTELor Geodesic Shooting with predefined templates. On that note, the aforementioned creation and selection of a customized DARTEL or Geodesic Shooting template will disable the third module for the voxel-based ROI analysis.

1.6 Brief of CAT12 major process steps

CAT12 PROCESSING STEPS IN DETAIL

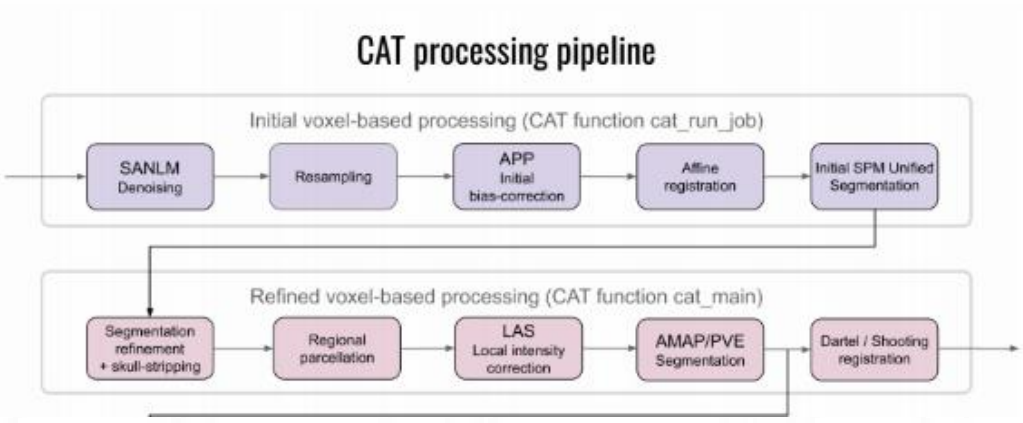


Fig 11: flowchart of CAT's processing pipeline.

(Source: http://www.neuro.uni-jena.de/cat12/CAT12-Manual.pdf)

The modules described in the previous section help understand the CAT's overall processing workflow, including its priors, templates, and atlases. Also, data processing in CAT can be separated into three main processes: [1] the initial voxel-based processing, (2) the main voxel-based processing, and (3) the surface-based processing (optional).

The 'initial voxel-based processing' begins with a spatial adaptive non-local means (SANLM) denoising filter, which is followed by internal resampling to properly accommodate low-resolution images and anisotropic spatial resolutions. The data are then bias-corrected and affine-registered (to further improve the outcomes of the following steps) followed by the standard SPM "unified segmentation". The outcomes of the latter step will provide the starting estimates for the subsequent refined voxel-based processing.

The 'refined voxel-based processing' uses the output from the unified segmentation and proceeds with skull-stripping of the brain. The brain is then parcellated into the left and right

hemisphere, subcortical areas, and the cerebellum. Furthermore, local white matter hyperintensities are detected (to be later accounted for during the spatial normalization and cortical thickness estimation). Subsequently, a local intensity transformation of all tissue classes is performed, which is particularly helpful to reduce the effects of higher gray matter intensities in the motor cortex, basal ganglia, or occipital lobe before the final adaptive maximum a posteriori (AMAP) segmentation.

1.6.1 Some Major toolbox of CAT12

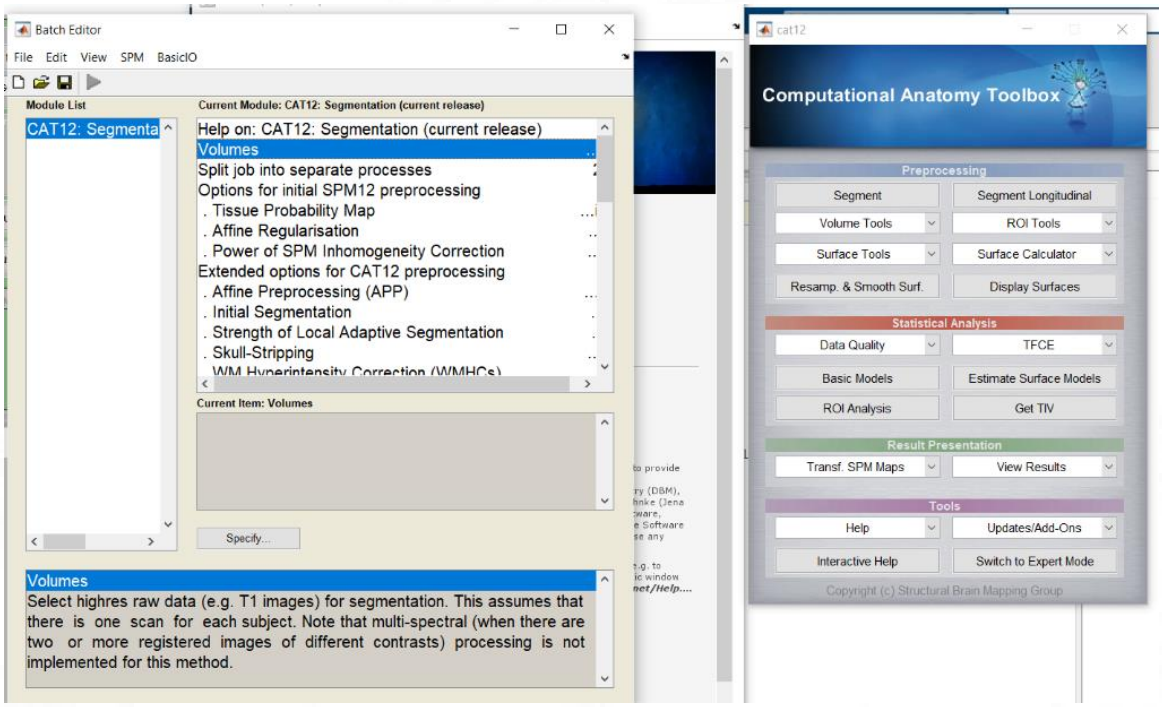


Fig 12 Volume

1.6.2 Volumes:

Select higher raw data (e.g. T1 images) for segmentation. This assumes that there is one scan for each subject. Note that multi-spectral (when there are two or more registered images of different contrasts) processing is not implemented for this method.

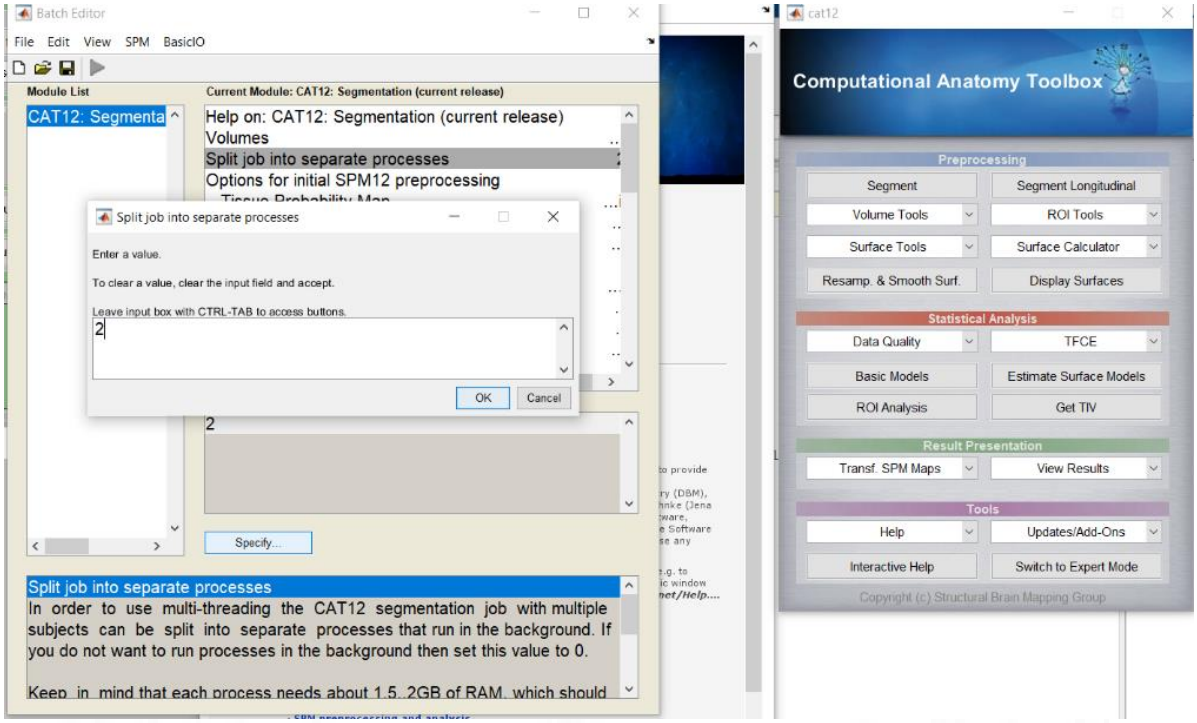


Fig 13: Split job into separate processes

1.6.3 Split job into separate processes

In order to use multi-threading the CAT12 segmentation job with multiple subjects can be split into seperate processes that run in the background if you do not want to run processes in the background then set this value to 0.

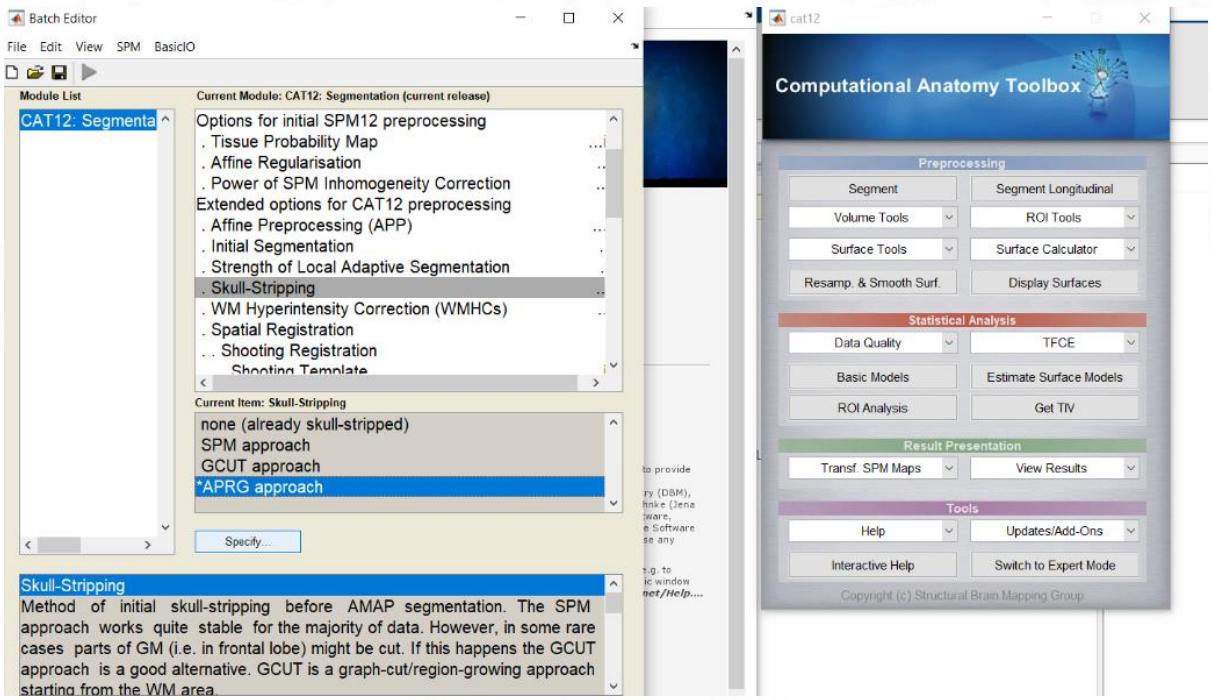


Fig 14: Skull -stripping

1.6.4 Skull -stripping

Method of initial skull-stripping before AMAP segmentation. The SPM approach works quite stable for the majority of data. However, in some rare cases parts of GM (in frontal lobe) might be cut.

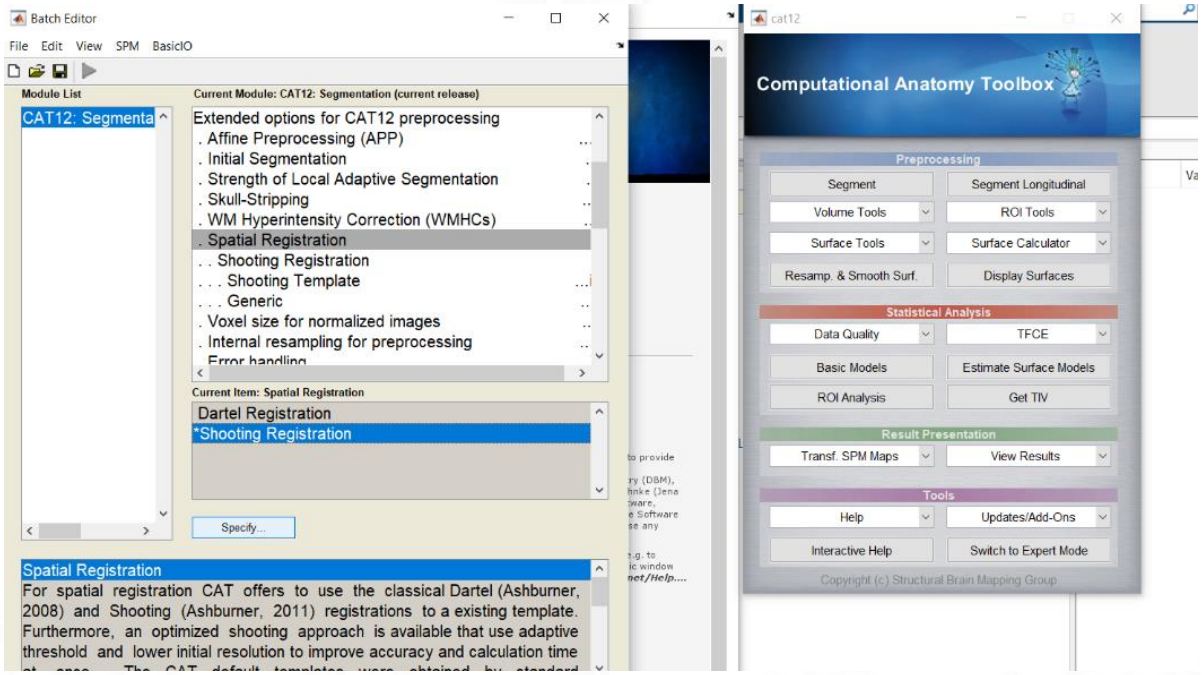


Fig 15:Spatial Registration

1.6.5Spatial Registration

For this registration CAT offers to use the classical Dartel and shooting registrations to a existing template. Furthermore, and optimised shooting approach is available that use adaptive threshold and lower initial resolution to improve accuracy and calculation time at once. CAT default templates were obtained by standard aur shooting registration of 555, lXl subjects between 20 and 80 years. The registration time is typically about 3, 10 and 5 minutes for Dartel, shooting and optimised shooting for the default registration resolution.

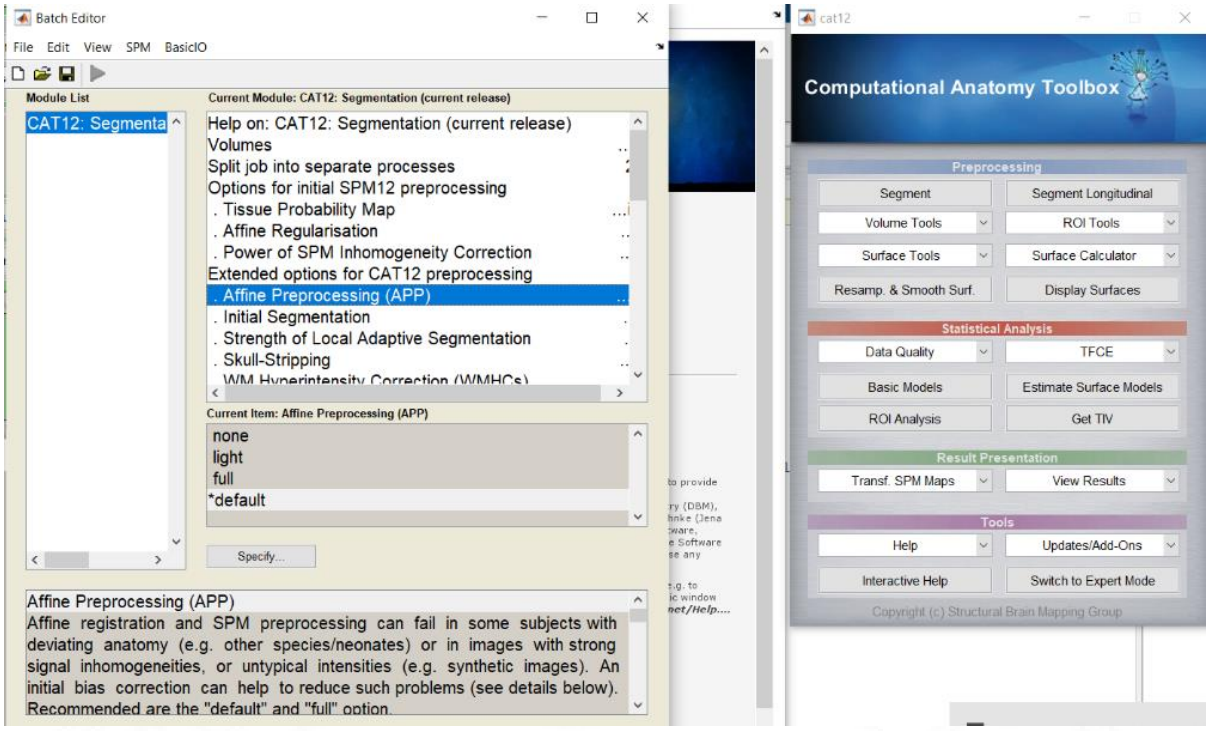


Fig 16: Affine Preprocessing

1.6.6 Affine Preprocessing

Affine registration and SPM preprocessing can fail in some subjects with deviating anatomy (e.g. other species/ neonates) or in images with strong signal inhomogeneities, or untypical intensities (e.g. synthetic images). An initial bias correction can help to reduce such problems.

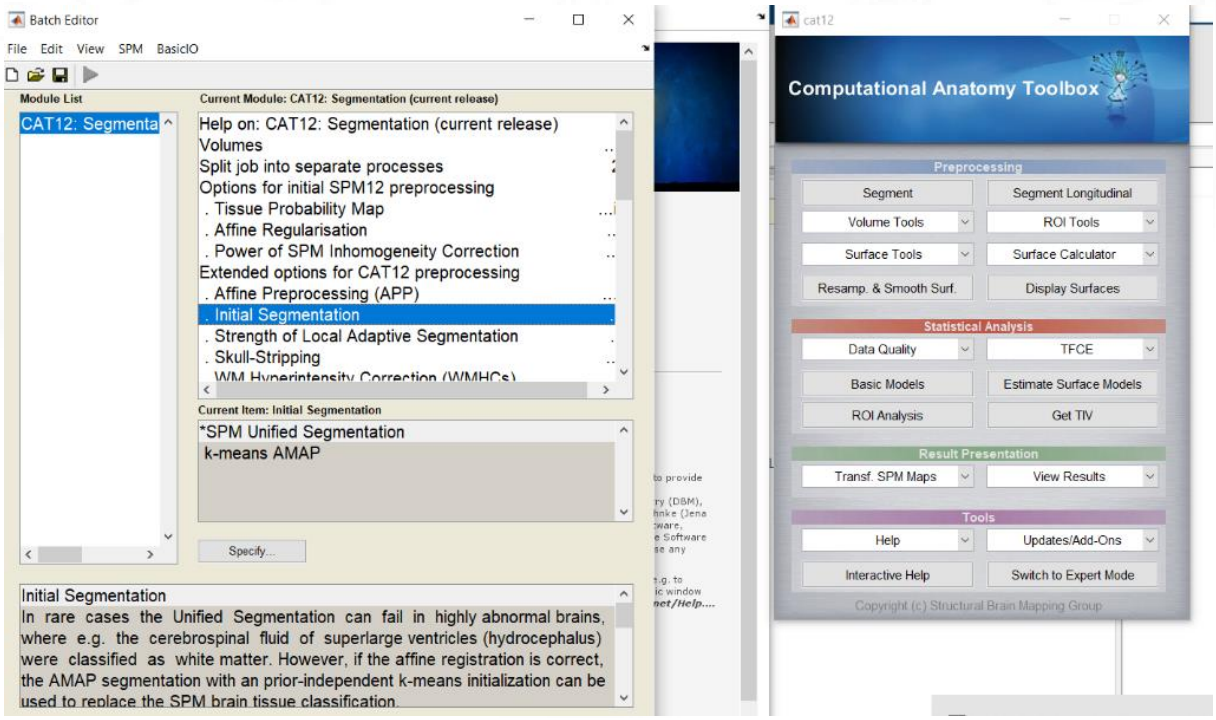


Fig **17 Initial Segmentation**

1.6.7 Initial Segmentation

In rare cases the Unified Segmentation can fail in highly abnormal brains, where e.g. the cerebrospinal fluid of super large ventricles (hydrocephalus) were classified as white matter. However, if the affine registration is correct, the AMAP segmentation with an prior-independent k-means initialisation can be used to replace the SPM brain tissue classification

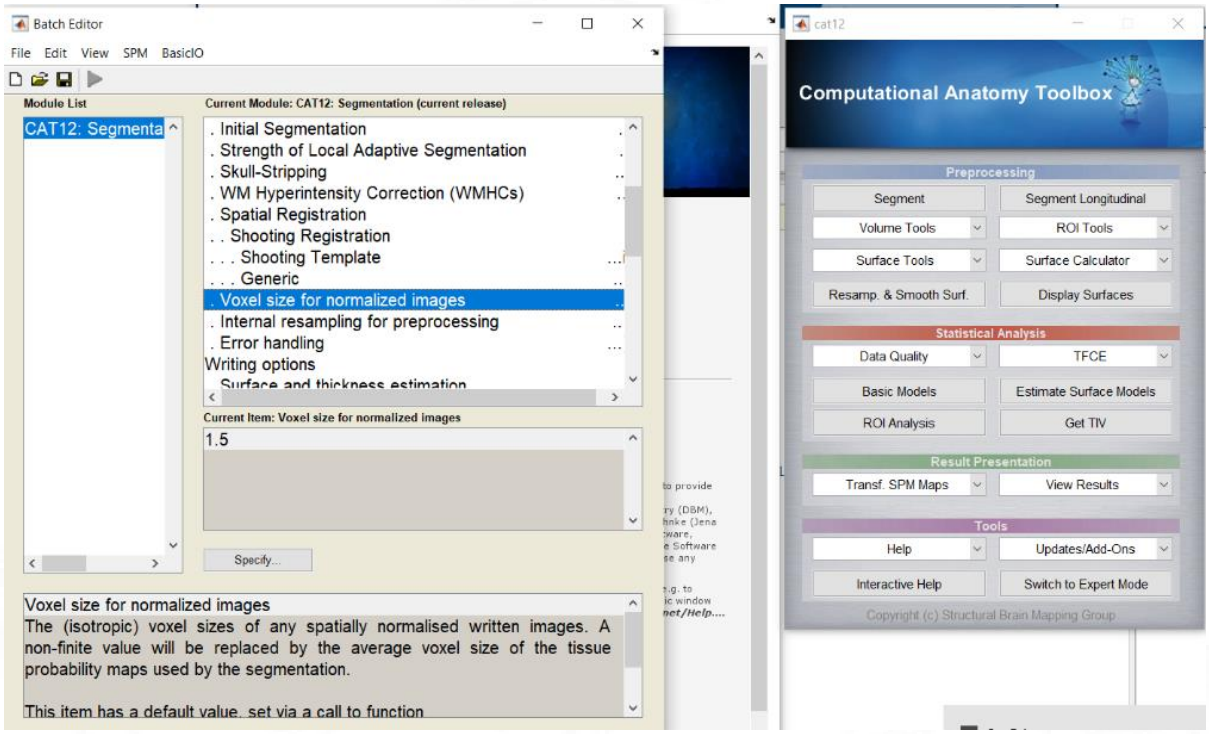


Fig 18: Voxel size for normalised images

1.6.8Voxel size for normalised images

The isotropic voxel sizes of any spatially normalised written images. A non finite value will be replaced by the average voxel size of the tissue probability maps used by the segmentation. This item has a default value set via a call to function.

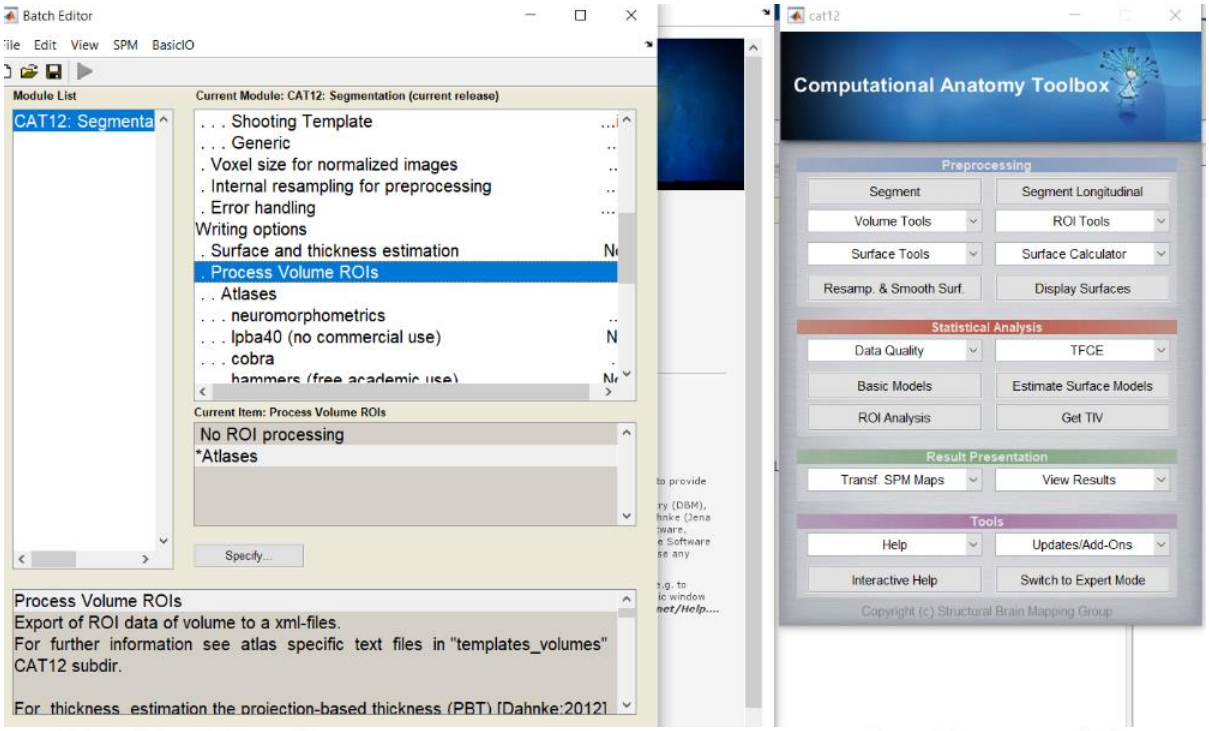


Fig 19 Atlases

1.6.9 Atlases

Writing options of Arrow Atlas maps this branch contains 5 items:

1. neuromorphic metrics
2. ipba 40
3. cobra
4. hammers
5. own Atlas maps

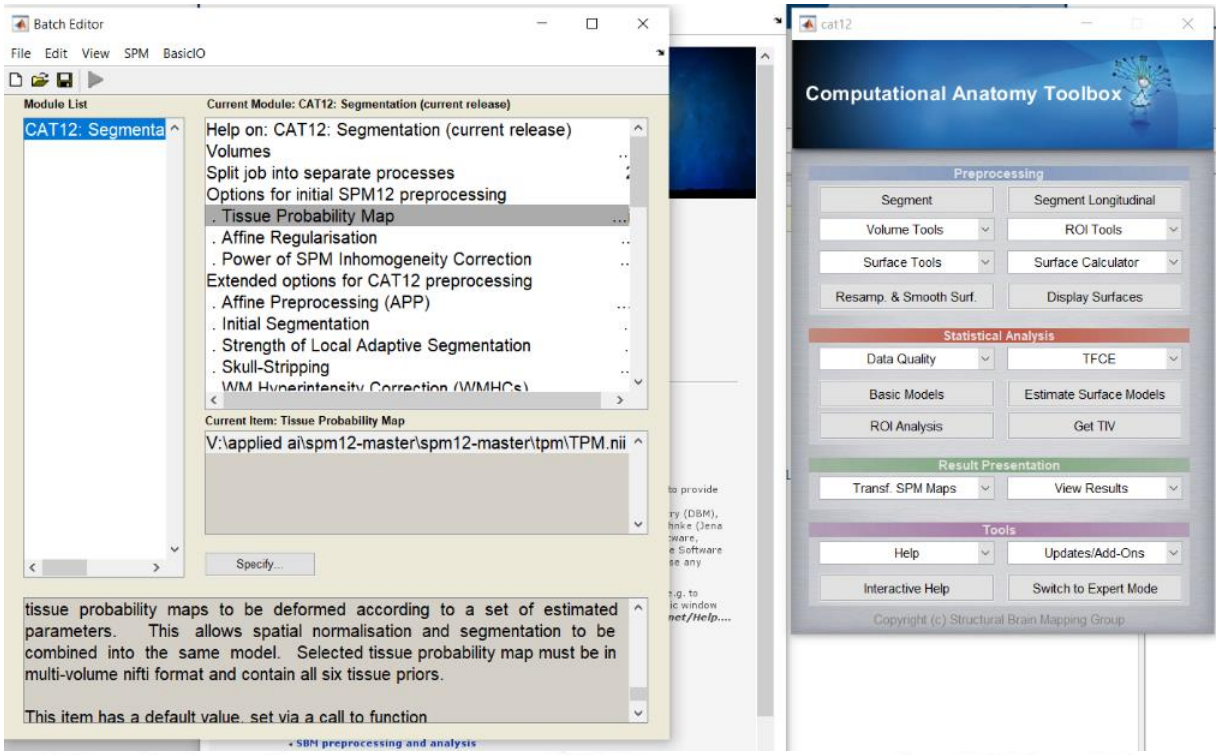


Fig 20 Tissue probability map

1.6.10 Tissue probability map

Select the tissue probability image that includes 6tissues probability classes for (1) grey matter , (2) white matter, (3) cerebrospinal fluid ,(4) bone ,(5) non brain soft tissue , and (6) the background. CAT uses the TPM only for initial SPM segmentation . Hence it is more independent and allows accurate and robust processing even with the standard TPM in case of strong anatomical differences, for example young brains. Nevertheless , for children data we recommend to use customized TPMs created using the Template-O-Matic toolbox.

1.6.11 Feature Extraction

Once all images had been spatially normalized, we used a region of interest (ROI) approach and computed the average Gray matter volume and CSF volume for 142 ROIs using SPM12 and CAT12 toolbox using MATLAB software. These 142 values, labeled ROI_1 through ROI_142, were our primary features.

1.7 Additional Features

We also took two extra features that can increase the accuracy of our model those are:

1) Age – Age is one the most important feature that can be included to categorise the AD group of patients. As Age is the biggest criterion to have this disease.

2) Sex – Sex is also one the important factor and we will see that it is also a important features for our dataset.

1.8 Column Descriptors

Neuromorphometrics (142 ROIs)		
#	ROIabbr	ROIname
1	l3thVen	Left Third Ventricle
2	r3thVen	Right Third Ventricle
3	l4thVen	Left Fourth Ventricle
4	r4thVen	Right Fourth Ventricle
5	lAcc	Left Accumbens
6	rAcc	Right Accumbens
7	lAmy	Left Amygdala
8	rAmy	Right Amygdala
9	lBst	Left Brainstem
10	rBst	Right Brainstem
11	lCau	Left Caudate
12	rCau	Right Caudate
13	lExtCbe	Left Exterior Cerebellum
14	rExtCbe	Right Exterior Cerebellum
15	lCbeWM	Left Cerebellum White Matter
16	rCbeWM	Right Cerebellum White Matter
17	lCbrWM	Left Cerebral White Matter
18	rCbrWM	Right Cerebral White Matter
19	lCSF	Left CSF
20	rCSF	Right CSF
21	lHip	Left Hippocampus
22	rHip	Right Hippocampus
23	lInfLatVen	Left Inferior Lateral Ventricle
24	rInfLatVen	Right Inferior Lateral Ventricle
25	lLatVen	Left Lateral Ventricle
26	rLatVen	Right Lateral Ventricle
27	lPal	Left Pallidum
28	rPal	Right Pallidum
29	lPut	Left Putamen
30	rPut	Right Putamen
31	lThaPro	Left Thalamus Proper
32	rThaPro	Right Thalamus Proper

33	lVenVen	Left Ventral Ventricle
34	rVenVen	Right Ventral Ventricle
35	lOC	Left Optic Chiasm
36	rOC	Right Optic Chiasm
37	lCbeLoCbe1-5	Left Cerebellar Lobule Cerebellar Vermal Lobules I-V
38	rCbeLoCbe1-5	Right Cerebellar Lobule Cerebellar Vermal Lobules I-V
39	lCbeLoCbe6-7	Left Cerebellar Lobule Cerebellar Vermal Lobules VI-VII
40	rCbeLoCbe6-7	Right Cerebellar Lobule Cerebellar Vermal Lobules VI-VII
41	lCbeLoCbe8-10	Left Cerebellar Lobule Cerebellar Vermal Lobules VIII-X
42	rCbeLoCbe8-10	Right Cerebellar Lobule Cerebellar Vermal Lobules VIII-X
43	lBasCbr+FobBr	Left Basal Cerebrum and Forebrain Brain
44	rBasCbr+FobBr	Right Basal Cerebrum and Forebrain Brain
45	lAntCinGy	Left Anterior Cingulate Gyrus
46	rAntCinGy	Right Anterior Cingulate Gyrus
47	lAntIns	Left Anterior Insula
48	rAntIns	Right Anterior Insula
49	lAntOrbGy	Left Anterior Orbital Gyrus
50	rAntOrbGy	Right Anterior Orbital Gyrus
51	lAngGy	Left Angular Gyrus
52	rAngGy	Right Angular Gyrus
53	lCal+Cbr	Left Calcarine and Cerebrum
54	rCal+Cbr	Right Calcarine and Cerebrum
55	lCenOpe	Left Central Operculum
56	rCenOpe	Right Central Operculum
57	lCun	Left Cuneus
58	rCun	Right Cuneus
59	lEnt	Left Entorhinal Area
60	rEnt	Right Entorhinal Area
61	lFroOpe	Left Frontal Operculum
62	rFroOpe	Right Frontal Operculum

63	lFroPo	Left Frontal Pole
64	rFroPo	Right Frontal Pole
65	lFusGy	Left Fusiform Gyrus
66	rFusGy	Right Fusiform Gyrus
67	lRecGy	Left Gyrus Rectus
68	rRecGy	Right Gyrus Rectus
69	lInfOccGy	Left Inferior Occipital Gyrus
70	rInfOccGy	Right Inferior Occipital Gyrus
71	lInfTemGy	Left Inferior Temporal Gyrus
72	rInfTemGy	Right Inferior Temporal Gyrus
73	lLinGy	Left Lingual Gyrus
74	rLinGy	Right Lingual Gyrus
75	lLatOrbGy	Left Lateral Orbital Gyrus
76	rLatOrbGy	Right Lateral Orbital Gyrus
77	lMidCinGy	Left Middle Cingulate Gyrus
78	rMidCinGy	Right Middle Cingulate Gyrus
79	lMedFroCbr	Left Medial Frontal Cerebrum
80	rMedFroCbr	Right Medial Frontal Cerebrum
81	lMidFroGy	Left Middle Frontal Gyrus
82	rMidFroGy	Right Middle Frontal Gyrus
83	lMidOccGy	Left Middle Occipital Gyrus
84	rMidOccGy	Right Middle Occipital Gyrus
85	lMedOrbGy	Left Medial Orbital Gyrus
86	rMedOrbGy	Right Medial Orbital Gyrus
87	lMedPoCGy	Left Medial Postcentral Gyrus
88	rMedPoCGy	Right Medial Postcentral Gyrus
89	lMedPrcGy	Left Medial Precentral Gyrus
90	rMedPrcGy	Right Medial Precentral Gyrus
91	lSupMedFroGy	Left Superior Medial Frontal Gyrus
92	rSupMedFroGy	Right Superior Medial Frontal Gyrus
93	lMidTemGy	Left Middle Temporal Gyrus
94	rMidTemGy	Right Middle Temporal Gyrus
95	lOccPo	Left Occipital Pole
96	rOccPo	Right Occipital Pole

97	lOccFusGy	Left Occipital Fusiform Gyrus
98	rOccFusGy	Right Occipital Fusiform Gyrus
99	lInfFroGy	Left Inferior Frontal Gyrus
100	rInfFroGy	Right Inferior Frontal Gyrus
101	lInfFroOrbGy	Left Inferior Frontal Orbital Gyrus
102	rInfFroOrbGy	Right Inferior Frontal Orbital Gyrus
103	lPosCinGy	Left Posterior Cingulate Gyrus
104	rPosCinGy	Right Posterior Cingulate Gyrus
105	lPCu	Left Precuneus
106	rPCu	Right Precuneus
107	lParHipGy	Left Parahippocampus Gyrus
108	rParHipGy	Right Parahippocampus Gyrus
109	lPosIns	Left Posterior Insula
110	rPosIns	Right Posterior Insula
111	lParOpe	Left Parietal Operculum
112	rParOpe	Right Parietal Operculum
113	lPoCGy	Left Postcentral Gyrus
114	rPoCGy	Right Postcentral Gyrus
115	lPosOrbGy	Left Posterior Orbital Gyrus
116	rPosOrbGy	Right Posterior Orbital Gyrus
117	lPla	Left Planum Polare
118	rPla	Right Planum Polare
119	lPrcGy	Left Precentral Gyrus
120	rPrcGy	Right Precentral Gyrus
121	lTem	Left Temporal
122	rTem	Right Temporal
123	lSCA	Left Subcallosal Area
124	rSCA	Right Subcallosal Area
125	lSupFroGy	Left Superior Frontal Gyrus
126	rSupFroGy	Right Superior Frontal Gyrus
127	lCbr+Mot	Left Cerebrum and Motor
128	rCbr+Mot	Right Cerebrum and Motor
129	lSupMarGy	Left Supramarginal Gyrus
130	rSupMarGy	Right Supramarginal Gyrus

131	lSupOccGy	Left Superior Occipital Gyrus
132	rSupOccGy	Right Superior Occipital Gyrus
133	lSupParLo	Left Superior Parietal Lobule
134	rSupParLo	Right Superior Parietal Lobule
135	lSupTemGy	Left Superior Temporal Gyrus
136	rSupTemGy	Right Superior Temporal Gyrus
137	lTemPo	Left Temporal Pole
138	rTemPo	Right Temporal Pole
139	lInfFroAngGy	Left Inferior Frontal Angular Gyrus
140	rInfFroAngGy	Right Inferior Frontal Angular Gyrus
141	lTemTraGy	Left Temporal Transverse Gyrus
142	rTemTraGy	Right Temporal Transverse Gyrus

Table 1Features

1.9 Steps performed after data extraction:

After getting dataset from the MATLAB file , we copied all those data from MATLAB file to excel and then we converted those excel file to csv file to perform different operation on that dataset.

After getting the dataset in csv format we started performing on Python

The steps are as following:

1.10 EXPLORATORY DATA ANALYSIS (EDA)

It is best to first analyse the data and then strive to glean as many insights as possible from it. Before getting their hands dirty with data, EDA is all about making sense of it.

In this part, we investigated the association between each aspect of MRI testing and the patient's dementia.We performed this Exploratory Data Analysis technique to clearly explain the association of data through a graph so that we may assume the correlations prior to data extraction or data analysis. It might assist us understand the nature of the data and afterwards choose the best analysis strategy for the model.

1.10.1 Data Pre-processing

This section focuses on data pre-processing techniques in Python. Learning algorithms have a preference for specific data types, on which they often perform very well. They are known to make incredibly rash predictions when unscaled or unstandardized characteristics are applied

to them. Algorithms like XGBoost and LGBM expressly require dummy encoded data, but decision trees don't seem to care (sometimes)!

In simple terms, pre-processing refers to the modifications performed on your data prior to feeding it to the algorithm. The scikit-learn package in Python has a pre-built capability called sklearn pre-processing. There are several more possibilities for pre-processing.

Rows with missing values are being removed.

Train/Validation/Test Sets Splitting.

1.10.2 Cross-validation

To determine the appropriate parameters for each model, Logistic Regression, SVM, Decision Tree, Random Forests, and AdaBoost, we use 5-fold cross-validation. Because accuracy is our performance metric, we identify the ideal tuning settings based on accuracy. Finally, we compare each model's accuracy, recall, and AUC.

1.10.3 Feature Scaling:

The approach of limiting the range of variables so that they may be compared on common grounds is known as feature scaling. It is used to continuous variables.

1.10.4Label Encoding

Sklearn is a powerful tool for encoding the levels of category characteristics into numerical values. Label Encoder encodes labels with values ranging from 0 to n classes.

All of our category characteristics are encoded. Using X train.head, you may inspect your updated data set. We'll look at the gender frequency distribution before and after encoding.

1.10.5Performance Measures

Our primary performance metric is the area under the receiver operating characteristic curve (AUC). We feel that in the case of medical diagnostics for non-life threatening terminal diseases, such as most neurodegenerative illnesses, a high true positive rate is critical so that all individuals with Alzheimer's are diagnosed as soon as feasible. However, we must ensure that the false positive rate is as low as feasible, since we do not want to misdiagnose a healthy adult as demented and initiate medical treatment. As a result, AUC appeared to be an excellent candidate for a performance metric.

What are the terms Sensitivity and Specificity?

A confusion matrix looks like this:

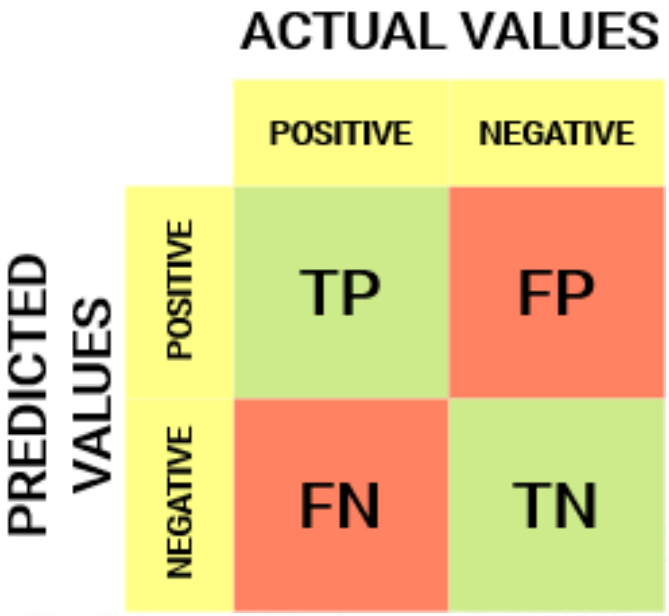


Fig 21 Confusion matrix

Sensitivity / Recall /True Positive Rate

$$Sensitivity = \frac{TP}{TP + FN}$$

Sensitivity informs us how many people in the positive class were accurately categorised.

FNR

$$FNR = \frac{FN}{TP + FN}$$

The False Negative Rate indicates how much of the positive class was erroneously categorised by the classifier.

We want to accurately categorise the positive class, hence we want a higher TPR and a lower FNR.

True Negative Rate/ Specificity

Specificity shows us how much of the negative class was accurately categorised.

False Positive Rate

The FPR shows us how much of the negative class was wrongly categorised by the classifier.

We want to accurately categorise the negative class, hence we want a greater TNR and a lower FPR. The most crucial are probably sensitivity and specificity.

What is the AUC-ROC curve?

The Receiver Operator Characteristic (ROC) curve is a binary classification issue assessment measure. It is a probability curve that displays the TPR against the FPR at various threshold

levels, separating the'signal' from the 'noise.' The Area Under the Curve (AUC) is a measure of a classifier's ability to discriminate between classes and is used to summarise the ROC curve. The greater the AUC, the better the model's ability in differentiating between positive and negative classifications.

AUC-ROC for Multi-Class Classification

As previously stated, the AUC-ROC curve is only applicable to binary classification situations. However, by employing the One vs All methodology, we may extend it to multiclass classification situations. So, if we have three classes, 0, 1, and 2, the ROC for class 0 will be created as a result of categorising 0 versus not 0, i.e. 1 and 2. The ROC for class 1 will be calculated by comparing class 1 to not 1, and so on.

1.10.6Different Machine Learning Algorithm Applied :

The goal of multinomial model was to distinguish among all three disease categories (Normal, MCI, and AD). We implemented different models using scikit-learn Python Machine Learning Library. In the case of the multinomial model, we used a one-vs-all scheme to train the data.

1.10.7Decision tree

A decision tree is a supervised learning technique (one with a predetermined goal variable) that is commonly employed in classification issues. It is applicable to both categorical and continuous input and output variables. We divide the population or sample into two or more homogenous groups (or sub-populations) using the most significant splitter, differentiator in input variables.

1.10.8RANDOM FOREST

Random forests are type of ensemble methods in machine learning. Random forest is most successful ensemble methods which sometimes had exhibited performance better than boosting and support vector machines. It is also proved to be effective classifier for classification of big data in biomedical, biotechnology, and medical imaging fields. Many competitions in kaggle had been won by using Random Forest classifier. When other classifier is hard to interpret, Random forest offers more intuitive illustrations . The random sampling and ensemble procedures used in RF allow it to make more accurate predictions and generalisations. This generalisation trait is derived from the bagging scheme, which enhances generalisation by reducing variance, whereas related approaches like as boosting do this via reducing bias. Randomization for increasing diversity is proved to be efficient method of classification.

1.10.10 Gradient Boosting Algorithms

GBM is a boosting method that is generally utilised when dealing with a large amount of dataset in order to make a prediction with a very high prediction power and accuracy. Boosting is the collection of learning techniques that generally add the predictions of numerous base estimators

to increase resilience over a single estimator. It combines a number of weak or mediocre predictors to create a powerful predictor.

1.10.11 XGBoost

The XGBoost have a very high predictive and probabilistic power, making it ideal choice of event accuracy since it generally contains both of the model, a linear model and a tree learning method, making the method about ten times quicker than the existing gradient boosting approaches.

1.10.12 Light GBM:

LightGBM is thevery highperforming gradient boosting techniques based on decision tree techniques that can be used for classification techniques, ranking them and a variety of other machine learning applications.LightGBM is a framework for GB that use tree-based learning techniques. It is intended to be widely dispersed and be efficient, with the some of the following benefits:training speed becomes more faster and have very higher efficiency, Lower memory usage, and very good accuracy.Support of parallel and GPU learning, and is very much capable of handling a very large data.

1.11 Feature Selection

It is an interaction of choosing the most critical and applicable highlights from a bunch of highlights in the given dataset.

For a dataset with d information includes, the element determination measure brings about k highlights to such an extent that k < d, where k is the littlest arrangement of huge and significant highlights.

1.11.1 Chi-square

It is utilized for the all-out highlights in the dataset. We regularly compute Chi-square between each component and the objective and select the ideal number of highlights with the best Chi-square scores. To effectively apply the chi-squared to test the connection between different highlights in the dataset and the objective variable, the accompanying conditions must be met: the factors must be clear cut, examined autonomously and qualities ought to have a normal recurrence more noteworthy than 5.

1.11.2 Recursive Feature Elimination (RFE)

RFE is a feature selection method in which in it generally select the features recursively in a small-small set of features. First, the estimator is generally trained on the initial set of features and the importance of each feature is obtained either through a coef_ attribute or through a feature_importances_ attribute.Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.We used this model to find the small set of features

from all the features that can perform in a better wayfor each of our models. With the lesser number of features , we can see some a little bit of improvement in all those different models we have used.

2. IMPORTANT STEPS

Important steps are used as following:

1. Import all the important libraries into the python file.
2. Then read the doc that consists of our dataset named as “Vishal_alz_detect.csv”
3. Then we found the shape of the dataset that is 823,289
4. Clean the column name by removing special character or space or symbols.
5. Removing the missing rows and duplicated rows.
6. Now we checked the file it looks like after removing the duplicate value.
7. After removing the duplicate value, we checked the shape of the dataset again and then we came to know that the dimension of the dataset is (289,287) that means we are having 289 patients with 287 different features.
8. Then we normally saw the statistics related to the features of dataset like we calculated the mean value of each feature.
9. We saw the information related to all the dataset like we checked the datatypes of all the features.
10. We checked the distribution of patients then we concluded that 159 belongs to MCI , 93 patients belong to CN and 37 patients belongs to AD.
11. Then we did label encoding , as we can’t use machine learning techniques on categorical variables that’s why we coded female as 0 and male as 1.
12. And we coded patient group also into coded form like CN as 0 , MCI as 1 and AD as 2.
13. Then we used seaborn library to show in graph how the patient distribution in the dataset.
14. Installed Light GBM ML algorithms which we will be using in future for patient condition prediction.
15. We removed features like patient Id and Image Id from the dataset as these features are of no use.
16. We distributed whole dataset into two parts in training dataset and test dataset in 67:33 proportion.
17. We used Decision Tree Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 86.45% accuracy on test dataset and 99.48% on training dataset.

18. We used Random Forest Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 91.66% accuracy on test dataset and 99.48% on training dataset.
19. We used Gradient boosting Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 89.58% accuracy on test dataset and 99.48% on training dataset.
20. We used Logistic regression Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 89.58% accuracy on test dataset and 99.48% on training dataset.
21. We plotted ROC-AUC curve for multi classification using logistic regression (One vs All) ML algorithms.
22. We used XG boosting Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 89.58% accuracy on test dataset and 99.48% on training dataset.
23. We used Light GBM Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 94.55 accuracy on test dataset and 99.48% on training dataset.
24. We used chi square to select top 5 features that can be used to predict the patient condition in a better way than using all features at a time as taking all features at a time is time consuming.
25. We used RFE to select top 5 features that can be used to predict the patient condition in a better way than using all features at a time as taking all features at a time is time consuming.
26. We plotted ROC-AUC curve for multi classification using logistic regression (One vs All) ML algorithms.
27. We plotted a chart of number of features selected vs cross validation scoreof selected features.
28. Installed one of the most important feature selection library developed by Microsoft.
29. Imported library SHAP and then we plotted the bar graph that is saying us which of the given features are important to classify the patient into different category. As we can see the feature CSF roccFusGy will help a lot in classifying a patient into AD group more than any features whatever we took.
30. Plotted the important features to correctly classify AD Group.
31. Plotted the important features to correctly classify CN Group.
32. Plotted the important features to correctly classify MCI Group.

CHAPTER 4

RESULTS

1. INTRODUCTION

This chapter summarises the results of the .analysis described in the previous chapters All our experiment were run with a 67% / 33% proportion of dataset (Training / testing) as described earlier , We also used Stratified k-fold cross-validation methods to calculate the accuracy of different models. We used this classification as our main goal:Normal / MCI / AD

2. EDA

2.1 Statistics of dataset

	Age	GM_l3thVen_	GM_r3thVen_	GM_l4thVen_	GM_r4thVen_	GM_lAcc_	GM_rAcc_	GM_lAmy_	GM_rAmy_	GM_lBst_
count	289.000000	289.000000	289.000000	289.000000	289.000000	289.000000	289.000000	289.000000	289.000000	289.000000
mean	76.273356	0.042531	0.033522	0.066225	0.076509	0.331842	0.333170	0.710713	0.681481	0.587371
std	7.121808	0.017928	0.016750	0.020472	0.020816	0.060658	0.060580	0.158534	0.165030	0.209235
min	55.000000	0.013582	0.008788	0.029157	0.038525	0.135233	0.161498	0.353952	0.314563	0.213282
25%	72.000000	0.028070	0.021043	0.052625	0.062000	0.284919	0.304557	0.629464	0.551302	0.442588
50%	76.000000	0.040864	0.030645	0.060694	0.073769	0.327820	0.333290	0.723004	0.719753	0.558760
75%	81.000000	0.048980	0.042384	0.080080	0.086486	0.373129	0.361917	0.842578	0.813988	0.667545
max	90.000000	0.095719	0.079800	0.142294	0.144082	0.535776	0.487392	1.111668	1.063428	1.164329

8 rows × 285 columns

Fig 22 Statistics of dataset

1. After removing the duplicate value, we checked the shape of the dataset again and then we came to know that the dimension of the dataset is (289,289) that means we are having 289 patients with 289 different features.
2. Then we normally saw the statistics related to the features of dataset like we calculated the mean value of each feature.
3. We checked the distribution of patients then we concluded that 159 belongs to MCI , 93 patients belong to CN and 37 patients belongs to AD.

```
Number of Normal people:  93
Number of MCI patient :  159
Number of AD patient :  37
Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and pass
```

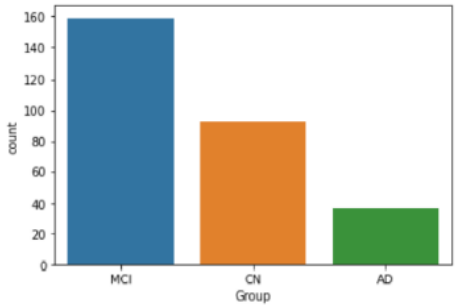


Fig 23 Distribution of Patients

1. We checked the distribution of patients then we concluded that 159 belongs to MCI , 93 patients belong to CN and 37 patients belongs to AD.
2. Then we used seaborn library to show in graph how the patient distribution in the dataset.

2.2 Machine Learning Algorithms:

2.2.1 Decision Tree

1. We used Decision Tree Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 86.45% accuracy on test dataset and 99.48% on training dataset.
2. With Stratified k-fold cross-validation we got accuracy of 91.69%

2.2.1.1 Confusion Matrix for Decision Tree:

	precision	recall	f1-score	support
AD	0.77	0.83	0.80	12
CN	0.84	0.87	0.86	31
MCI	0.92	0.89	0.90	53
accuracy			0.88	96
macro avg	0.84	0.86	0.85	96
weighted avg	0.88	0.88	0.88	96

Fig 24: Confusion Matrix

2.2.1.2 ROC-AUC Curve

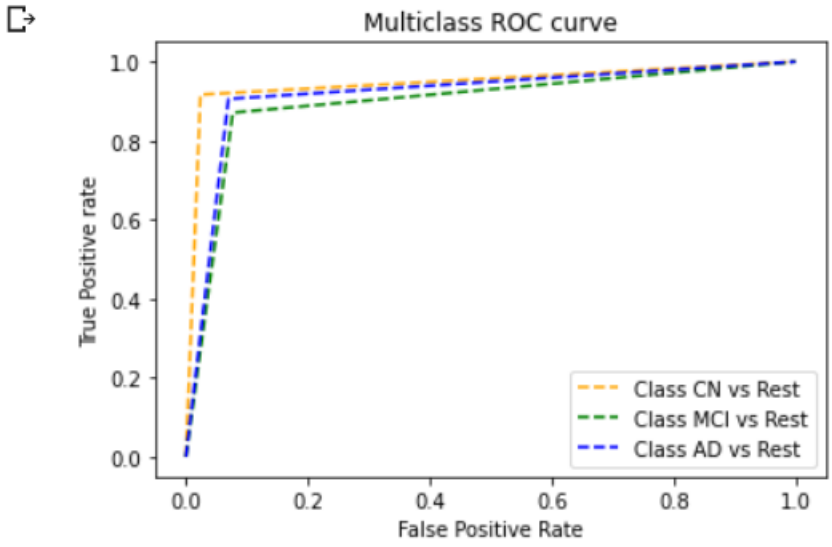


Fig 25: ROC-AUC Curve

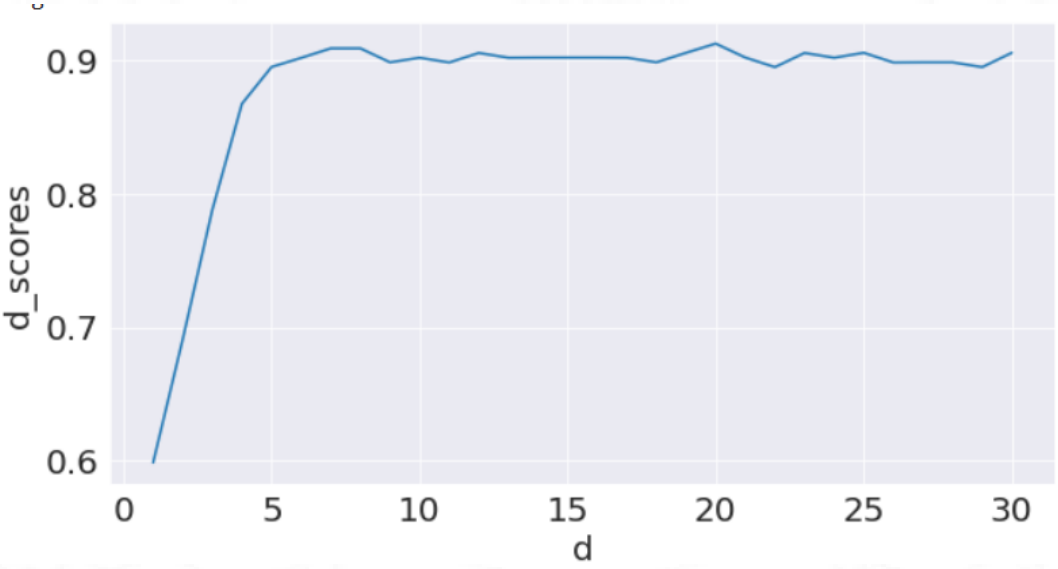


Fig 26: Plot of depth vs score of decision tree

1. Here, d indicates the depth of decision tree and d_scores indicates the accuracy of decision tree

2.2.2 Random forest:

1. We used Random Forest Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 91.66% accuracy on test dataset and 99.48% on training dataset.
2. With Stratified k-fold cross-validation we got accuracy of 92.38%

2.2.2.1 ROC-AUC Curve

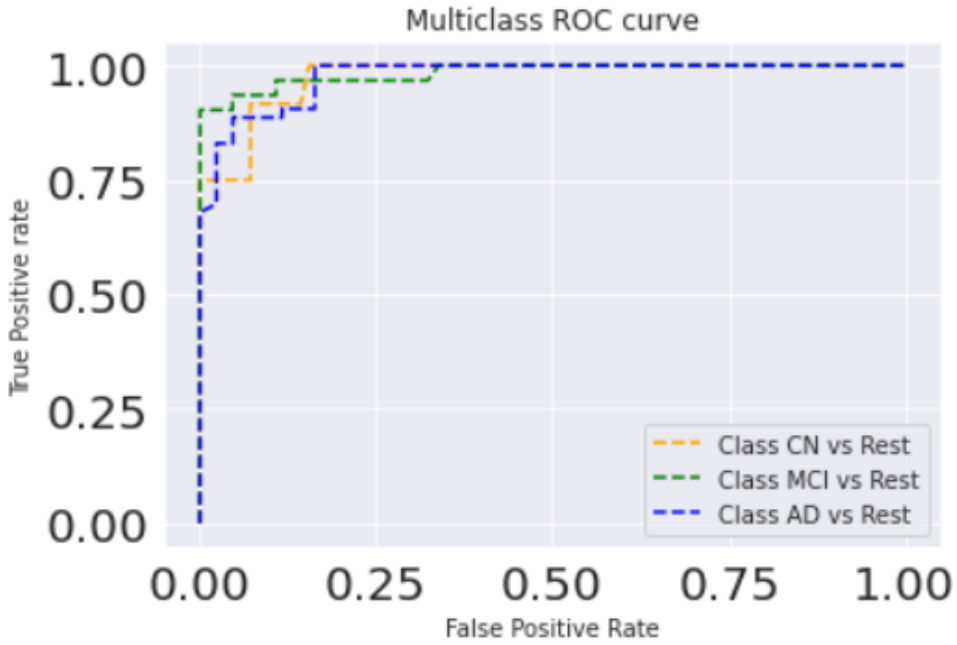


Fig 27 ROC-AUC Curve

2.2.2.2 Confusion matrix

	precision	recall	f1-score	support
AD	1.00	0.75	0.86	12
CN	0.88	0.94	0.91	31
MCI	0.91	0.92	0.92	53
accuracy			0.91	96
macro avg	0.93	0.87	0.89	96
weighted avg	0.91	0.91	0.91	96

Fig 28 Confusion matrix

2.2.3 Gradient Boosting Classifier:

1. We used Gradient boosting Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 89.58% accuracy on test dataset and 99.48% on training dataset.
2. With Stratified k-fold cross-validation we got accuracy of 91.35%

2.2.3.1 ROC-AUC Curve

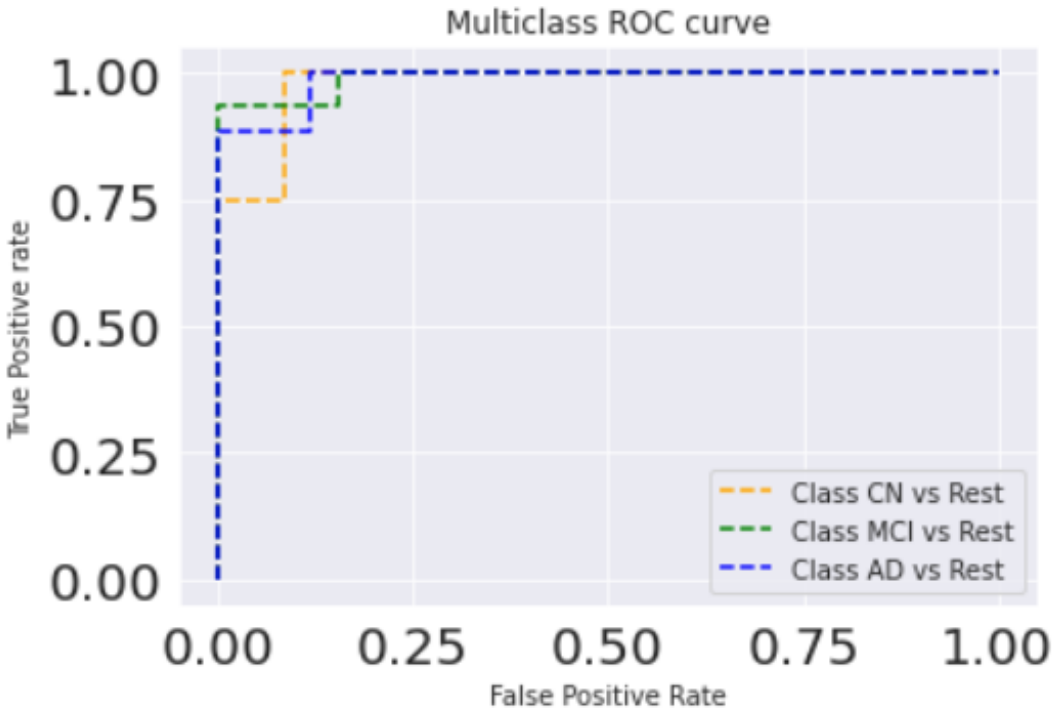


Fig: 29:AUC-ROC Curve

2.2.4 Logistic Regression

1. We used Logistic regression Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 88.54% accuracy on test dataset and 98.44% on training dataset.
2. With Stratified k-fold cross-validation we got accuracy of 94.11%

2.2.4.1 ROC-AUC Curve:

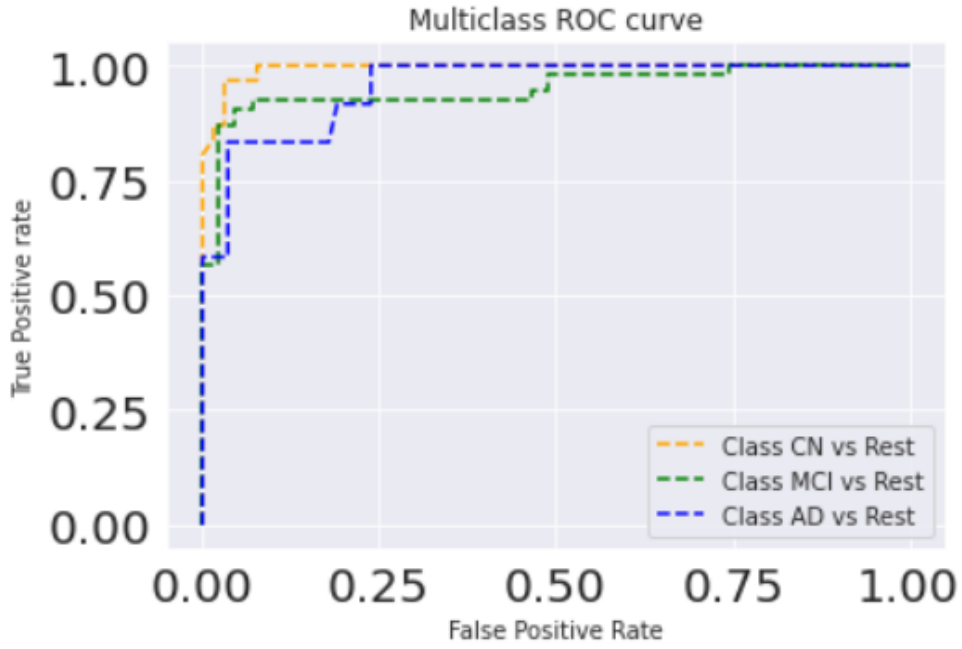


Fig 30: ROC-AUC Curve

1. We plotted ROC-AUC curve for multi classification using logistic regression (One vs All) ML algorithms.

2.2.4.2 Confusion Matrix

	precision	recall	f1-score	support
AD	1.00	0.67	0.80	12
CN	0.93	0.90	0.92	31
MCI	0.90	0.98	0.94	53
accuracy			0.92	96
macro avg	0.94	0.85	0.88	96
weighted avg	0.92	0.92	0.91	96

Fig 31: Confusion Matrix

2.2.5 XG Boosting Classifier:

1. We used XG boosting Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 89.58% accuracy on test dataset and 99.48% on training dataset.
2. With Stratified k-fold cross-validation we got accuracy of 91.69%

2.2.5.1 Confusion Matix:

	precision	recall	f1-score	support
AD	0.82	0.75	0.78	12
CN	0.97	0.94	0.95	31
MCI	0.91	0.94	0.93	53
accuracy			0.92	96
macro avg	0.90	0.88	0.89	96
weighted avg	0.92	0.92	0.92	96

Fig 32: Confusion Martix

2.2.5.2 ROC-AUC Curve:

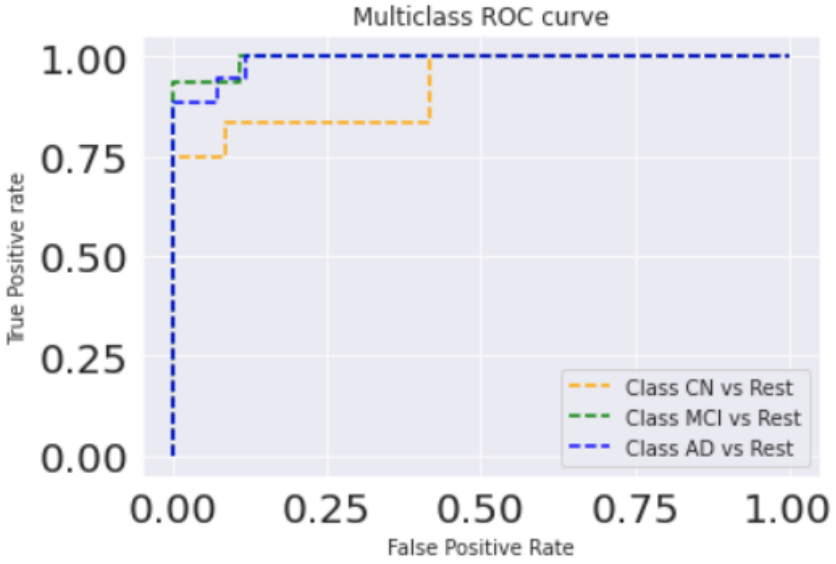


Fig 33: ROC-AUC Curve

2.2.6 Light GBM

2.2.6.1 Confusion Matrix

	precision	recall	f1-score	support
AD	0.90	0.75	0.82	12
CN	1.00	0.94	0.97	31
MCI	0.91	0.98	0.95	53
accuracy			0.94	96
macro avg	0.94	0.89	0.91	96
weighted avg	0.94	0.94	0.94	96

```
from sklearn.metrics import precision_score,f1_score
sc1=np.round(precision_score(y_test,pred,average='macro'),4)
sc2=np.round(precision_score(y_test,pred,average='micro'),4)
f1=np.round(f1_score(y_test,pred,average='weighted'),4)
print(f'Weighted F1-Score: {np.round(np.mean(f1),4)}, Macro precision score {np.round(np.mean(sc1),4)}, Micro precision score {np.rou
```

```
Weighted F1-Score: 0.9364, Macro precision score 0.9374, Micro precision score 0.9375
```

Fig 34:Confusion Matrix

1. We used Light GBM Machine learning technique to predict the patient dementia group and then calculate the accuracy, we got 94.55 accuracy on test dataset and 99.48% on training dataset.
2. With Stratified k-fold cross-validation we got accuracy of 93.42%

2.2.6.2 ROC-AUC Curve:

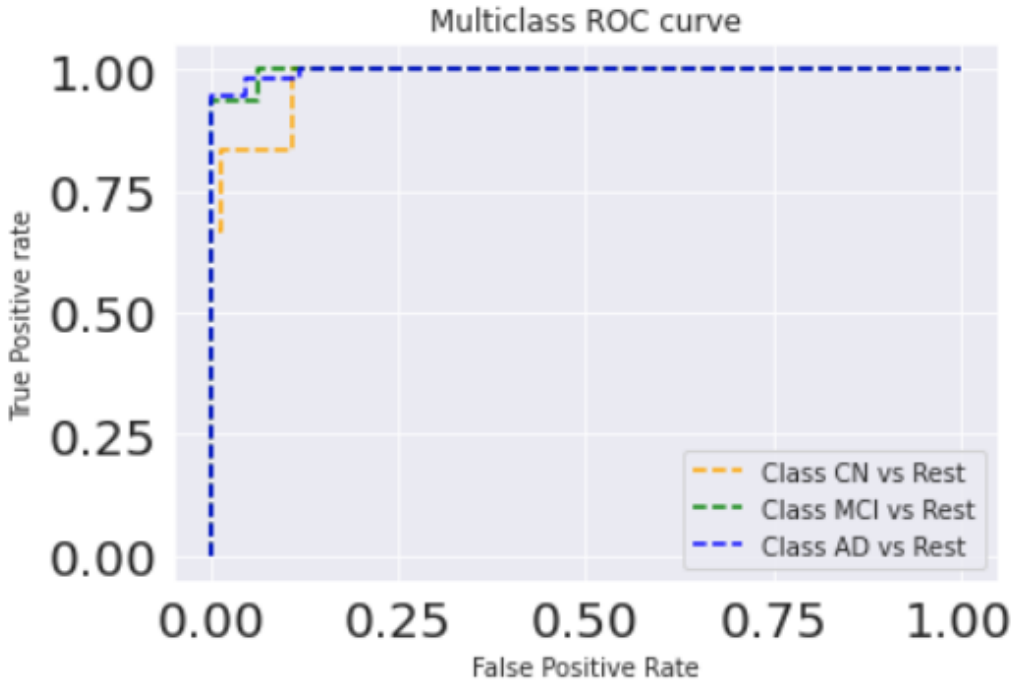


Fig 35 ROC-AUC Curve

2.2.7 Comparison of accuracy :

Machine Learning Algorithms	Accuracy on Testing Dataset on Whole Features	Accuracy with Stratified k-fold
Decision Tree	89.55	91.69
Random Forest	91.48	92.35
Logistic Regression	89.55	94.11
XG boosting	89.55	91.69
Gradient Boosting	89.55	91.38
Light GBM	94.55	93.64

Table 2 Comparison of accuracy of different machine learning techniques

2.3 Feature Selection

2.3.1 Using chi-square algorithm

We used chi square to select top 5 features that can be used to predict the patient condition in a better way than using all features at a time as taking all features at a time is time consuming.

2.3.2 RFE

1. We used RFE to select top 5 features that can be used to predict the patient condition in a better way than using all features at a time as taking all features at a time is time consuming.
2. We got best five features as
   [GM_rExtCbe(Right Exterior Cerebellum)

GM_lFro0pe(Left Frontal Operculum)

CSF_lCau,(Left Caudate)

CSF_rCau (Right Caudate)

CSF_l0ccFusGy_ (Left Occipital Fusiform Gyrus)]

We plotted ROC-AUC curve for multi classification using logistic regression (One vs All) ML algorithms.

2.4 Cross Validation

```
# Plot number of features VS. cross-validation scores
import matplotlib.pyplot as plt
plt.figure()
plt.xlabel("Number of features selected")
plt.ylabel("Cross validation score of number of selected features")
plt.plot(range(1, len(rfecv.grid_scores_) + 1), rfecv.grid_scores_)
plt.show()
```

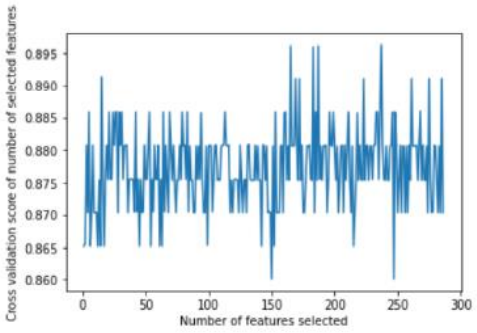


Fig 36 Plot betweennumber of features selected vs cross validation score of selected features.

We plotted a chart of number of features selected vs cross validation scoreof selected features.

2.5 Feature Importance Overall

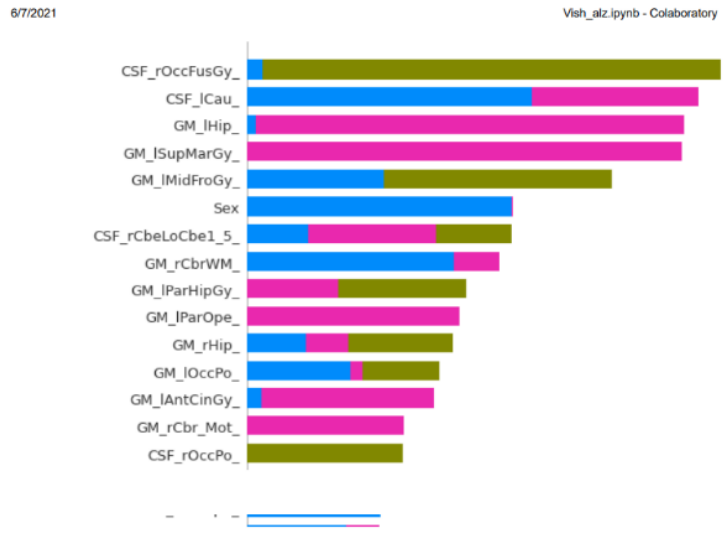


Fig 37:Feature importance overall

1. We plotted the bar graph that is saying us which of the given features are important to classify the patient into different category. As we can see the feature CSF roccFusGy will help a lot in classifying a patient into AD group more than any features whatever we took.
2. We got these 5 features as the best as to categorise the group overall. Those are:
    1. CSF of Right Occipital Fusiform Gyrus.
    2. CSF of Left Caudate

3. GM of Left Hippocampus
4. GM of Left Supramarginal gyrus
5. GM of Left Frontal Gyrus

2.5.1 Feature importance for AD

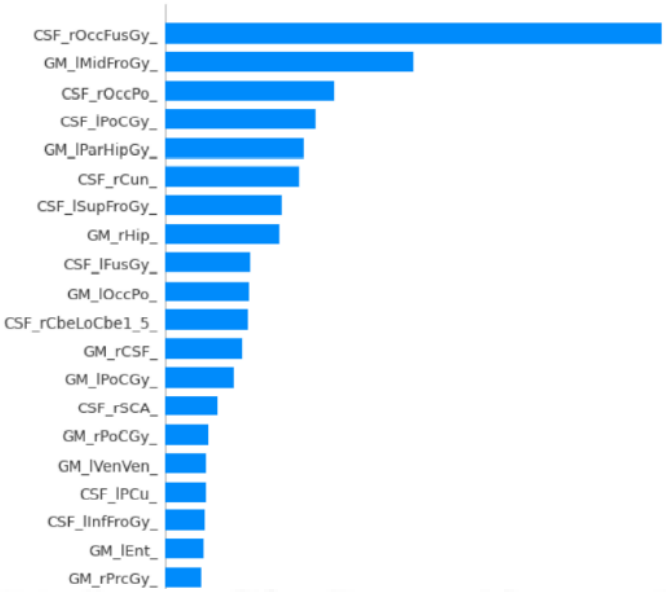


Fig 38 Code for feature importance for AD

These are the important features to correctly classify AD Group

1. CSF of right occipital fusiform gyrus
2. GM of left middle frontal gyrus
3. CSF of right occipital pole
4. CSF of left para hippocampus gyrus
5. CSF of right hippocampus

2.5.1.1 Feature CSF_r0ccFusGy importance to categorise AD

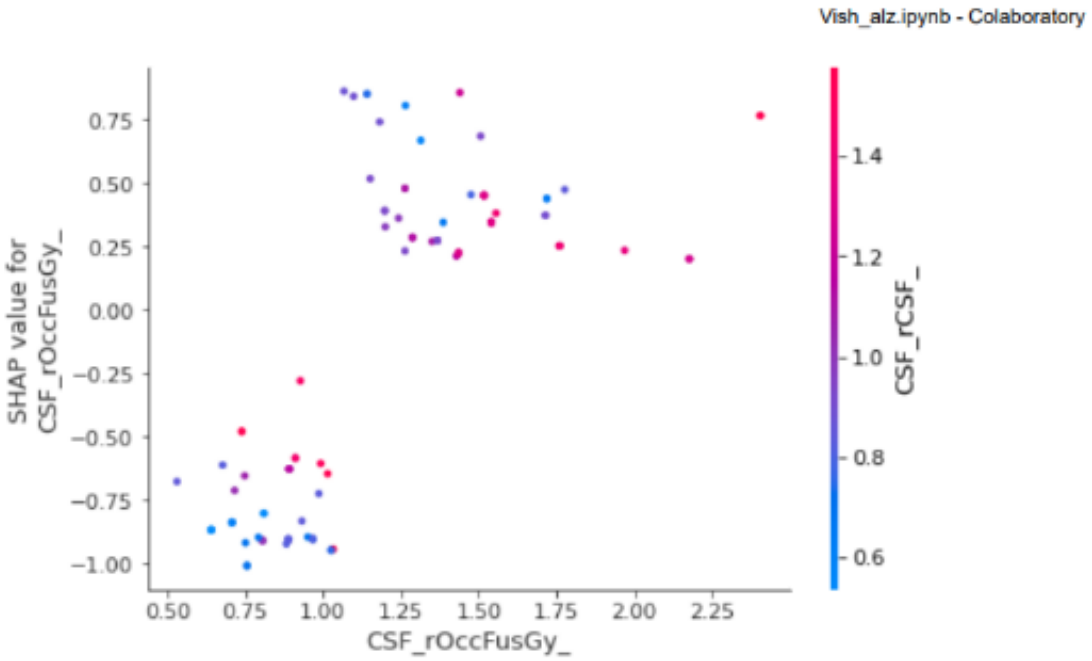


Fig 39 Plot to see "CSF rOccFusGy" importance to categorise AD

1. This Graph Signifies that patient with " CSF of Right Occipital Fusiform Gyrus." more than 1.1 are having higher chances to be in AD Group

2.5.1.2Feature “CSF rOccPo”importance to categorise AD

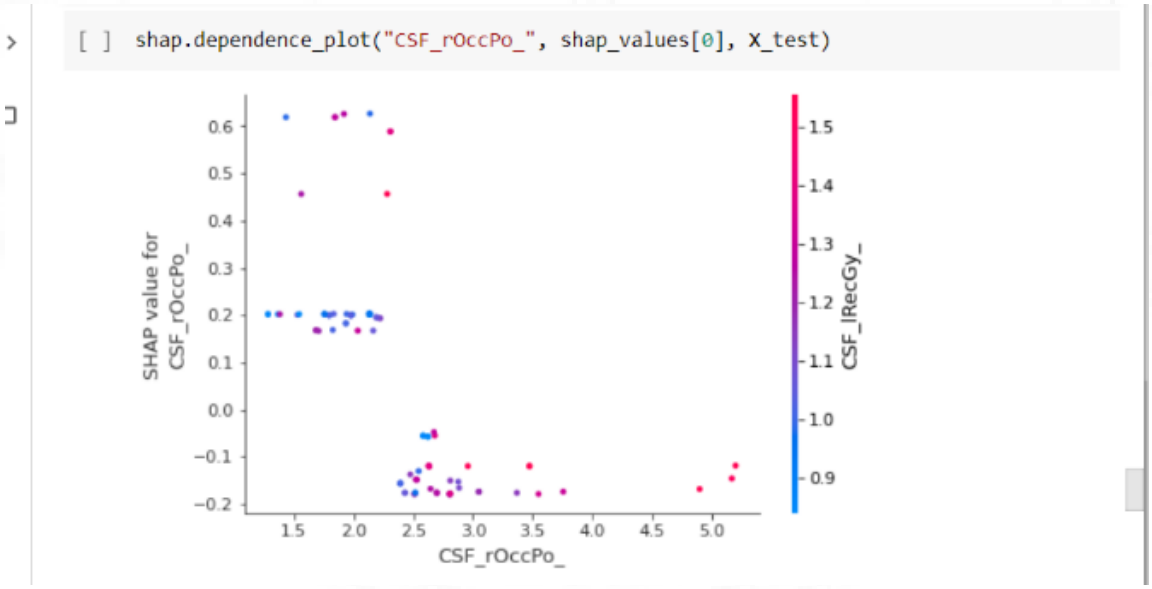


Fig 40 Plot to see "CSF rOccPo" importance to categorise AD

This Graph Signifies that patient with "CSF of right occipital pole" less than 2.2 are having higher chances to be in AD Group

2.5.1.3 Feature CSF_lMidFroGy importance to categorise AD

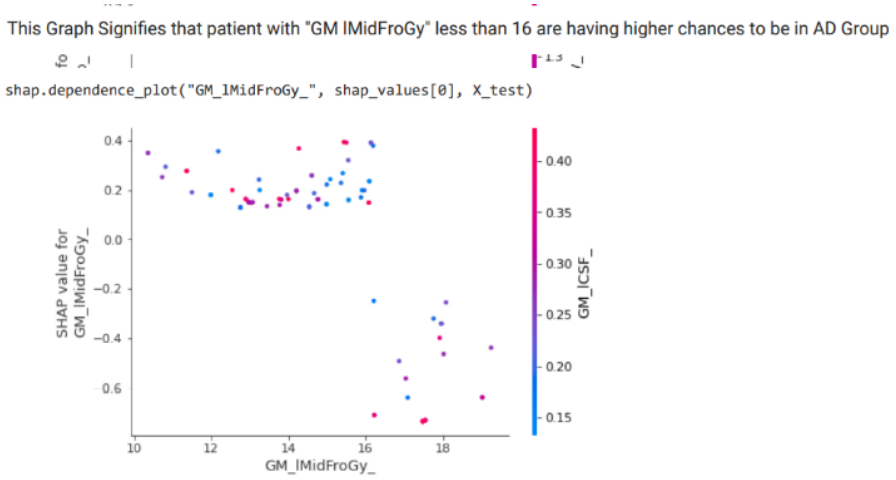


Fig 41 Plot to see CSF_lMidFroGy importance to categorise AD

This Graph Signifies that patient with "GM left middle frontal gyrus" less than 16 are having higher chances to be in AD Group

2.5.2 Feature importance for CN

2.5.2.1 Feature importance for CN category patients.

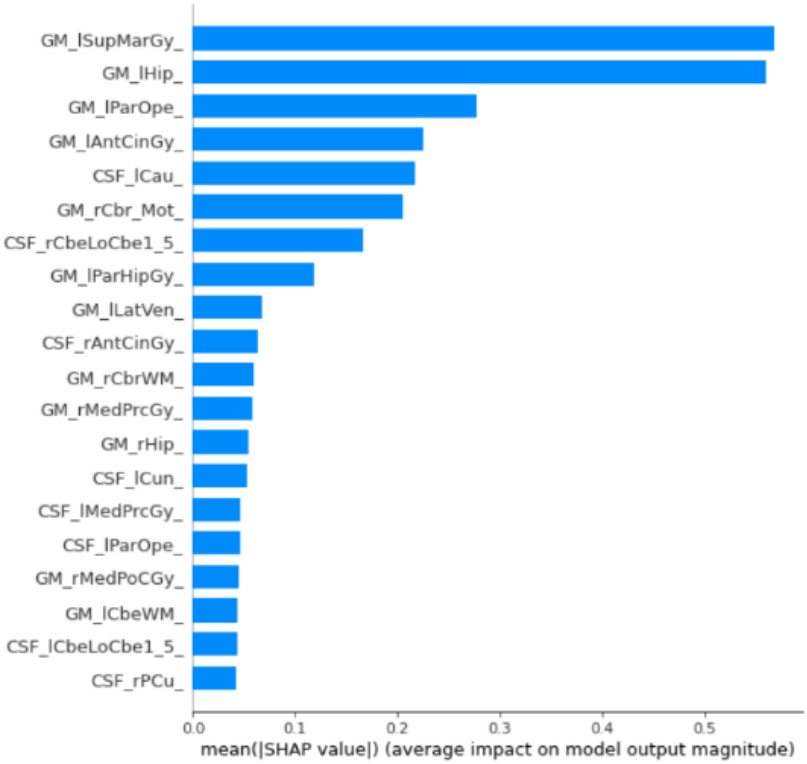


Fig 42 Code for feature importance for CN

These are the important features to correctly classify CNGroup

1. GM of left supramarginal gyrus
2. GM of left hippocampus
3. GM of left parential operculum
4. GM of left anterior cinguli gyrus
5. CSF of left caudate

2.5.2.2 Feature GM_lHip importance to categorise CN

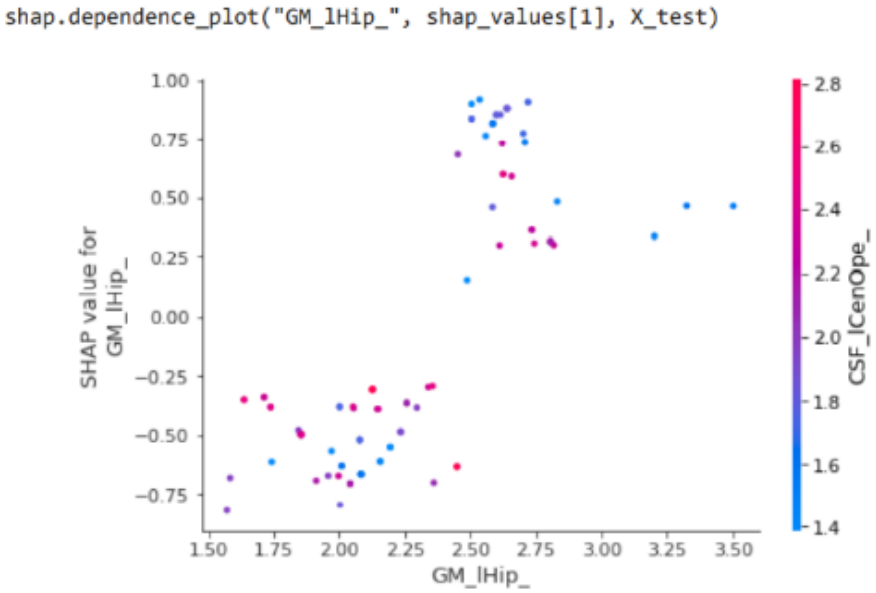


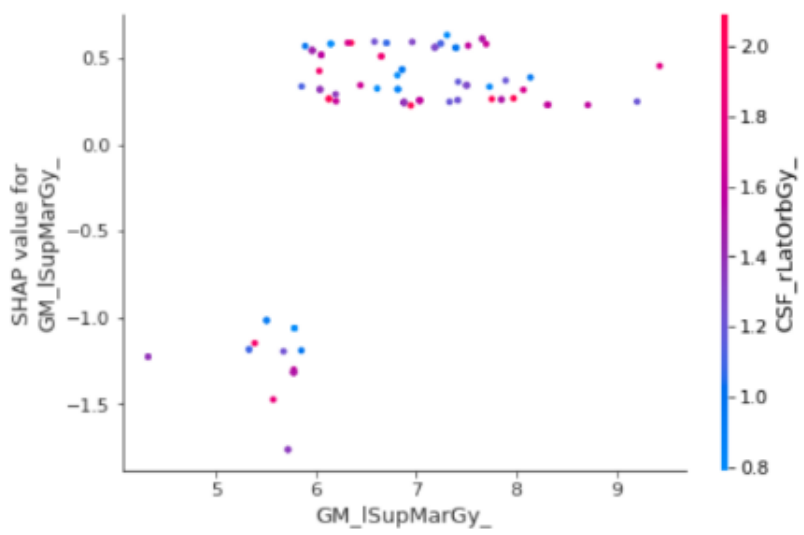
Fig 43 Plot to see GM_lHip importance to categorise CN

This Graph Signifies that patient with "GM left hippocampus" more than 2.5 are having higher chances to be in CN Group.

2.5.2.3 Feature GM lSupMarGyimportance to categorise CN

This Graph Signifies that patient with "GM lSupMarGy" more than 6 are having higher chances to be in CN Group

```
shap.dependence_plot("GM_lSupMarGy_", shap_values[1], X_test)
```



,

Fig 44 Plot to see GM lSupMarGy importance to categorise CN

This Graph Signifies that patient with "GM left supramarginal gyrus" more than 6 are having higher chances to be in CN Group

2.5.2.4 Feature GM lAntCinGy importance to categorise CN

```
shap.dependence_plot("GM_lAntCinGy_", shap_values[1], X_test)
```

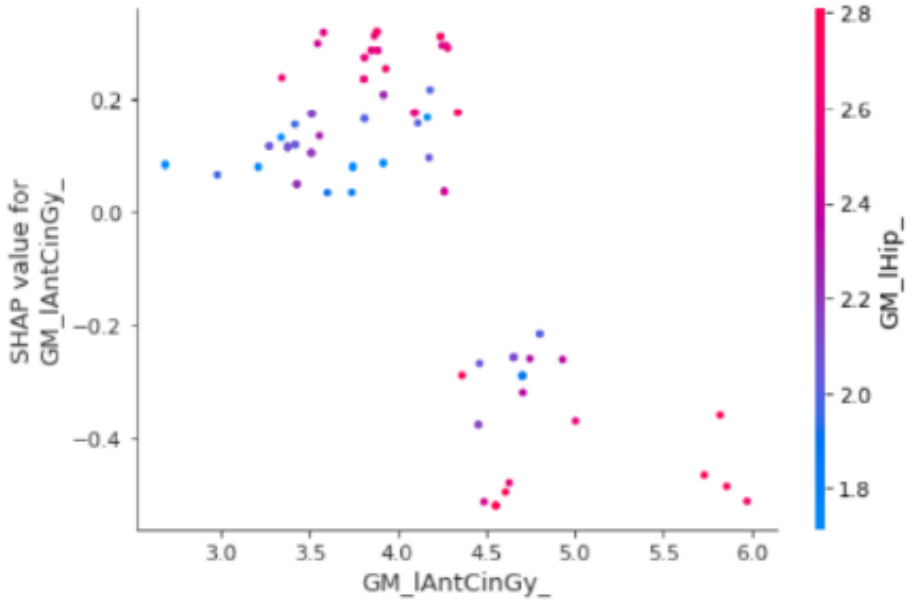


Fig 45 Plot to see GM lAntCinGy importance to categorise CN

This Graph Signifies that patient with "GM left anterior cinguli gyrus" less than 4.4 are having higher chances to be in CN Group

2.5.3 Feature importance for MCI

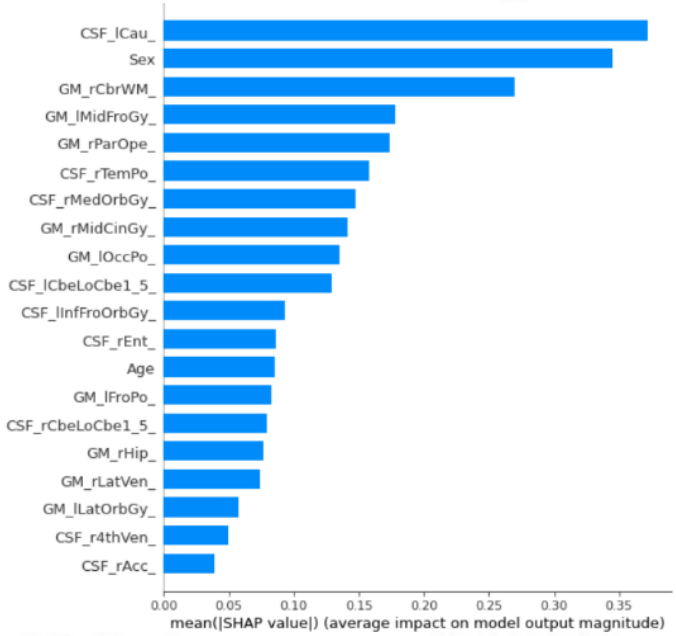


Fig 46 plot to see feature importance for MCI

These are the features important for MCI group

1. CSF of left Caudate
2. Sex
3. Right Cerebral White matter
4. GM of Left middle frontal gyrus
5. GM of right parental operculum

2.5.3.1 Feature “Sex” importance to categorise into MCI

```
shap.dependence_plot("Sex", shap_values[2], X_test)
```

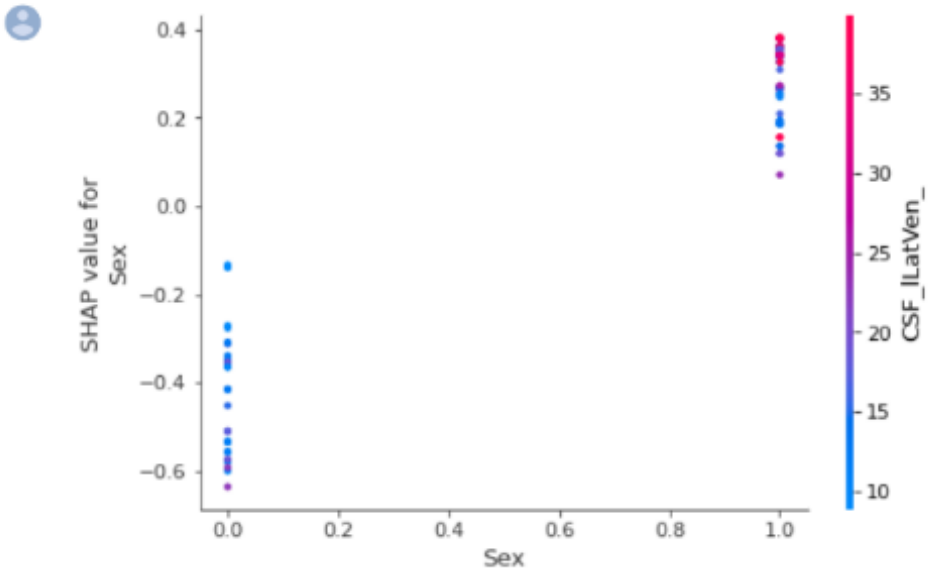


Fig 47 Feature "Sex" importance to categorise into MCI

This Graph Signifies that patient with "Sex" Male are having higher chances to be in MCI Group

2.5.3.2 Feature "CSF lCau" importance to categorise into MCI

is Graph Signifies that patient with "CSF lCau" more than 0.35 are having higher chances to
in MCI Group

```
shap.dependence_plot("CSF_lCau_", shap_values[2], X_test)
```

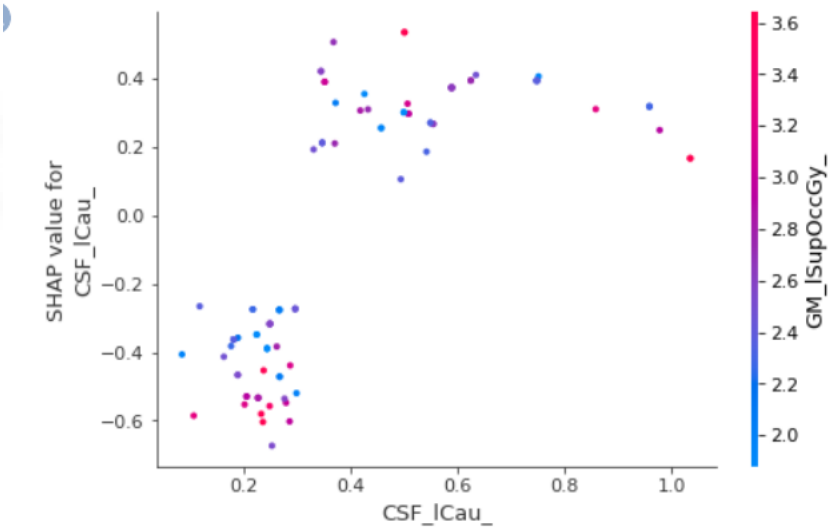


Fig 48 Plot to see thefeature "CSF lCau" importance to categorise into MCI

This Graph Signifies that patient with **CSF of left caudate**" more than 0.35 are having higher chances to be in MCI Group.

2.5.3.3 Feature " GM rCbrWM " importance to categorise into MCI

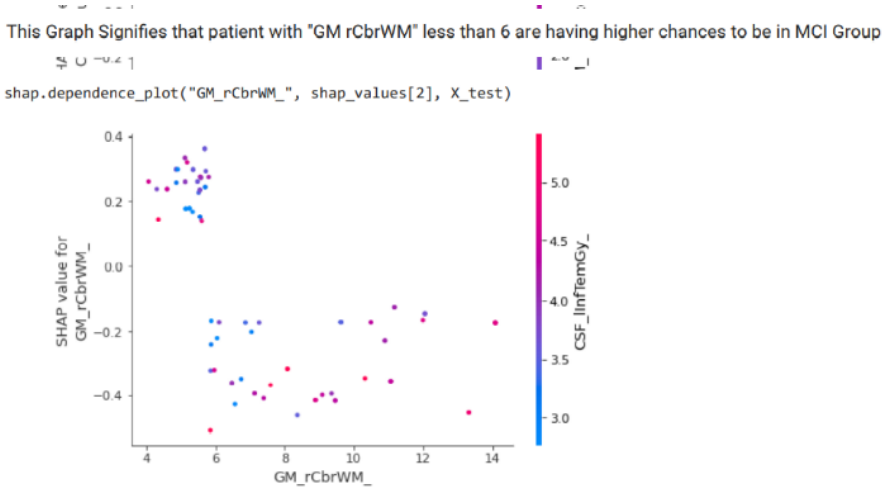


Fig **49 Plot to see the feature "GM rCbrWM"importance to categorise into MCI**

This Graph Signifies that patient with "GM right cerebral white matter" less than 6 are having higher chances to be in MCI Group.

4.7. Important features to categorise each category:

Important features to correctly classify AD Group	Important features to correctly classify CN Group	Important features to correctly classify CN Group
1. CSF of right occipital fusiform gyrus 2. GM of left middle frontal gyrus 3. CSF of right occipital pole 4. CSF of left para hippocampus gyrus 5. CSF of right hippocampus	1. GM of left supramarginal gyrus 2. GM of left hippocampus 3. GM of left parential operculum 4. GM of left anterior cinguli gyrus 5. CSF of left caudate	1. CSF of left Caudate 2. Sex 3. Right Cerebral White matter 4. GM of Left middle frontal gyrus 5. GM of right parental operculum

Table 3 Important features to categorise each group.

CONCLUSION

This research has been done on 289 patients , we got the SMRI dataset from the ADNI website and then with the help of all those patient data we extracted the features using MATLAB with the help of SPM12 and CAT12 toolbox. Then with the help of different machine learning techniques we calculated the accuracy and that can be shown from the following table :

Machine Learning Algorithms	**Accuracy on Testing Dataset on Whole Features**	**Accuracy with Stratified k-fold**
Decision Tree	89.55	91.69
Random Forest	91.48	92.35
Logistic Regression	89.55	94.11
XG boosting	89.55	91.69
Gradient Boosting	89.55	91.38
Light GBM	94.55	93.64

And with the help of different machine learning feature selection techniques like SHAP and RFE we come to conclusion that out of all 142 ROIs some of them are very important to classify them into different groups those features can be shown as following :

1. CSF of right occipital fusiform gyrus
2. GM of left middle frontal gyrus
3. CSF of right occipital pole
4. CSF of left para hippocampus gyrus
5. CSF of right hippocampus
6. GM of left supramarginal gyrus
7. GM of left hippocampus
8. GM of left parential operculum
9. GM of left anterior cinguli gyrus
10. CSF of left caudate

DISCUSSION

Using structural MRI data , we are able to classify the patients with more than 90% accuracy using multi class model with the help of six different machine learning algorithms. RFE and SHAP were the two different feature selection algorithms used in our research , that helped us to know that whether with the help of top features also we will be able to classify the patients. Our research shows that machine learning algorithms can be applied to the neuroimaging data to successfully help the doctors while during diagnosis of dementia patient. Feature extraction from structural MRI can take an average of 40 to 45 minutes per patient using SPM12 and CAT12 toolbox and then with the help of Machine learning algorithms we can help the doctors. Our method presents a faster way of diagnosis.

The presence of features that are consistently important across all models is shown by observing feature weights. Some of the important features are as followsHippocampus, Occipital fusiform gyrus, Caudate, Para hippocampus gyrus, Anterior cinguli gyrus

Unsurprisingly, earlier molecular and clinical investigations have linked several of these areas to Alzheimer's disease and other neurological illnesses that induce cognitive impairment.

Combining structural MRI data with other neuroimaging data, such as functional MRI (fMRI) and positive emission tomography (PET), may improve the model's performance to the point where machine learning models can outperform and replace the traditional clinical diagnostic model.

APPENDIX : CODE USED

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

import os
df=pd.read_csv('Vishal_alz_detect.csv')
df.shape
#Clean the column name by removing speacial character/space/symbols
df.columns = df.columns.str.replace(r"[^a-zA-Z\d\_]+", "_")
#Removing missing rows and duplicated rows
df=df.dropna().drop_duplicates().reset_index(drop=True)
df.head()
df.shape
df.describe()
df.describe(include=object)
df.info()
df.Group.value_counts()
df.Sex=df.Sex.map({'F':0,'M':1})
df.Group=df.Group.map({'CN':0,'MCI':1,'AD':2})
df.Group.value_counts(normalize=True)
ax = sns.countplot(df['Group'],label="Count")       # M = 212, B = 357
CN, MCI, AD = df.Group.value_counts()
print('Number of Normal people: ',MCI)
print('Number of MCI patient : ',CN)
print('Number of AD patient : ',AD)

from sklearn.model_selection import train_test_split
feat=list(df.columns.drop(['Image_Data_ID','Subject','Group']))
#Train test split
X_train, X_test, y_train, y_test = train_test_split( df[feat],
                                                     df['Group'],
                                                     test_size=0.33,
                                                     stratify=df['Group'],
```

```
                                    shuffle=True,
                                    random_state=42)
#The following code is for Decision Tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
dtc=DecisionTreeClassifier(max_depth=10)
dtc.fit(X_train,y_train)
accuracy_score(y_test,dtc.predict(X_test))

from sklearn.metrics import confusion_matrix
confusion_matrix(y_test,dtc.predict(X_test))

from sklearn.metrics import classification_report
print(classification_report(y_test,dtc.predict(X_test)))

from sklearn.metrics import roc_curve


pred_prob = dtc.predict_proba(X_test)

# roc curve for classes
fpr = {}
tpr = {}
thresh ={}
mmap=pd.Series(np.arange(0,len(dtc.classes_)),index=dtc.classes_ ).to_dict()

n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(np.where(y_test.apply(lambda x: mmap[x])==i,1,0), pre
d_prob[:,i])

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class CN vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class MCI vs Rest')
```

```
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue', label='Class AD vs Rest')
plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.savefig('Multiclass ROC',dpi=300);

from sklearn.model_selection import cross_val_score

from sklearn.model_selection import StratifiedKFold

cv = StratifiedKFold(n_splits=4, shuffle=True, random_state=1)

scores = cross_val_score(dtc, df[feat], df['Group'], scoring = 'accuracy', cv = cv)

print(scores.mean())

import matplotlib

import matplotlib.pyplot as plt

import seaborn as sns
matplotlib.rc('xtick', labelsize=20)

matplotlib.rc('ytick', labelsize=20)
%matplotlib inline
d_rad_range = range(1, 31)
d_scores = []
for d in d_rad_range:
  dt = DecisionTreeClassifier(max_depth=d)
  scores = cross_val_score(dt, df[feat], df['Group'], cv=10, scoring='accuracy')
  d_scores.append(scores.mean())
  plt.figure(figsize=(10, 5))

sns.set_style("darkgrid")
```

```
plt.plot(d_rad_range, d_scores)

plt.xlabel('d', size=20)

plt.ylabel('d_scores', size=20)

#The following code is for the Random Forest
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier()
# fit the model with the training data
rfc.fit(X_train,y_train)
# number of trees used
print('Number of Trees used : ', rfc.n_estimators)
# predict the target on the train dataset
rfc_predict_train = rfc.predict(X_train)
print('\nTarget on train data',rfc_predict_train)

# Accuray Score on train dataset
rfc_accuracy_train = accuracy_score(y_train,rfc_predict_train)
print('\naccuracy_score on train dataset : ', rfc_accuracy_train)

# predict the target on the test dataset
rfc_predict_test = rfc.predict(X_test)
print('\nTarget on test data',rfc_predict_test)

# Accuracy Score on test dataset
rfc_accuracy_test = accuracy_score(y_test,rfc_predict_test)
print('\naccuracy_score on test dataset : ', rfc_accuracy_test)

from sklearn.metrics import classification_report
print(classification_report(y_test,rfc.predict(X_test)))

from sklearn.metrics import roc_curve
```

```
pred_prob = rfc.predict_proba(X_test)

# roc curve for classes
fpr = {}
tpr = {}
thresh ={}
mmap=pd.Series(np.arange(0,len(rfc.classes_)),index=rfc.classes_ ).to_dict()

n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(np.where(y_test.apply(lambda x: mmap[x])==i,1,0), pre
d_prob[:,i])

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class CN vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class MCI vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue', label='Class AD vs Rest')
plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.savefig('Multiclass ROC',dpi=300);

from sklearn.model_selection import cross_val_score

from sklearn.model_selection import StratifiedKFold

cv = StratifiedKFold(n_splits=4, shuffle=True, random_state=1)

scores = cross_val_score(rfc, df[feat], df['Group'], scoring = 'accuracy', cv = cv)

print(scores.mean())
```

```
#The following code is for Gradient Boosting
from sklearn.ensemble import GradientBoostingClassifier
gbc = GradientBoostingClassifier(n_estimators=100,max_depth=5)

# fit the model with the training data
gbc.fit(X_train,y_train)

# predict the target on the train dataset
gbc_predict_train = gbc.predict(X_train)
print('\nTarget on train data',gbc_predict_train)

# Accuray Score on train dataset
gbc_accuracy_train = accuracy_score(y_train,gbc_predict_train)
print('\naccuracy_score on train dataset : ', gbc_accuracy_train)

# predict the target on the test dataset
gbc_predict_test = gbc.predict(X_test)
print('\nTarget on test data',gbc_predict_test)

# Accuracy Score on test dataset
gbc_accuracy_test = accuracy_score(y_test,gbc_predict_test)
print('\naccuracy_score on test dataset : ', gbc_accuracy_test)

from sklearn.metrics import roc_curve


pred_prob = gbc.predict_proba(X_test)

# roc curve for classes
fpr = {}
tpr = {}
thresh ={}
mmap=pd.Series(np.arange(0,len(gbc.classes_)),index=gbc.classes_ ).to_dict()
```

```
n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(np.where(y_test.apply(lambda x: mmap[x])==i,1,0), pre
d_prob[:,i])

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class CN vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class MCI vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue', label='Class AD vs Rest')
plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.savefig('Multiclass ROC',dpi=300);

from sklearn.model_selection import cross_val_score

from sklearn.model_selection import StratifiedKFold

cv = StratifiedKFold(n_splits=4, shuffle=True, random_state=1)

scores = cross_val_score(gbc, df[feat], df['Group'], scoring = 'accuracy', cv = cv)

print(scores.mean())
# multi-class classification
from sklearn.multiclass import OneVsRestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score



# fit model
clf = OneVsRestClassifier(LogisticRegression(solver='lbfgs', max_iter=1000))
```

```
clf.fit(X_train, y_train)
# predict the target on the train dataset
clf_predict_train = clf.predict(X_train)
print('\nTarget on train data',clf_predict_train)

# Accuray Score on train dataset
clf_accuracy_train = accuracy_score(y_train,clf_predict_train)
print('\naccuracy_score on train dataset : ', clf_accuracy_train)

# predict the target on the test dataset
clf_predict_test = clf.predict(X_test)
print('\nTarget on test data',clf_predict_test)

# Accuracy Score on test dataset
clf_accuracy_test = accuracy_score(y_test,clf_predict_test)
print('\naccuracy_score on test dataset : ', gbc_accuracy_test)

from sklearn.metrics import classification_report
print(classification_report(y_test,clf.predict(X_test)))

from sklearn.model_selection import cross_val_score

from sklearn.model_selection import StratifiedKFold

cv = StratifiedKFold(n_splits=4, shuffle=True, random_state=1)

scores = cross_val_score(clf, df[feat], df['Group'], scoring = 'accuracy', cv = cv)

print(scores.mean())

pred = clf.predict(X_test)
pred_prob = clf.predict_proba(X_test)

# roc curve for classes
fpr = {}
```

```
tpr = {}
thresh ={}

n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(y_test, pred_prob[:,i], pos_label=i)

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class CN vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class MCI vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue', label='Class AD vs Rest')
plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.savefig('Multiclass ROC',dpi=300);

#The following code is for XGBOOST
from xgboost.sklearn import XGBClassifier
xgb = XGBClassifier()

# fit the model with the training data
xgb.fit(X_train,y_train)

# predict the target on the train dataset
xgb_predict_train = xgb.predict(X_train)
print('\nTarget on train data',xgb_predict_train)

# Accuray Score on train dataset
xgb_accuracy_train = accuracy_score(y_train,xgb_predict_train)
print('\naccuracy_score on train dataset : ', xgb_accuracy_train)

# predict the target on the test dataset
xgb_predict_test = xgb.predict(X_test)
```

```
print('\nTarget on test data',xgb.predict_test)

# Accuracy Score on test dataset
xgb_accuracy_test = accuracy_score(y_test,xgb_predict_test)
print('\naccuracy_score on test dataset : ', xgb_accuracy_test)

from sklearn.metrics import roc_curve


pred_prob = xgb.predict_proba(X_test)

# roc curve for classes
fpr = {}
tpr = {}
thresh ={}
mmap=pd.Series(np.arange(0,len(xgb.classes_)),index=xgb.classes_ ).to_dict()

n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(np.where(y_test.apply(lambda x: mmap[x])==i,1,0), pre
d_prob[:,i])

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class CN vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class MCI vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue', label='Class AD vs Rest')
plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.savefig('Multiclass ROC',dpi=300);

from sklearn.metrics import roc_curve
```

```
pred_prob = xgb.predict_proba(X_test)

# roc curve for classes
fpr = {}
tpr = {}
thresh ={}
mmap=pd.Series(np.arange(0,len(xgb.classes_)),index=xgb.classes_ ).to_dict()

n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(np.where(y_test.apply(lambda x: mmap[x])==i,1,0), pre
d_prob[:,i])

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class CN vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class MCI vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue', label='Class AD vs Rest')
plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.savefig('Multiclass ROC',dpi=300);

from sklearn.metrics import classification_report
print(classification_report(y_test,xgb.predict(X_test)))

from sklearn.model_selection import cross_val_score

from sklearn.model_selection import StratifiedKFold

cv = StratifiedKFold(n_splits=4, shuffle=True, random_state=1)
```

```
scores = cross_val_score(xgb, df[feat], df['Group'], scoring = 'accuracy', cv = cv)

print(scores.mean())

from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
# find best scored 5 features
select_feature = SelectKBest(chi2, k=5).fit(X_train, y_train)
print('Score list:', select_feature.scores_)
print('Feature list:', X_train.columns)

from sklearn.feature_selection import RFE
# Create the RFE object and rank each pixel
rfc = RandomForestClassifier()
rfe = RFE(estimator=rfc, n_features_to_select=5, step=1)
rfe = rfe.fit(X_train, y_train)

from sklearn.feature_selection import RFECV

# The "accuracy" scoring is proportional to the number of correct classifications
rfc = RandomForestClassifier()
rfecv = RFECV(estimator=rfc, step=1, cv=5,scoring='accuracy')   #5-fold cross-validation
rfecv = rfecv.fit(X_train, y_train)

print('Optimal number of features :', rfecv.n_features_)
print('Best features :', X_train.columns[rfecv.support_])

# Plot number of features VS. cross-validation scores
import matplotlib.pyplot as plt
plt.figure()
plt.xlabel("Number of features selected")
plt.ylabel("Cross validation score of number of selected features")
plt.plot(range(1, len(rfecv.grid_scores_) + 1), rfecv.grid_scores_)
plt.show()
```

```
from lightgbm import LGBMClassifier
lgbc=LGBMClassifier(learning_rate=0.1,
                    n_estimators=50,
                    num_leaves=8)
lgbc.fit(X_train,y_train)
pred=lgbc.predict(X_test)

from sklearn.metrics import roc_curve



pred_prob = lgbc.predict_proba(X_test)

# roc curve for classes
fpr = {}
tpr = {}
thresh ={}
mmap=pd.Series(np.arange(0,len(lgbc.classes_)),index=lgbc.classes_ ).to_dict()

n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(np.where(y_test.apply(lambda x: mmap[x])==i,1,0), pre
d_prob[:,i])

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class CN vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class MCI vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue', label='Class AD vs Rest')
plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.savefig('Multiclass ROC',dpi=300);

from sklearn.model_selection import cross_val_score
```

```
from sklearn.model_selection import StratifiedKFold

cv = StratifiedKFold(n_splits=4, shuffle=True, random_state=1)

scores = cross_val_score(lgbc, df[feat], df['Group'], scoring = 'accuracy', cv = cv)

print(scores.mean())

from sklearn.metrics import classification_report

#Multi-class classification report
print(classification_report(y_test, pred))

from sklearn.metrics import precision_score,f1_score
sc1=np.round(precision_score(y_test,pred,average='macro'),4)
sc2=np.round(precision_score(y_test,pred,average='micro'),4)
f1=np.round(f1_score(y_test,pred,average='weighted'),4)
print(f'Weighted F1-
Score: {np.round(np.mean(f1),4)}, Macro precision score {np.round(np.mean(sc1),4)}, Micr
o precision score {np.round(np.mean(sc2),4)}\n')

import shap
shap_values = shap.TreeExplainer(lgbc).shap_values(X_test)
shap.summary_plot(shap_values, X_test, plot_type="bar",class_names=list(lgbc.classes_))

import shap
shap_values = shap.TreeExplainer(lgbc).shap_values(X_test)
shap.summary_plot(shap_values[0], X_test,plot_type="bar",class_names=list(lgbc.classes_)[
0])

shap.dependence_plot("CSF_rOccFusGy_", shap_values[0], X_test)
shap.dependence_plot("CSF_rOccPo_", shap_values[0], X_test)

shap.dependence_plot("GM_lMidFroGy_", shap_values[0], X_test)
```

```
import shap
shap_values = shap.TreeExplainer(lgbc).shap_values(X_test)
shap.summary_plot(shap_values[1], X_test,plot_type="bar",class_names=list(lgbc.classes_)[
1])

shap.dependence_plot("GM_lHip_", shap_values[1], X_test)

shap.dependence_plot("GM_lSupMarGy_", shap_values[1], X_test)

shap.dependence_plot("GM_lAntCinGy_", shap_values[1], X_test)

import shap
shap_values = shap.TreeExplainer(lgbc).shap_values(X_test)
shap.summary_plot(shap_values[2], X_test,plot_type="bar",class_names=list(lgbc.classes_)[
2])
shap.dependence_plot("Sex", shap_values[2], X_test)

shap.dependence_plot("CSF_lCau_", shap_values[2], X_test)

shap.dependence_plot("GM_rCbrWM_", shap_values[2], X_test)
```

REFERENCES

1. *Alzheimer's Disease, available at* https://www.writework.com/essay/alzheimer-s-disease.
2. MAYO CLINIC, *available at* https://www.mayoclinic.org/.
3. *Structural MRI Imaging*, UC SAN DIEGO, *available at* http://fmri.ucsd.edu/Howto/3T/structure.html.
4. *Structural MRI Imaging*, UC SAN DIEGO, *available at* http://fmri.ucsd.edu/Howto/3T/structure.html.
5. *Grey Matter, available* at https://en.wikipedia.org/wiki/Grey_matter.
6. *Available at* https://en.wikipedia.org/wiki/Cerebrospinal_fluid#/media/File:Blausen_0216_CerebrospinalSystem.png.
7. *What is Alzheimer's Disease? Symptoms & Causes*, *available at* alz.org.
8. Ron Brookmeyer et al, *Forecasting the global burden of Alzheimer's disease*, ALZHEIMER'S ASSOCIATION, *available* at https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2007.04.381.
9. Chris Hinrichs et al, *Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population*, EUROPE PMC, *available at* http://europepmc.org/article/MED/21146621.
10. Tiraboschi et al., *The importance of neuritic plaques and tangles to the development and evolution of AD*, NEUROLOGY 62.11, 1984–1989 (2004). *Available at*https://pubmed.ncbi.nlm.nih.gov/15184601/.
11. Earlier Diagnosis, ALZHEIMER'S ASSOCIATION, *available at*https://www.alz.org/alzheimers-dementia/research_progress/earlier-diagnosis.
12. Marilyn S. Albert et al., *The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease*, PMC, 270–279 (2012). *Available at* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3312027/
13. Yoon-Suk Han , Wyatt Hong , *Classification of Alzheimer's Patients Using Structural MRI Data*
14. Earlier Diagnosis, ALZHEIMER'S ASSOCIATION, *available at*https://www.alz.org/alzheimers-dementia/research_progress/earlier-diagnosis.
15. Yi Ren Fung , Ziqiang Guan, *Alzheimer's Disease Brain MRI Classification: Challenges and Insights ,* University of Massachusetts Amherst (2019)
16. Pascal Vincent et al., *Extracting and composing robust features with denoising autoencoders*, ICML '08: PROCEEDINGS OF THE 25TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, ACM DIGITAL LIBRARY, 1096–1103 (2008). *Available at*https://dl.acm.org/doi/10.1145/1390156.1390294
17. J C Morris et al., *Cerebral amyloid deposition and diffuse plaques in "normal" aging: Evidence for presymptomatic and very mild Alzheimer's disease*, NEUROLOGY 46.3, 707–719 (1996). *Available at*https://pubmed.ncbi.nlm.nih.gov/8618671/.
18. Tiraboschi et al., *The importance of neuritic plaques and tangles to the development and evolution of AD*, NEUROLOGY 62.11, 1984–1989 (2004). *Available at* https://www.researchgate.net/publication/8521699_The_Importance_of_neuritic_plaques_and_tangles_to_the_development_and_evolution_of_AD.

19. M P Laakso et al., *Hippocampal volumes in Alzheimer's disease, Parkinson's disease with and without dementia, and in vascular dementia: An MRI study*, NEUROLOGY 46.3, 678–681, (1996). *Available at* https://pubmed.ncbi.nlm.nih.gov/8618666/.
20. Oskar Hansson et al., *Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study*, 5(3), THE LANCET NEUROLOGY, 228–234 (2006).
21. Jieping Ye et al., *Machine learning approaches for the neuroimaging study of Alzheimer's Disease*, 44(4), COMPUTER, 99–101 (2011).
22. *Ibid*.
23. Gary L. Wenk, *Neuropathologic changes in Alzheimer's disease*, 64 Suppl 9, THE JOURNAL OF CLINICAL PSYCHIATRY, 7–10 (2003).
24. *Study Design,* ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE, *Available at*http://adni.loni.usc.edu/study-design/.
25. *What is Alzheimer's Disease? Symptoms & Causes, available at* alz.org.
26. MAGNETIC RESONANCE IMAGING, *available at* https://www.journals.elsevier.com/magnetic-resonance-imaging/.
27. Basic Principles of MRI, *available at* sfu.ca.
28. Available at *2.2 Basic Principles of MRI (sfu.ca)*
29. Arpit Raut, *A Machine Learning Based Approach for Detection of Alzheimer's Disease Using Analysis of Hippocampus Region from MRI Scan* , VIT , UNIVERSITY OF MUMBAI , INDIA
30. Basic Principles of MRI, *available at* sfu.ca.
31. Ahsan Bin Tufail, *Binary Classification of Alzheimer Disease using sMRI Imaging modality and Deep Learning,* CORNELL UNIVERSITY, (2020), *available at* https://arxiv.org/abs/1809.06209.
32. *Structural MRI*, INTERNATIONAL PSYCHOGERIATRICS,23 Suppl 2(S2):S13-24,(2011). *Available at* https://www.researchgate.net/publication/51467509_Structural_MRI.
33. *Structural MRI,* THE UNIVERSITY OF EDINBURGH, *available at*https://www.ed.ac.uk/clinical-sciences/edinburgh-imaging/research/themes-and-topics/medical-physics/imaging-techniques/structural-mri.
34. Mike P. Wattjes , *Structural MRI* , VU MEDICAL CENTER, AMSTERDAM
35. *Structural MRI,* THE UNIVERSITY OF EDINBURGH, *available at*https://www.ed.ac.uk/clinical-sciences/edinburgh-imaging/research/themes-and-topics/medical-physics/imaging-techniques/structural-mri.
36. Kimberly A. Stigleret al., *Structural and Functional MRI Studies of Autism Spectrum Disorders,* STRUCTURAL MAGNETIC RESONANCE IMAGING, SCIENCE DIRECT,(2013), *available at*https://www.sciencedirect.com/topics/medicine-and-dentistry/structural-magnetic-resonance-imaging.
37. *Available at* https://www.scirp.org/reference/ReferencesPapers.aspx?ReferenceID=1341795.
38. Bruno Dubois et al, *Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria*, NATIONAL LIBRARY OF MEDICINE, 734-46 (2007), *available at*https://pubmed.ncbi.nlm.nih.gov/17616482/.
39. Frisoni GB et al., *The clinical use of structural MRI in Alzheimer disease*, EUROPE PMC, 67-77 (2010), *available at*https://europepmc.org/article/MED/20139996.

40. *Structural MRI*, (2011), *available at*https://www.cambridge.org/core/journals/international-psychogeriatrics/article/abs/structural-mri/72026464ED51670F2C1DB9B0761207BE.
41. *Available at*https://www.researchgate.net/publication/51467509_Structural_MRI.
42. Michael W. Weiner et al., *The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception*, ALZHEIMER'S ASSOCIATION, (2011) *available at* https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2011.09.172.
43. *Available at* 31-1394871866.pdf.
44. Dan Claudiu Cireşan et al, *Deep, Big, Simple Neural Nets for Handwritten Digit Recognition*, 22(2),Neural Computation(2010), *available at* https://ieeexplore.ieee.org/document/6797043.
45. Tom M. Mitchell et al., *Machine Learning, available at* http://www.cs.cmu.edu/~tom/mlbook.html.
46. Sandhya Joshi, *Classification and treatment of different stages of alzheimer's disease using various machine learning methods*, 2(1), INTERNATIONAL JOURNAL OF BIOINFORMATICS RESEARCH, 44-52 (2010).
47. *Classification of Alzheimer's Disease using Machine Learning Techniques, available at* https://www.researchgate.net/publication/335064063_Classification_of_Alzheimer%27s_Disease_using_Machine_Learning_Techniques.
48. Sandhya Joshi, Vibhudendra Simha G.G. , *Classification and treatment of different stages of alzheimer's disease using various machine learning methods* , VCE, BANGALORE, INDIA
49. *Classification and Investigation of Alzheimer Disease Using Machine Learning Algorithms*, 13 (13), Biosc.Biotech.Res.Comm., 15-20, (2020), *available at*https://bbrc.in/wp-content/uploads/2021/03/Spcial-Issue-13-13-3.pdf.
50. Randy L. Buckneret al., *A unified approach for morphometric and functional data analysis inyoung, old, and demented adults using automated atlas-based headsize normalization: reliability and validation against manual measurement of total intracranial volume*, 23, NEUROIMAGE, 724–738(2004).
51. Yi Ren Fungand et al., *Alzheimer's Disease Brain MRI Classification: Challenges and Insights*, UNIVERSITY OF MASSACHUSETTS AMHERST, (2019).
52. Ivana Despotovit, *MRI Segmentation of the Human Brain: Challenges,Methods, and Applications*, COMPUTATIONAL AND MATHEMATICAL METHODS IN MEDICINE, 1-23, (2015).
53. Yoon-Suk Han and Wyatt Hong, *Classification of Alzheimer's Patients Using Structural MRI Data*, (2013).
54. Chris Hinrichs et al., *Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population*, 55, NEUROIMAGE, 574-589 (2011).
55. Arpita Raut and Vipul Dalal, *A Machine Learning Based Approach for Detectionof Alzheimer's Disease Using Analysis of Hippocampus Region from MRI Scan*, Proceedings of the IEEE 2017 INTERNATIONAL CONFERENCE ON COMPUTING METHODOLOGIES AND COMMUNICATION, 236-242 (2017).

56. Osman Hegazy, *A Machine Learning Model for Stock MarketPrediction*, 4(12), INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TELECOMMUNICATIONS, 17-23 (2013).
57. Charles K. Fisher, *Machine learning for comprehensive forecasting ofAlzheimer's Disease progression*, SCIENTIFIC REPORTS (2019).
58. Heung-Il Suk and Dinggang Shen, *Deep Learning-Based Feature Representation for AD/MCI Classification*, MED IMAGE COMPUT COMPUT ASSIST INTERV., 583–590 (2013).
59. Lucia Billeci et al., *Machine Learning for the Classification ofAlzheimer's Disease and Its Prodromal Stage UsingBrain Diffusion Tensor Imaging Data:A Systematic Review*, PROCESSES (2020).
60. A. Collie and P. Maruff, *The neuropsychology of preclinical Alzheimer's disease and mildcognitive impairment*, 24, NEUROSCIENCE AND BIOBEHAVIORAL REVIEWS, 365–374 (2000).
61. Saman Sargolzaei et al., *Estimating Intracranial Volume in Brain Research:An Evaluation of Methods*, 13, NEUROINFORM, 427–441 (2015).
62. Vânia Tavares et al., *Comparing SPM12 and CAT12 segmentation pipelines: a brain tissuevolume-based age and Alzheimer's disease study*, JOURNAL OF NEUROSCIENCE METHODS, 334 (2020).
63. ANALYTICS VIDHYA, *available at* https://www.analyticsvidhya.com/blog/tag/feature-selection/ (visited on April 24, 2021).
64. *Feature Selection Techniques in Machine Learning*, ANALYTICS VIDHYA (OCTOBER 10, 2020)*available at*https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/(visited on April 24, 2021).
65. *Alzheimer's Disease Neuroimaging Initiative*, *available at*http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Data_Use_Agreement.pdf(visited on April 21, 2021).
66. *LightGBM Classifier in Python*, KAGGLE, *available at*https://www.kaggle.com/prashant111/lightgbm-classifier-in-python(visited on April 21, 2021).
67. RENDYK, *Distinguish between Tree-Based Machine Learning Algorithms*, ANALYTICS VIDHYA (April 15, 2021)*available athttps://www.analyticsvidhya.com/blog/2021/04/distinguish-between-tree-based-machine-learning-algorithms/*(visited on April 20, 2021).
68. Akshay Sharma, *Machine Learning 101: Decision Tree Algorithm For Classification*,ANALYTICS VIDHYA (February 25, 2021) *available athttps://www.analyticsvidhya.com/blog/2021/02/machine-learning-101-decision-tree-algorithm-for-classification/*(visited on April 11, 2021).
69. Vishesh Arora, *13 Most Important Pandas Functions for Data Science*, ANALYTICS VIDHYA (May 13, 2021) *available athttps://www.analyticsvidhya.com/blog/2021/05/pandas-functions-13-most-important/*(visited on April 12, 2021).
70. Syed Danish, *Practical Guide on Data Preprocessing in Python using Scikit Learn*, ANALYTICS VIDHYA (July 18, 2016) *available*

athttps://www.analyticsvidhya.com/blog/2016/07/practical-guide-data-preprocessing-python-scikit-learn/(visited on April 02, 2021).

71. Pranjal Pandey, *Data Preprocessing : Concep*ts, TOWARDS DATA SCIENCE (November 25, 2019) *available at* https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825(visited on April 03, 2021).
72. *Changes in version CAT12.8 (1830),* STRUCTURAL BRAIN MAPPING GROUP, *available at* http://www.neuro.uni-jena.de/cat12-html/cat_versions.html(visited on April 03, 2021).
73. *Detecting Early Alzheimer's Using Mri Data And Machine Learning,* KAGGLE *available at*https://www.kaggle.com/hyunseokc/detecting-early-alzheimer-s(visited on April 03, 2021).
74. ALLEN BRAIN MAP*, available at* https://portal.brain-map.org(visited on April 03, 2021).
75. *Available at https://i.redd.it/piexn7yw4ob31.jpg*(visited on April 03, 2021).
76. Prasad Patil, *What is Exploratory Data Analysis?*, TOWARDS DATA SCIENCE (March 23, 2018) *available at* https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15(visited on April 03, 2021).
77. *Alzheimers Multiclass-Classification*, KAGGLE *available at*https://www.kaggle.com/kenconstable/alzheimer-s-multi-class-classification(visited on April 30, 2021).
78. ANALYTICS VIDHYA, *available at*https://www.analyticsvidhya.com/blog/tag/lightgbm/(visited on April 03, 2021).
79. Pranjal Khandelwal, *Which algorithm takes the crown: Light GBM vs XGBOOST?*ANALYTICS VIDHYA(June 2, 2017*) available at*https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/(visited on April 03, 2021).