

# MSc Dissertation Report

## **Detection of possible tumour registration(s) and its spatial coordinates on brain from MRI 3D Medical Images using ANN (Multi- Layer Perceptron)**

A dissertation submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of **Master of Science in Big Data Analytics**

Student Name	<b>Raveendrababu Pasumarthi</b>
Student ID	<b>C0028603</b>
Supervisor	<b>Dr. Hemlata Sharma</b>
Date of Submission	<b>8<sup>th</sup> Sep 2022</b>

This dissertation does NOT contain confidential material and thus can be made available to staff and students via the library.

## ACKNOWLEDGEMENTS

With immense gratitude, I would like to express my sincere thanks to the faculty of **Dept. of Computing, Sheffield Hallam University (SHU)**, particularly **Dr. Hemlata Sharma** for their continuous support, encouragement, valuable suggestions, and feedback throughout my research and made the work possible. I would like to extend my thanks with due respect to **Dr. Arul N Selvan**, a Researcher in SHU and Philosopher for sharing their brainchild thought and idea of brain tumor registrations and corresponding spatial coordinates prediction using machine learning. I would also like to extend my sincere thanks to **BRaTS Team** (Bhakti Baheti, et. al. 2022) for accepting my registration for the dataset and for their interaction during dataset processing. I would like to extend further my thanks to my brother **Chandra Sekhar Pasumarthi** for sparing their time to discuss on Deep Learning and Python. I must be thankful to **Vaibhav Ranjan**, my classmate for their valuable tips and thought process discussions during the time of dissertation. I must also be thankful to the **Srikanth Ankisetty**, my friend for their moral support to stay strong throughout the research work and my family (wife **Vajra Kumari** and daughter **Vaidurya Ruth**) for being the driving force of my activities during the dissertation period

## **ABSTRACT:**

Prediction, detection of possible tumour affected regions of brain using MR Images using segmentation and AI are being extensively under study and gained the ground in exponential pace thus contributing to medicine domain. Identifying possible spatial correspondences across the longitudinal scans due to the deformation and change of appearance of brain tissue is still a challenge. An attempt has been made to predict the possible tumour spatial coordinates on highly curated (by domain experts) 3d MRI dataset (Bhakti, Diana, Hamed, & Satrajit, 2021) that has base(pre-operative) and follow up scan's combination for at least 140 patients (total of 280 image files of t1ce modality 3d MRI scans of NIfTI format) that only supplied the ground truth landmark details of tumour for each scan. The dataset was released by the BRaTS team in 2022 to the registered participants as part of the BRaTS Challenge (Bhakti Baheti et al, 2022). The dataset was successfully pre-processed (NIfTI format files to flat files) by converting anatomical coordinate system to world coordinate system and applied istumor label based on the landmark files provided along with dataset. Results using keras Multi-Layer Perceptron (MLP) for binary label has achieved an accuracy of 89.27 with 25 % of data augmentation with confusion matrix of recall 0.38 for the istumor 1 and overall weighted average recall 0.98 and f1-score 0.99. This novel attempt of predicting tumor correspondences and spatial coordinates on the medical MRI brain images using the ANN Multi- Layer Perceptron sequential model with Keras is obtaining optimistic results. There is a huge scope to develop enhanced models by utilising the key activities performed in the current study in turn helps in identification and diagnosis of tumor registrations on the whole brain not just in the vicinity of the tumor region (Bhakti, Diana, Hamed, & Satrajit, 2021) at the early stages

**Keywords:** NIfTI, istumor, anatomical, BRaTS, correspondences, nii, pre-processing, verification, validation, prediction

# CONTENTS

<b>ACKNOWLEDGEMENTS</b> .....	ii
<b>ABSTRACT:</b> .....	iii
<b>1. Introduction</b> .....	1
<b>1.1 Project Aim</b> .....	2
<b>1.2 Project Scope</b> .....	2
<b>1.3 Research Question</b> .....	3
<b>1.4 Objectives</b> .....	3
<b>1.5 Project Benefits</b> .....	3
<b>1.6 Project Deliverables</b> .....	4
<b>1.7 Summary of the Research Study</b> .....	4
<b>2 Literature Study and Review</b> .....	5
<b>2.1 Discussion on existing studies on brain tumour detection</b> .....	6
2.1.1 A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network .....	7
2.1.2 Brain tumor detection from MRI images using deep learning techniques .....	9
2.1.3 Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM.....	11
2.1.4 A Novel Deep Learning Method for Recognition and Classification of Brain Tumors from MRI Images.....	15
2.1.5 The Brain Tumor Sequence Registration Challenge: Establishing Correspondence between Pre-Operative and Follow-up MRI scans of diffuse glioma patients .....	18
<b>2.2 Anatomical co-ordinate system</b> .....	20
<b>2.3 Literature Review Summary</b> .....	22
<b>3 Research Design</b> .....	22
<b>3.1 Research Ethics</b> .....	22
<b>3.2 Research framework</b> .....	23
<b>3.3 Research methodology</b> .....	25
<b>3.4 Machine Learning</b> .....	28
3.4.1 Artificial Intelligence .....	28
3.4.2 Machine Learning .....	29
<b>3.4.3 Classification and Classifiers</b> .....	29
<b>3.5 Classifiers used in the current study</b> .....	33

3.5.1 Support Vector Machines .....	33
3.5.2 Random Forest Classifier .....	34
3.5.3 Decision Tree classifier .....	35
3.5.4 Multi-Level Perceptron (MLP) – ANN .....	36
<b>4Data Analysis .....</b>	<b>40</b>
<b>4.1 Dataset .....</b>	<b>40</b>
<b>4.2 Pre-processing .....</b>	<b>41</b>
<b>4.3 Verification and Validation .....</b>	<b>45</b>
<b>4.4 Data Augmentation .....</b>	<b>47</b>
<b>5 Mode Development.....</b>	<b>48</b>
<b>5.1 Non-linear SVM .....</b>	<b>48</b>
<b>5.2 Random Forest Model.....</b>	<b>49</b>
<b>5.3 Decision Trees.....</b>	<b>51</b>
<b>5.4 Deep Learning - ANN Multi Level Perceptron .....</b>	<b>52</b>
<b>6 Results and Discussion.....</b>	<b>55</b>
<b>6.1 Major Findings .....</b>	<b>55</b>
<b>6.2 Conclusion.....</b>	<b>56</b>
<b>6.3 Recommendations.....</b>	<b>57</b>
<b>6.4 Limitations &amp; Future Scope.....</b>	<b>57</b>
<b>7 References .....</b>	<b>58</b>
<b>Appendix A: Research Project Plan.....</b>	<b>64</b>
<b>Appendix B: Completed Research ethics form.....</b>	<b>74</b>
<b>Appendix C: Publication Procedure Form .....</b>	<b>81</b>
<b>Appendix D: Dataset Source .....</b>	<b>82</b>
<b>Appendix E: Source Code .....</b>	<b>83</b>

# 1. Introduction

The human brain, in the list of five vital organs that are essential for living, the human brain places at first place out and there is no active life without brain (Zawn Villines, 2020). The human brain is the controlling and commanding center for the central nervous system (Tanya Lewis, Ashley P. Taylor, 2021). It receives the commands from the organs that are responsible for senses and orders muscles to respond. It helps human beings (as other mammals) to be creative, adaptive, being social and makes complex human beings (Larry Abbott, 2020). Like other human organs, the brain is no exception from diseases and brain tumors toll 85-95% of the brain or spinal tumors (Cancer.Net Editorial Board, 2021). Human brain can be scanned for study and analysis with various methods/scans like EEG, fMRI, PET, MRI, CT or DTI depending on type interest (Helen Phillips, 2006). MRI (Magnetic Resonance Images) are particularly good at capturing images of brain at any plane contrast soft tissue (R F Kilcoyne, M L Richardson, B A Porter, D O Olson, T K Greenlee, W Lanzer)

Registration of baseline pre-operative (treatment-naïve) and follow-up brain tumor MRI scans is challenging, yet a clinically important task for a multitude of reasons. Brain tissue shows heavy deformations induced by the apparent tumor (also known as mass effect) that following its resection are relaxed due to the relieving pressure from the resected tissue. Such deformations affect the whole brain (including the lateral ventricles) and are not limited to the vicinity of the tumour. This is particularly important as the relationship of the tumour to the lateral ventricles and the deformations to the rest of the brain tissue are important factors in prognosis and treatment planning. Further changes in the peritumoral edematous/infiltrated tissue, potential tumour recurrence, as well as treatment related changes, also affect the brain tissue elasticity. The resected tissue/tumor also relates to missing correspondences, and inconsistent intensity profiles between the follow up and the baseline pre-operative scans (Bhakti, Diana, Hamed, & Satrajit, 2021).

Taking all the above into consideration, finding spatial correspondences between two longitudinal scans of brain tumour patients, i.e., the registration between the baseline pre-operative and follow-up MRI scans, can advance mechanistic understanding for these tumors. Specifically, for tumour infiltration and potential recurrence, further contributing to the

generation of predictive modelling for related pathophysiological processes, but also in understanding biophysical dynamic and plasticity characteristic of brain tissues, as well as for neurosurgical planning (CBICA image processing portal.).

University of Pennsylvania is conducting every year brain tumor problem challenges and this year they are conducting their first Brain Tumor Sequence Registration (BRaTS-Reg) challenge (Bhakti Baheti et al., 2021), as part of the challenge, the dataset was provided for data science enthusiasts upon registration and request. In the current work, both the pre-operative and follow up scans are combined without distinction so that machine treats all the 280 MRI files as separate files during the training process. Prediction evaluation will use the separate follow up files (10 in number) from a different set provided as part of model validation

### 1.1 Project Aim

The current study aims on predicting the spatial correspondences of tumour using a dataset that contains pair of 3d MRI scans (preoperative and follow up) of 140 patients where ground truth is confirmed and landmarked by the domain experts. The 140 patient's dataset was released by the BRaTS challenge team to its participants in the current year 2022 (Bhakti Baheti et. al 2021) The focus of the work starts from collecting consented raw dataset (CBICA image processing portal.) along with corresponding landmarked ground truths to the pre-processing of the NIfTI format 3d MRI datasets into machine readable format for learning, labelling, development of supervised machine learning/deep learning models to the outcome publishing and conclusion.

### 1.2 Project Scope

The essential items in the scope of this project are conduct research on situation, complication and resolution mechanism, literature review, identify research gap to extract a problem area,

collect the dataset of MRI images, perform the exploratory data analysis, pre-process the dataset and develop a suitable supervised machine learning/ deep learning model that predicts the possible brain tumour registrations and their spatial coordinates. The scope extends to develop various possible supervised models within the given timeframe of the project, evaluating and analysing the outcome of each model extensively and concludes with the best model. As the collected dataset was provided has ground truth landmarks, the selection choices were considered as label prediction machine learning and deep learning supervised learning models [Refer the paper]. The scope also considers highlighting the key outcomes, limitations, future works that can be potentially contributed to enhance or extend the research further.

### 1.3 Research Question

How to predict possible tumour affected regions and their spatial coordinates of the brain from the 3D magnetic resonance images (MRI) using supervised learning methods of Artificial Intelligence?

### 1.4 Objectives

s

The following objectives were considered to achieve and obtain the expected outcome in the process of resolution of the chosen research question

- To explore and investigate extensively the classification algorithms for binary label
- To examine the attributes (features) to predict tumour regions from MRIs.
- To develop and train a machine learning model(s) using a suitable technique based on the examined and explored options using appropriate software tools.
- To identify the best performing model among all

### 1.5 Project Benefits

The MRI dataset that was provided to develop the prediction models consists of pre-operative and follow-up (post operative) data with the aim of predicting potential new follow-up spatial



coordinates of brain tumour, however the current study does not identify the distinction of pre-operative or post-operative but predicts on any given MRI NIfTI file the possible brain tumour effected region and its coordinates and the following benefits were identified

- A novel mechanism that quickly provides results to the domain experts to assist in their analysis
- Bridges the gap between existing models that were built on small data set as the current models were developed based on the domain expert provided ground truth landmarks
- Provides the system to the current novice users (Diagnostic Radiography final year students at SHU) to evaluate with their current tumour visualisation process
- Pre-processing techniques that were used in converting the NIfTI files to csv files will be helpful for the future work
- The dataset is highly useful with BRaTS citation in future study as actual collection of the dataset has very stringent process to be followed

## 1.6 Project Deliverables

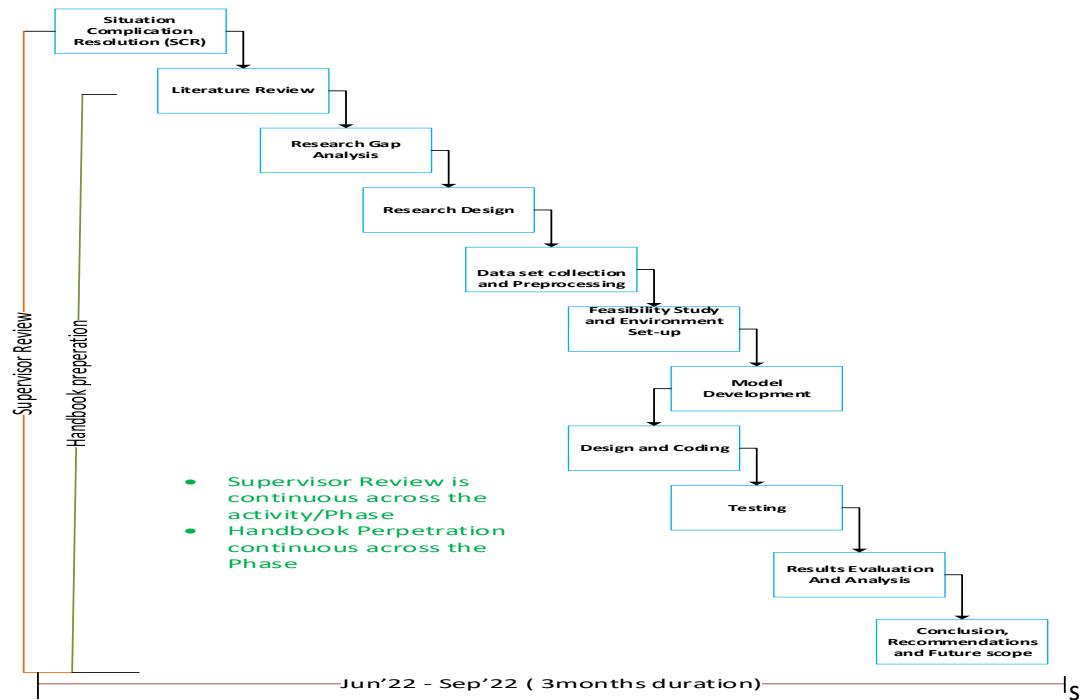
- A machine learning model that predicts brain tumor correspondences and their spatial coordinates from the Brain MR Images of NIfTI format

## 1.7 Summary of the Research Study

The initial proposal of the research plan was provided in the Appendix A, and the Research Ethics form in Appendix B. The research was kicked off with the SCR (situation complication resolution) framework that helped to identify an area to conduct the research. The SCR frame had helped to begin the literature review and narrow down the area of interest. The literature review in turn led to analyse the research gaps in the chosen area and helped to formulate the research question and its objectives. Upon the literature review complete, the next step was to collect the dataset with BRaTS Team consent. The research methodology phase had provided a blueprint and design to develop the supervised machine learning /deep learning models. Once

the design was completed, it was next step to pre-process the dataset (NIfTI to CSV format) and develop and train the learning models with appropriate train and test sets. The testing and evaluation/validation the model followed by the model development. The accuracy of developed models was analysed to in order make conclusions, recommendations and define the future work by highlining the limitations of current study/work/research. The project flow was explained in the Figure 1.1 that depicts various phases and timelines of the project

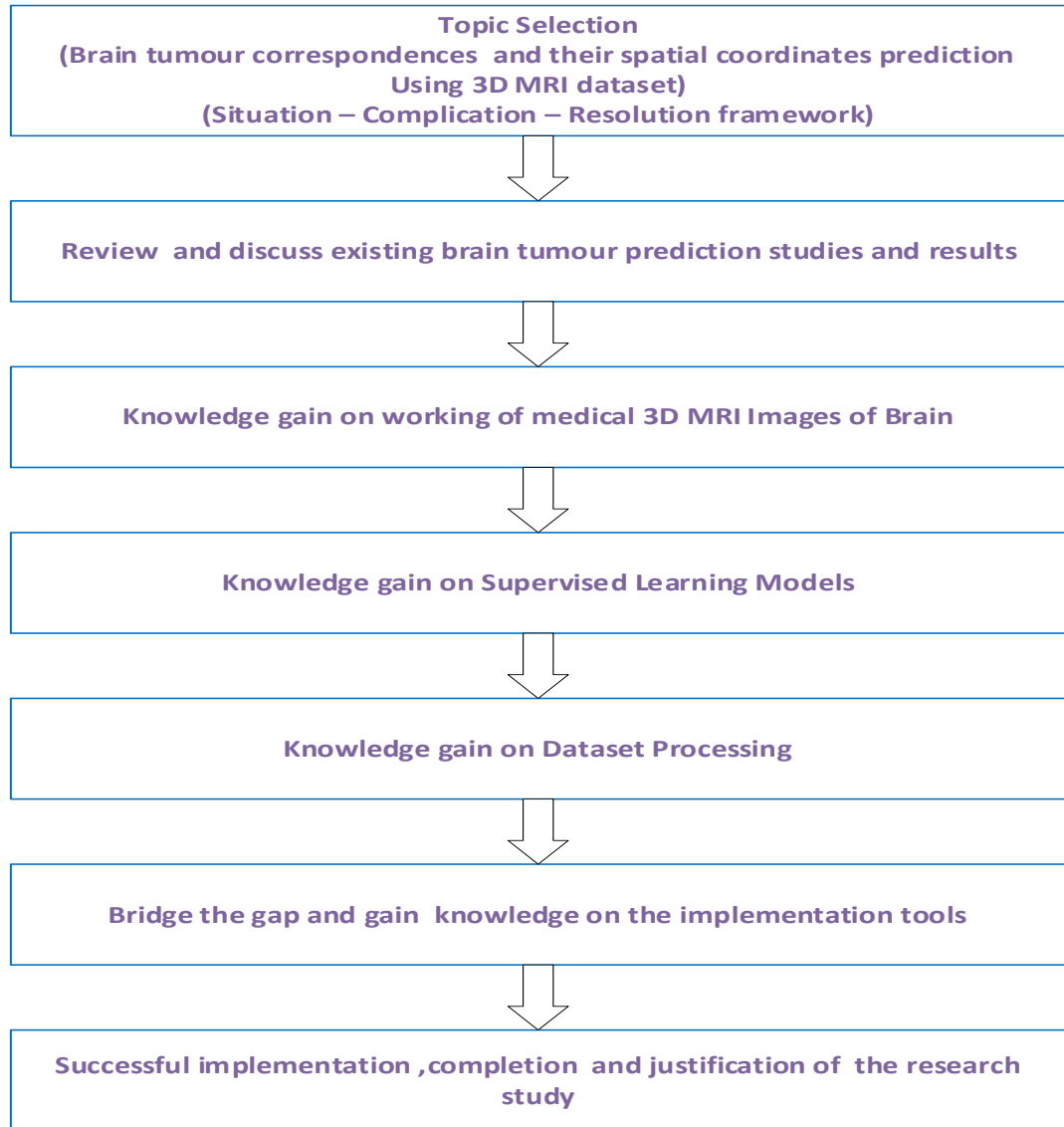
*Figure 1.1: project Overview of various phases involved*



## 2 Literature Study and Review

Literature study and review is an important and crucial phase of any research activity as it helps the gain the required knowledge and provides insight into the existing studies. It bridges the knowledge gap and produces knowledge that is required for the successful completion of chosen study (Hannah Snyder, 2019). The key areas of literature review flow of this project can be seen in the Figure 2.1.

*Figure 2.1: Literature review flow*



## 2.1 Discussion on existing studies on brain tumour detection

The section discusses extensively on the various selected studies conducted on the brain tumour region or segmentation prediction along with the BRaTS Challenge 2022 on spatial coordinates or correspondences prediction.

- A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network (Francisco Javier Díaz-Pernas et al., 2021)
- Brain tumor detection from MRI images using deep learning techniques (P Gokila Brindha et al., 2021)
- Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM (Nilesh Bhaskarrao Bahadure et al., 2017)
- A Novel Deep Learning Method for Recognition and Classification of Brain Tumors from MRI Images (Momina Masood et al., 2021)
- The Brain Tumor Sequence Registration Challenge: Establishing Correspondence between Pre-Operative and Follow-up MRI scans of diffuse glioma patients (Bhakti Baheti et al., 2021)

#### **2.1.1 A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network**

The article talks on prediction of segmented region and classification based on a deep learning model developed using the T1-CE modality MRI Images that were collected as part of study.

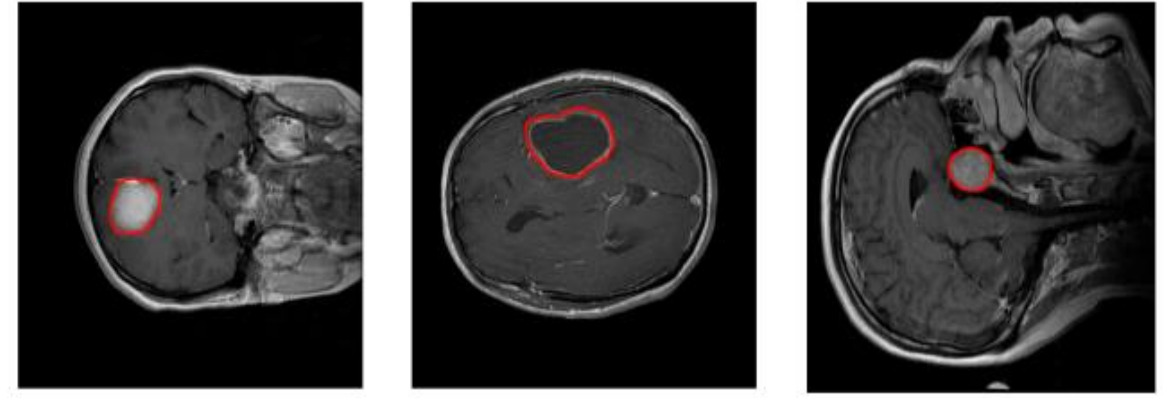
As per (Francisco Javier Díaz-Pernas et al., 2021), various techniques of medical imaging were highlighted in the paper based on the location, shape, type and size of tumours. The important imaging techniques mentioned were CT (Computed Tomography), SPECT (Single Photon Emission Computed Tomography), PET (Positron Emission Tomography), MRS (Magnetic Resonance Spectroscopy) and MRI (Magnetic Resonance Images) and reason for MRI images selection was highlighted as the distinctive exhibition of characteristics.

The (Francisco Javier Díaz-Pernas et al., 2021) also collected the four MRI modalities used in diagnosis called T1 (T1 weighted), T2 (T2 weighted), T1-CE (contract-enhanced T1 weighted)

and FLAIR (Fluid Attenuated Inversion Recovery) . Out of all the four modalities, T1-CE was considered best suited for the tumor region prediction and classification

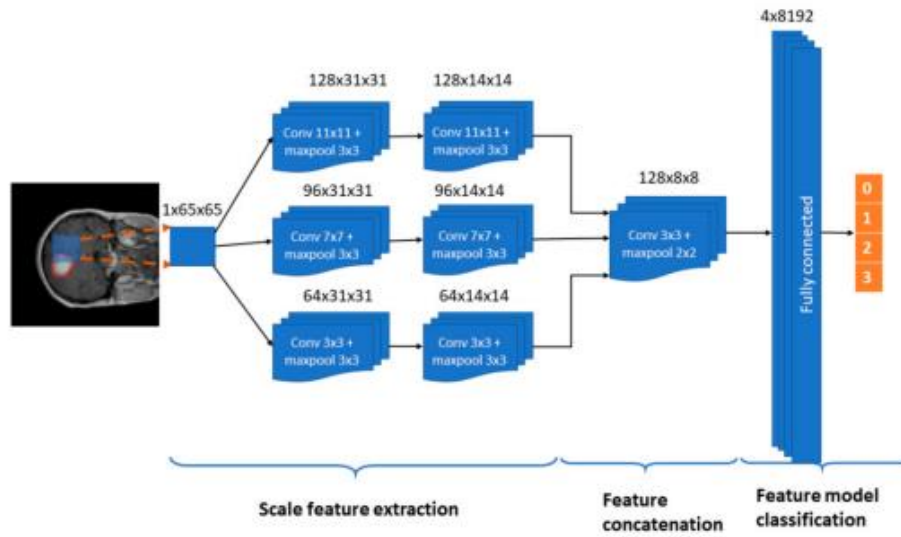
The classifications of tumours that were considered in this deep learning approach are pituitary, meningiomas, gliomas. The common views of that can show tumor region were highlighted in the picture Figure 2.2

*Figure 2.2: brain tumor regions*



The architecture that was proposed predicts and classifies four regions viz., healthy region (0), meningioma (1), glioma (2) , pituitary (3) can be seen in the Figure 2.3

*Figure 2.3: architecture of multi scale CNN*



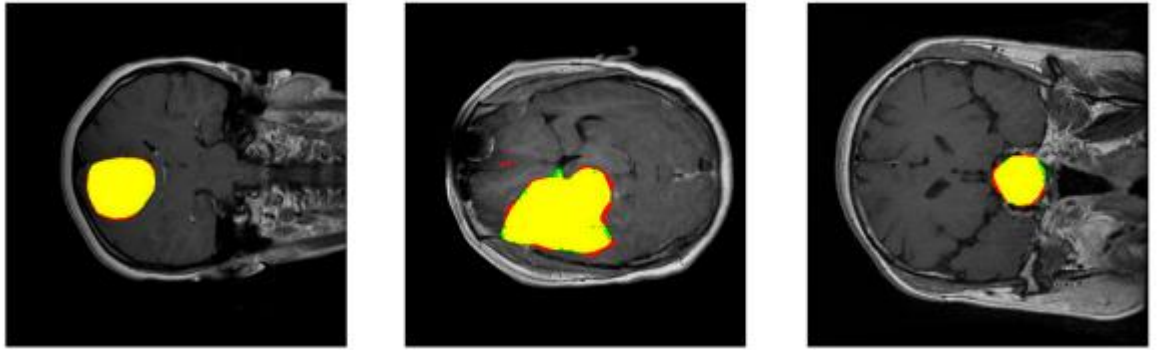
The accuracy using multiscale CNN achieved were 0.973 when the authors (Francisco Javier Díaz-Pernas et al., 2021), compared with previous accuracies by a different study was 0.912. The confusion matrix of the classification was given in the Table 2.1

*Table 2.1: confusion matrix of multi scale CNN*

True\Predicted	Meningioma	Glioma	Pituitary Tumor	Sensitivity	Non-Classified
Meningioma	659	4	3	0.93	42
Glioma	7	1414	1	0.99	4
Pituitary tumor	1	3	911	0.98	15
Tumor classification accuracy			0.973		

Predicted classifications were compared with ground truth regions. The red colour highlighted regions are predicted regions, whereas the green colour regions are tumours of ground truth and yellow coloured are intersections are shown in Figure 2.4

*Figure 2.4: brain tumor regions predicted vs ground truth*

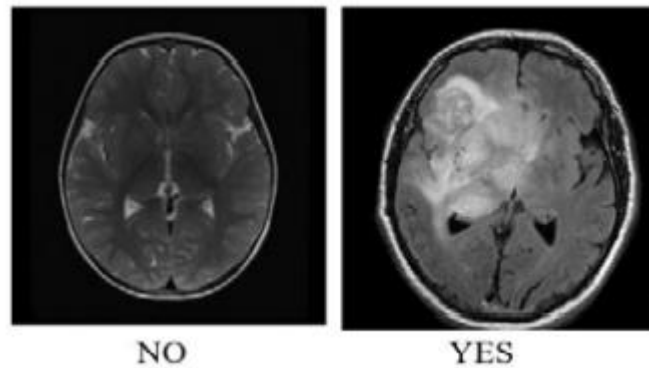


Review of this paper helped basic idea on the brain tumor dataset processing, unleashed the CNN methods of full image processing. The paper also helped to get a confirmation on the modality of scans that can be used as a candidate for data pre-processing

### 2.1.2 Brain tumor detection from MRI images using deep learning techniques

In the paper, (P Gokila Brindha<sup>1</sup>, M Kaviraj , P Manivasakam , P Prasanth, 2021) uses set of MRI images that are with tumor and with no tumor to predict the brain tumor with the application of Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) . Though there was no mention of which type or classification of brain tumour that authors were predicting, it has been highlighted that there were 2065 images used (out of 2065 images, there are 1085 images that has tumour and 980 images that are of no-tumor images). There was a mention of source image sizes shapes that are heterogenous (630X630,225X225) and were converted into the shape of 256x256. An example of the MRI images with tumour and no-tumor were shown in the Fig 2.5

*Figure 2.5: MR Images of no tumor and tumor*

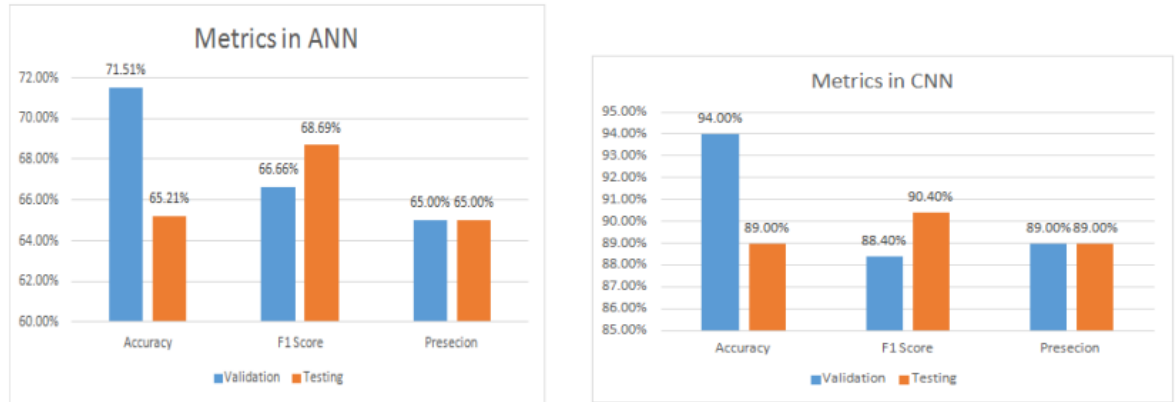


Details of CNN model were highlighted in the article whereas ANN details not very clear. The following technical details of CNN model, that helps to utilise in the current project were captured and presented

- The CNN Model: Sequential
- Input Image shape: 256x256
- Activation function: ReLU
- Max Pooling: 2 x 2 window
- Drop-out function: 20%
- No. of Epochs: 200
- Method: trial and error
- Implemented language: Python
- Lab executed: Google colab

Validation accuracies of both the ANN and CNN models were captured and based on the experimental results obtained it was concluded that CNN model was performed well with the accuracy score of 89% whereas ANN accuracy score shown as 65.21%. Complete accuracy scores comparison was shown in the Figure 2.6

*Figure 2.6: Metrics ANN vs CNN*



Study of this paper not only helped on the labelling of the images, but also provided insights into the Artificial Neural Network techniques as the current study has more relevancy for the ANN. ANN is one of the choices for the domain expert curated MRI medical images

### **2.1.3 Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM**

The paper on the Image analysis and feature extraction using BWT and SVM by (Nilesh Bhaskarrao Bahadure et al., 2017) aims to extract the features such as abnormal tissues (possible tumours) and normal tissues (White Matter, Gray Matter , CerebroSpinal Fluid) from the tumor segmented region using the Support Vector Machine(SVM).

The authors (Nilesh Bhaskarrao Bahadure et al., 2017) have investigated the BWT (Berkeley Wavelet Transformaton) method of tumor segmentation of brain using SVM.



Various classifications and grades of brain tumor were presented in the below table and the categorisation and classification of the brain tumours mention in the paper were highly useful for the current study and helped the brain tumor categories understanding shown in table 2.2

*Table 2.2: Brain tumor types and classification grades*

<b>Brain Tumor Type</b>	<b>Classification</b>	<b>Stage</b>	<b>Grade</b>	<b>Grade Level</b>
The gliomas	Benign tumours	possess a slow growth	I	low-grade
Meningiomas			II	
Glioblastoma	Malignant tumours	possess a rapid growth	III	high-grade
Astrocytomas			IV	

Though there was no mention of which brain tumor type images were used for the study, there was a mention of the 3 out of 4 available modalities of MRI Images. The modalities used for the study were weighted T1, weighted T2 and weighted FLAIR images. The total images that were used are 135, are from 15 channels with each channel has 9 slices or images ( per patient 9 images , so the total images covers 15 patients) . The thickness of each slice was mentioned as 1mm with voxel (volumetric pixel) size as 0.78 mm x 0.78 mm x 0.5mm. The size of MRI images was 512 x 512 Pixel and were converted into the gray scale images.

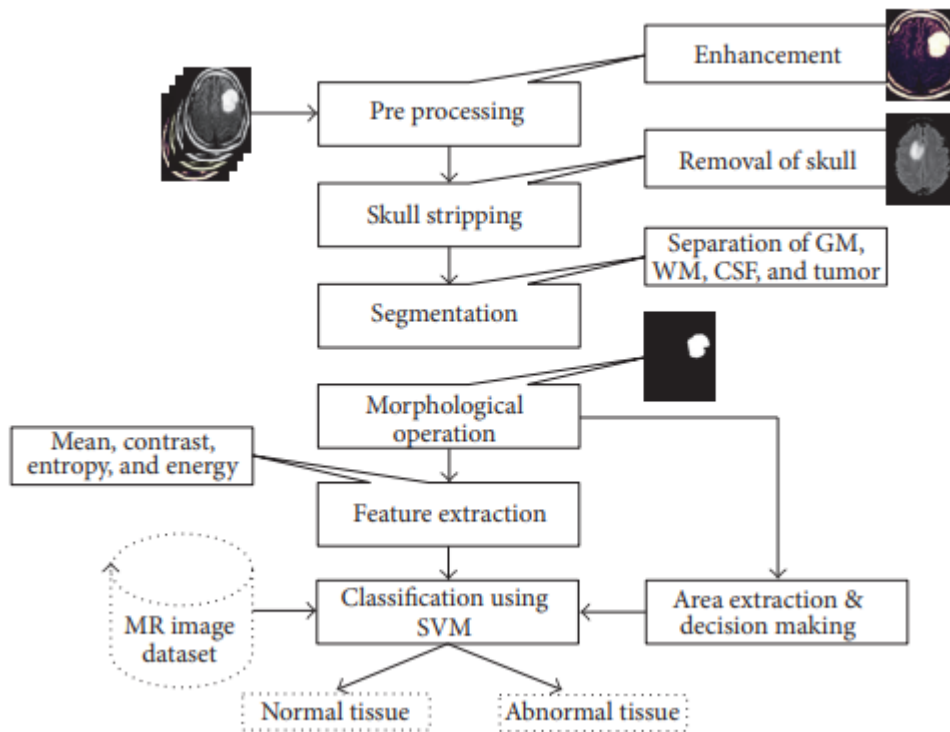
There was detailed mention of background and related work on the brain tumor image analysis viz.,

- Hybrid type approach (combines the Fuzzy C Means ,seed region , Jaccard algorithm) on tumor segmentation of brain with accuracies S 90% at noise 3% and 6%
- Neural Network based classifier for the segmentation of white matter , gray matter CSF and tumour regions with 83% accuracy . there was no mention of the specific classifier that was used
- Automatic tumor classification using FFT and MRMR with 98.9% accuracy
- Tissue segmentation algorithms using wavelets and neural networks

- Brain tumor classification and segmentation using RBF (radial basis function) based SVM ,PCA (Principal Component Analysis) with 94% accuracy
- Brain tumour classification using ANN texture-primitives features with 100% accuracy
- Brain MR Images bias field estimation and segmentation using jaccard similarity index claiming the accuracy from 83% - 95%. Localisation of spatial information with fuzzy clustering was used to segment WM,GM and CSF

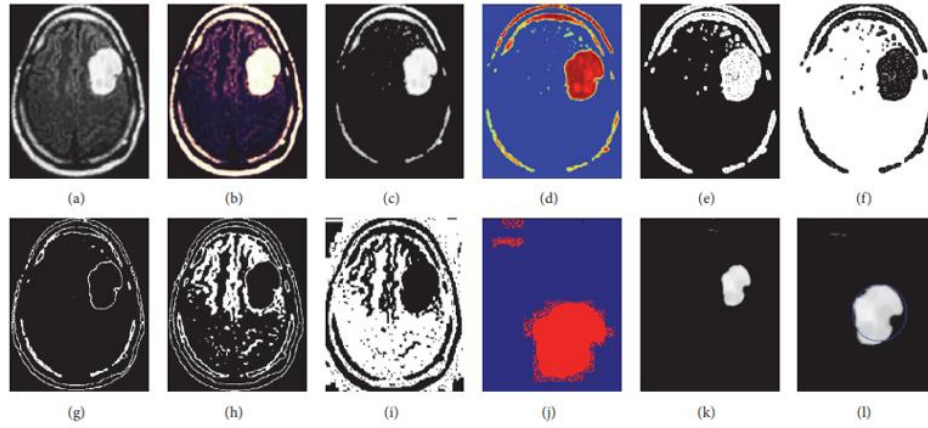
The Pre-processed Images were applied with morphological operation to get the segmented region and then SVM was applied to identify the abnormal tissue or normal tissue. The various steps involved in the applied algorithm were outlined in the below diagram (Figure 2.7)

Figure 2.7: Applied algorithm flow



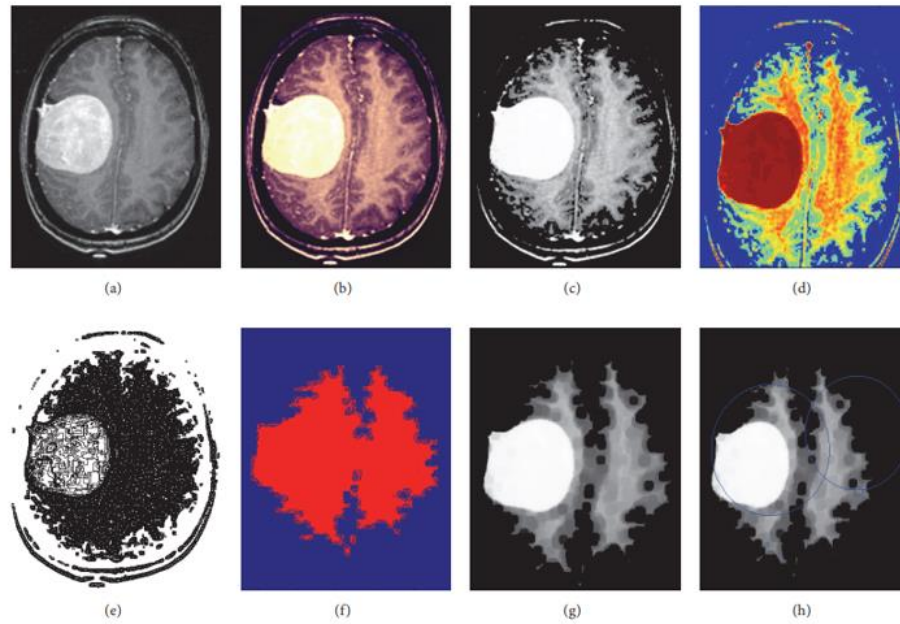
The BWT was used to extract the area of the brain tumor region and the classification was applied to identify features such as normal or abnormal issues.

Figure 2.8: tumor area extracted MR Image



Segmented and area extracted result of brain MR image. (a) Original image. (b) Enhanced image. (c) Skull-stripped image. (d) Wavelet transpose image. (e) Intense segmented image. (f) Inverse intense image. (g) Gray matter. (h) White matter. (i) CSF. (j) Dice overlap image. (k) Eroded image. (l) Area extracted image.

Figure 2.9: Experimental results of MR Image2



Experimental results of image 2. (a) Original image. (b) Enhanced image. (c) Skull-stripped image. (d) Wavelet decompose image. (e) Intense segmented image. (f) Dice overlap image. (g) Tumor region. (h) Area extracted tumor region.

The following classification models were developed with feature extraction and without feature extraction and selected the best working model as SVM with accuracy scores of 90.54 without

feature extraction and 96.51 with feature extraction as per the (Nilesh Bhaskarrao Bahadure et al., 2017)

- Adaptive Fuzzy Inference System
- Back Propagation
- K Nearest Neighbours (K-NN)
- Supporting Vector Machine (SVM)

The accuracies obtained from the applied classification models in the study were tabulated in the Figure 2.10.

*Figure 2.10: accuracies of the applied classification models*

Classifiers	Accuracy (%) without feature extraction	Accuracy (%) with feature extraction
ANFIS	86.14	90.04
Back Propagation	80.29	85.57
SVM (proposed classifier)	90.54	96.51
K-NN	84.55	87.06

The Image analysis and feature extraction paper literature review has helped the current study in terms of domain knowledge on the brain tumor classification, dataset properties and application of classifiers. The literature review of this paper also helped with brain tumor segmentation related studies as the paper captures enormous studies conducted on brain tumor segmentation

Literature review of this paper also helped to understand better the SVM model, which is one of the selection choices for the model development.

#### 2.1.4 A Novel Deep Learning Method for Recognition and Classification of Brain Tumors from MRI Images

The authors (Momina Masood et al., 2021) have developed a novel deep learning model to detect and classify the brain tumours in the form of segmentation using the Mask R CNN (

Mask Region based Convolutional Neural Network) with the architecture densenet-41 backbone with the 96.3% segmentation accuracy and 98.34% classification accuracy

As per the (Momina Masood et al., 2021), the selection models for the detection of brain tumour are two types

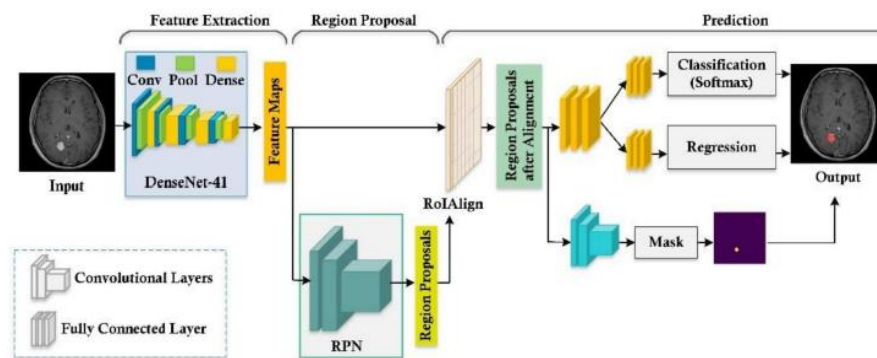
1. Machine learning models: These models need the features extracted from the MR Images and need to medical expert intervention to craft the images. The following are key models that can be applied on features crafted dataset

- I. SVM s
- II. Fuzzy Clustering means
- III. PCA (Principal Component Analysis)
- IV. Decision Trees
- V. Conditional Random Forest

The machine learning models are likely to produce errors unless features are not well crafted by the domain experts and in the case of large dataset as well.

2. Deep learning models: These models can extract features automatically and offers selection. CNN is widely used and most popular deep learning model for Image processing that includes MR Images due to their weighted pooling nature (Momina Masood et al., 2021), The architecture of Mask RCNN network is presented in the Figure 2.11

*Figure 2.11: Architecture of Mask RCNN*

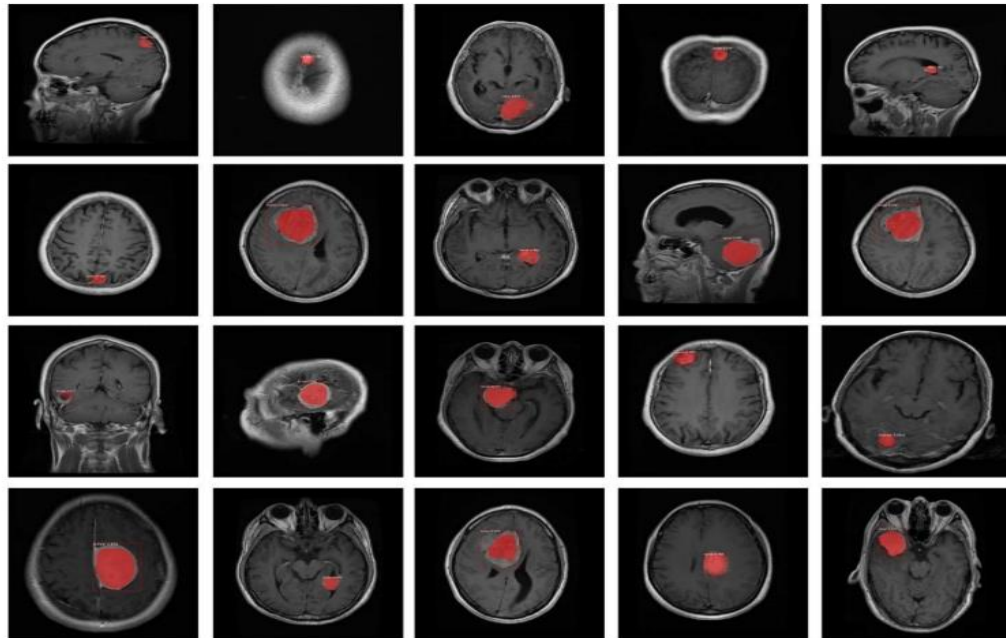


The dataset: There are total 3064 MR Images of brain that belongs to 233 patients were used. Of all the images, 708 are of meningioma class, 930 are pituitary and 1426 are glioma. A second dataset that consisted of 253 images (MRI) were also used for the model development. Both the datasets were available publicly and the size of the images are 512 x 512 pixels each. The MR Images are of T1-weighted modality , the same modality is used in the current study.

As per the (Momina Masood et al., 2021), Mask RCNN was developed using DenseNet- 41 and ResNet-50 for accurate and automated brain tumor segmentation and DenseNet- 41 performed better compared to ResNet-50

The results obtained as part of segmentation are shown in the Figure 2.12.

*Figure 2.12: brain tumor segmentation results*



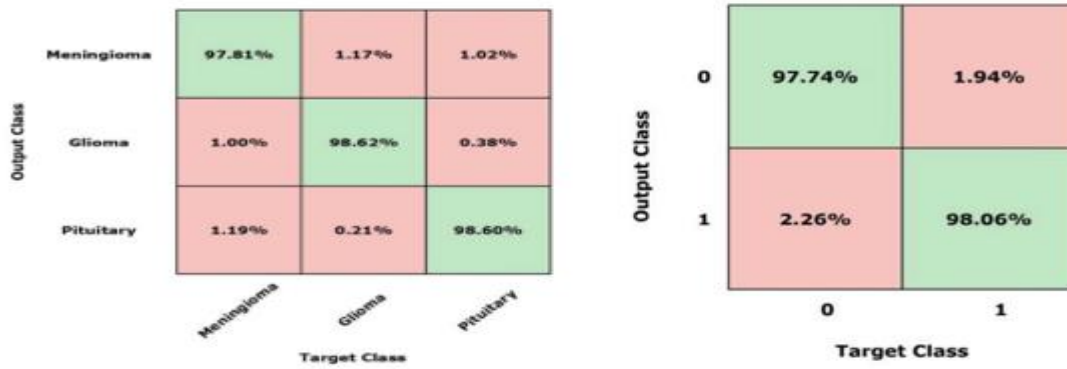
Accuracy scores are tabulated in the Figure 2.13

Figure 2.13: Accuracy scores of various methods applied

Method	Evaluation Metrics				
	Accuracy	mAP	Dice	Sensitivity	Time(s)
RCNN [51]	0.920	0.910	0.870	0.950	0.47
Faster RCNN [56]	0.940	0.940	0.910	0.940	0.25
Proposed (Resnet-50)	0.959	0.946	0.955	0.953	0.20
Proposed (Densenet-41)	0.963	0.949	0.959	0.953	0.20

Confusion Matrices of tumor classification and brain tumor prediction are presented in the Figure 2.14

Figure 2.14: confusion matrix of predicted brain tumor classes vs ground truth classes



The usage of supervised machine learning models for the ground truth-based dataset for the current study was identified from the literature review of this paper

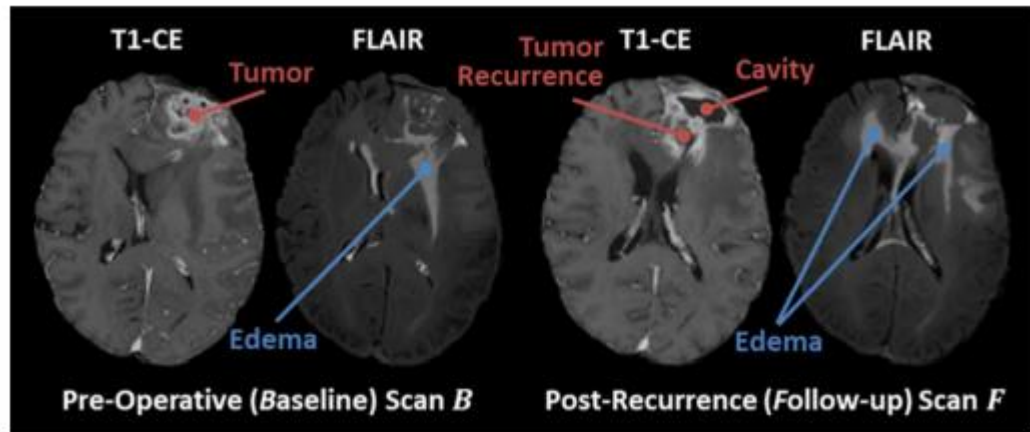
### 2.1.5 The Brain Tumor Sequence Registration Challenge: Establishing Correspondence between Pre-Operative and Follow-up MRI scans of diffuse glioma patients

The paper aim is to provide the details of a data science challenge on BRaTS-Reg (The Brain Tumor Sequence Registration) to the participants. The authors (Bhakti Baheti et al.,2021) have highlighted the need of this data science challenge and bests outcomes will be published in public domain as a benchmark for the deformable registration algorithms.



The objective of the challenge is to address the image registration problems in the Brain MR Images. Image registration is a process of establishment of correspondences between multiple images with the display of targeted object (Barbara Zitova, 2018). The Isocitrate Dehydrogenase (IDH)-wildtype glioblastoma tumor is the most serious and malignant brain tumor deforms the peripheral tissues impacts central nervous system as per World Health Organization (Bhakti Baheti et al.,2021). Finding registrations of image in the pre-operative and post-operative MR images is yet challenging due to deformation nature of the tumour mass effect. The pre-operative MRI scans are also called the baseline scans and the post operative scans are called the follow-up scans. An example of tumour and its recurrences in a baseline and follow-up scans are explained in the Fig. with the help of two modalities T1-Contrast Enhanced(T1-CE) and FLAIR (Fluid Attenuated Inversion Recovery). The T1-CE is considered as the best modality for tumor region analysis whereas the FLAIR exhibits best for the Edema (fluid) in the Figure 2.15

*Figure 2.15: Two modalities t1-ce and FLAIR with tumor and edema*

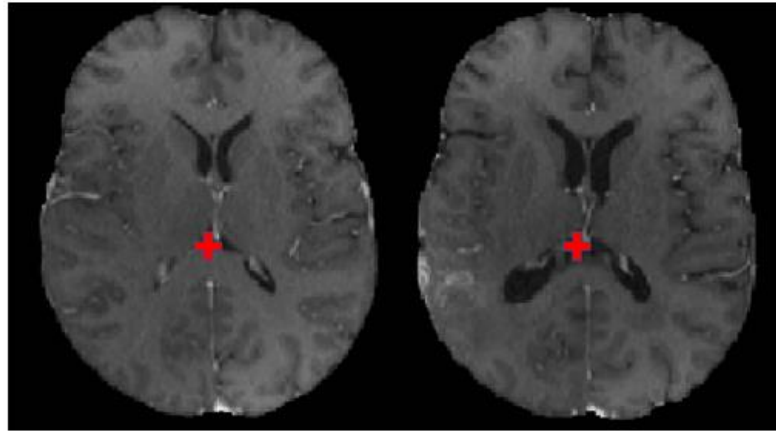


The dataset used in the challenge consists of 250 patients, with a pair of MRI scans (baselined and the follow-up) for each patient. Each scan will have 4 modalities to highlight the tumor region, edema etc) as it is difficult get the complete details of brain matter with one modality. The dataset is mpMRI (multi-parametric MRI contains 4 modalities viz., T1, T1-CE, T2 and FLAIR). The baseline and follow-up scans are captured at 2 or separate time points and the



time gap between to scans ranges from 27 days to 37 months. The mpMRI dataset is multi-institutional, patient dispersion of unique demographics. Each scan in the dataset was provided with landmark coordinates curated and hand-crafted by domain experts after thorough testing in the form of csv format along with mpMRI in NIfTI format (Neuroimaging Informatics Technology Initiative) with the resolution of  $1 \text{ mm}^3$  (isotropic). An example of curated landmark in the pair of scans of T1-CE modality can be seen in the Figure 2.16

*Figure 2.16: base(pre-operative) scan and follow-up(post-operative) scans with tumor highlighted*

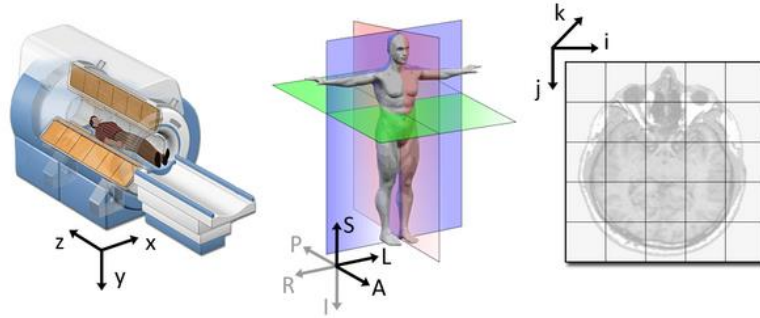


Current study research questions, procurement of dataset and pre-processed data validation were all based this paper. Future scope of current study will also be influenced by this paper as objective of this paper is about BRaTS challenge and participation details

## 2.2 Anatomical co-ordinate system

As the dataset provided was in the NIfTI format for the current study, it was necessary to gain knowledge on the coordinate system that was employed on the NIfTI images. The most important coordinate system used in the imaging techniques in the medical field is anatomical coordinate system or patient coordinate system (Coordinate systems.) The building blocks of anatomical coordinate system are as shown in the Figure 2.17)

Figure 2.17: Anatomical coordinate system



- Superior and Inferior. The plane axial plane separates towards Head (Superior) and Feet (Inferior) respectively
- Anterior and Posterior. Coronal Plane which is perpendicular to the ground and separates the front (Anterior) and back (Posterior)
- Left and Right. The sagittal plane from the Left to Right

The two anatomical coordinate systems that are widely used are

- LPS (Left, Posterior, Superior)
- RAS (Right, Anterior, Superior)

General conversion formula from anatomical coordinate system to world image coordinate system is shown in the Formula no. 1.

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & A_{13} & t_1 \\ A_{21} & A_{22} & A_{23} & t_2 \\ A_{31} & A_{32} & A_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} i \\ j \\ k \\ 1 \end{pmatrix} \quad (1)$$

Formula no. 1 (Coordinate systems.)

Where  $(i\ j\ k)'$  is image coordinate vector ,  $(x_1\ x_2\ x_3)'$  is anatomical vector ,  $(t_1\ t_2\ t_3)'$  is a translation vector of  $3 \times 1$  and A is a  $3 \times 3$  matrix contains information about space directions and scaling of axis

## 2.3 Literature Review Summary

Literature review was conducted in order to identify the research gap initially that has provided an opportunity to refine and frame the current research problem question. Upon the identification of research problem, detailed literature was reviewed and captured key relevant areas in terms of the problem domain (brain tumor prediction), and various implementation methods that are available in current public domain. The decisions that were made in data set processing, model development after extensive review of the various journal articles, publications, books and other sources. For example, the choice of supervised machine learning models was extracted after the review of A novel deep learning method for classification Brain Tumors (Momina Masood et al., 2021). The decision to use T1-CE modality from the dataset is obtained after the review of BRaTS challenge paper (Bhakti Baheti et al., 2021). There were interactions made with the BRaTS team to get confirmation on the landmark coordinates processing and conversion from the anatomical coordinate system to the world coordinate system. Various supervised learning machine learning models that are suitable for the current dataset which is unique in terms of volume and formats (NIfTI format MR Images along with csv files that contain coordinates of ground truth landmark)

## 3 Research Design

### 3.1 Research Ethics

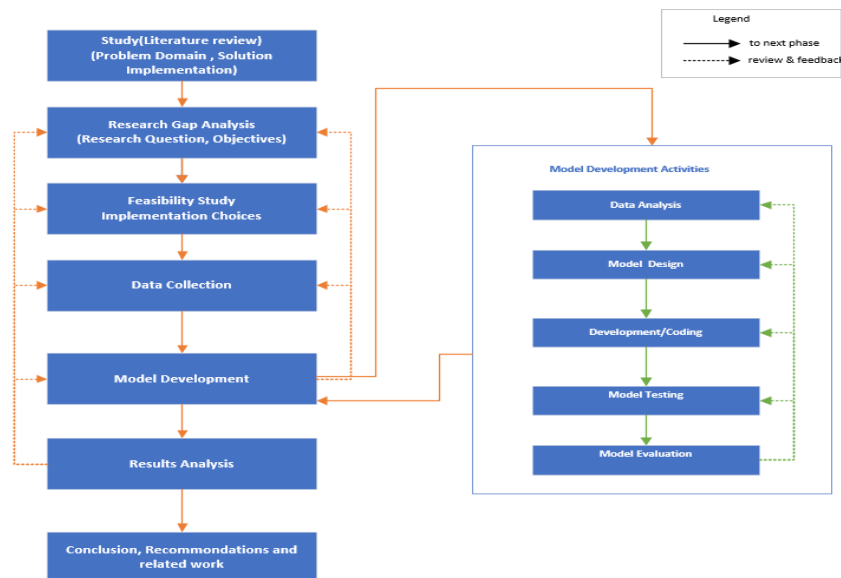
Sheffield Hallam University scrutinize all the research activities that are conducted as part of academics. To ensure the standards of ethics are maintained on high pedestal, there is a ethical

policy that is under consistent review to ensure high standards are met (Sheffield Hallam University, 2020) . The dataset involved in the current study was collected from BRaTS team through registration and approval from their website (Bhakti Baheti, et. al. 2022). Throughout the current research study, it was ensured Sheffield Hallam University standards were met and adhered to the University ethical policy. Research ethics form was completed with duly signed by the student and the supervisor was enclosed in the section APPENDIX B

### 3.2 Research framework

A research framework is a systematic procedure or structure that helps to classify and organise sequence of steps to follow on a complex problem or study. It provides a structured approach to resolve a problem, though it cannot provide precise answers to the practices, tools to be used in resolving the problem but it allows space to add various options and choices of practices, tools are available (thinkinsights.net, 2020). The research framework or various phases that were embraced during the research study are highlighted in the flow chart Figure 3.1

*Figure 3.1: Various steps involved (research framework) in the current study*



Literature Review was conducted in order to get familiarity on the domain area under which research study can be performed on. As part this phase, after various domain areas exploration

it was decided to conduct study on image processing. With various options of current studies on image processing, the next step was to identify the area of study

The next phase was research gap analysis, that helped to identify the research question and the objectives. As part of the phase, various ideas (facial fake expressions detection, deepfakes analysis, fake image prediction, image segmentation etc) were brainstormed and identified research gaps in the current studies. Based on the availability of datasets on image segmentation, it was considered brain tumor prediction as the research study. The research gap analysis phase has provided the question and objectives that lead to explore on uniqueness of the study and implementation choices

The next phase after research gap analysis was to study the feasibility and implementation choices and data collection. This phase has addressed the question of problem uniqueness, availability of dataset and the software tool (programming language) that can be potentially used for the develop the solution for the selected research study. The BRaTS Challenge Team (Bhakti, Diana, Hamed, & Satrajit, 2021) had provided a dataset upon registration and approval a highly expert curated NIfTI Images for brain tumor registration correspondences and spatial coordinates prediction problem. The programming language was selected as Python as this language was learned as part of first semester of the course and with the fact the language has various advances due to its open-source nature (more details of the language were provided in the research plan Appendix B)

The phase Model development is the key activity in the implementation of the chosen research study. The phase has various sub phases are activities like data analysis and pre-processing, model design for the selection of available and best suitable prediction models, development/coding activity, testing of the model in terms of correctness, completeness and quality of the solution that was proposed to build. Once the testing activity completes, the developed models were evaluated to obtain experimental results like the accuracy scores, confusion matrix and ROC curves for each of the developed models

Once the various models developed, the next steps or phases were to analysis, compare and reconcile all the models and to select the best working model to consider as the solution for the research study.

The final phase of the research framework is to provide a conclusion and recommendations based on the observations; results achieved. It was also highlighted to provide the limitations of the current study so that it helps to conduct future work

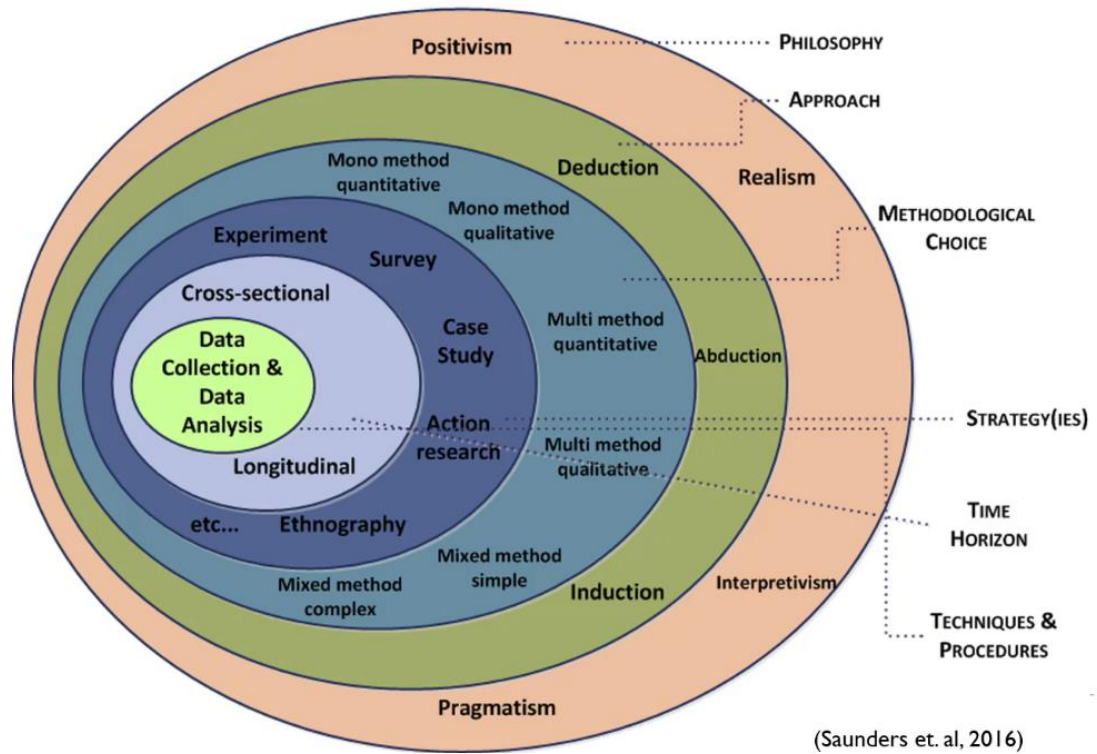
While following the research framework, based on the outcome of each phase, previous phases were revisited on need basis. E.g., while dataset pre-processing, to ensure the pre-processing is correct, I had to reach out to the dataset team (Bhakti Baheti, et. al. 2022) to get a confirmation on the correctness of the data pre-processing activity

### 3.3 Research methodology

Research methodology is a relationship between the current research study and existing studies in the same selected problem area. The methods followed while approaching the resolution of the research question like data gathering, data pre-processing, selection choices of algorithms, model development and results analysis. The design and methods must look robust with error free execution and must provide the details of system requirements and configuration required, execution procedure, user manual with clear exhibition of the various variables that influence the data and outcome. Methodology must provide clear justifications to each, and every decision made in every phase of answering the research question. (oxbridgeessays.com, 2021)

The current research has followed the research onion model to identify the various stages that has been passed through during the resolution of research question. The research onion diagram was explained in the Fig., . (Mark Saunders, Philip Lewis, & Adrian Thornhill, 2009) that offers various philosophies to follow, various approaches to choose, methodologies to select, strategies that be adopted and data collection and analysis processes that can be opted relevantly

Figure 3.2: research onion layers



The current study has systematically identified one suitable method from the research onion layers and explained in detail.

**Philosophy:** A research philosophy is considered as set of assumptions or beliefs made depending on the nature of the problem. The justification to the philosophy is made based on how the research question embraced during the resolution of the problem (ukessays.com, 2021)

Positivism was considered the suitable philosophy for the current research problem as there are studies exist independent of the current study

**Approach:** The second outer layer in the research onion model is the research approach. The approach will communicate to the data collection and analysis about the decisions that are made (David Phair (PhD) et.al. 2021)

Deductive approach was selected for the current study as this approach was suitable and opted for the initial point is well existed and for positivism philosophies. This approach is also best suited because of the study results can be examined with the already existing research studies

Methodological choice: This layer in the research onion model offers various quantitative and qualitative techniques that can be of mono, multi or mixed type methods.

Multi method quantitative approaches were considered for the current study as there were multiple models were developed and analysed with accuracy and selected the best accurate model

Strategy(ies): The research strategies can be of experimental, surveys, or of case study, action research or ethnography.

Experimental Strategy was automatically considered as the research strategy for the current study as the results obtained will be validated against an existing ground truth result. The relationship of the obtained results will be compared and validated against the existing results

Time Horizon: It is the layer which decided the time frame or time window within which the chose study is to be completed. There are two types of time horizons viz., cross-sectional time horizon and longitudinal time horizon

Cross sectional time horizon suits the current study is the duration of the dissertation study is 3 months.

Techniques & Procedures: The core layer of the onion research model is the techniques and procedures. The layer helps to find various techniques for the collection of data and analysis of data. This layer allows to practically get into actual activities like data collection, design, development, testing and evaluation of the actual software

The classification models considered to address the research question are explained in the next section in systematic way



### 3.4 Machine Learning

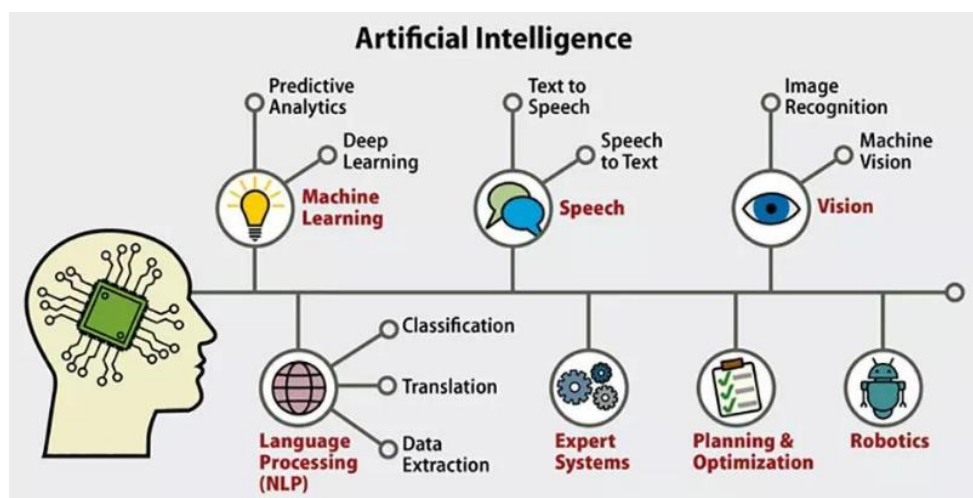
Artificial Intelligence and its applications are very crucial for the current study as the research problem was answered using the supervised learning of machine learned area. The section highlights and unleashes at high level the AI discipline to the detailed discussion on actual classification models that were applied on the current research question

#### 3.4.1 Artificial Intelligence

Artificial Intelligence (AI) is the computing systems capability that imitates learning, problem solving, and thinks cognitively just like human beings. (Artificial intelligence (AI) vs machine learning (ML).) Various disciplines or applications that fall under Artificial Intelligence are (Pradeep Kumar, 2021)

- Robotics
- Expert Systems
- Machine Learning
- Neural Network
- Fuzzy Logic
- NLP (Natural Language Processing)

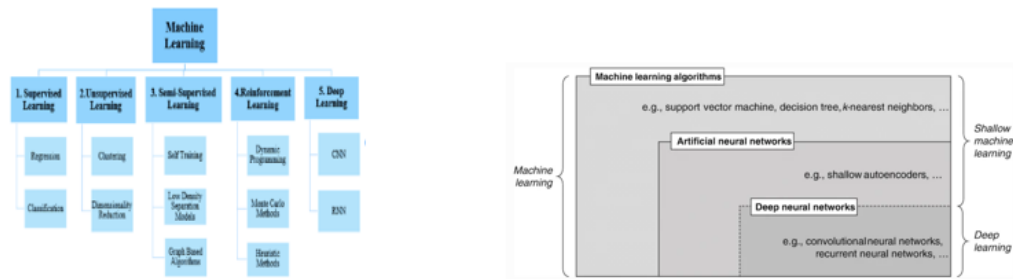
*Figure 3.3: Types of artificial intelligence*



### 3.4.2 Machine Learning

Machine Learning is one of the applications of Artificial Intelligence. The process of the applying the mathematical models of data and make the computing system to learn without direct instruction. This process makes the computing system to continue learning by itself, by experience (Artificial intelligence (AI) vs machine learning (ML).). Machine learning further high level divisions, types and applications are briefly highlighted in the Figure 3.4

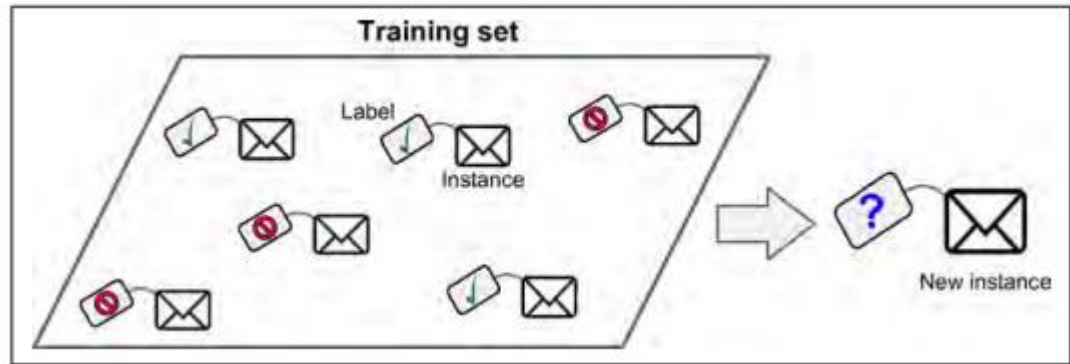
Figure 3.4: Machine learning types



### 3.4.3 Classification and Classifiers

Classification is one of the most popular prediction tasks of supervised learning type of machine learning. Classification is usually conducted on the labelled dataset that can be of binary class (Yes/No or spam/ham or Good/Bad) or mutli-class/multi-label (Planet species etc ) (Jason Brownlee, 2020) . A typical example of classification problem is explained in the Figure 3.5

Figure 3.5: Classification Problem



The example visualises a dataset that consists of genuine and spam emails that are labelled as spam or ham. The developed learning machine will be able to predict the new instance of email whether it is genuine or spam email. Current study is leveraging the supervised learning classification techniques also called the classifiers to predict the brain tumor class istumor Yes or No.

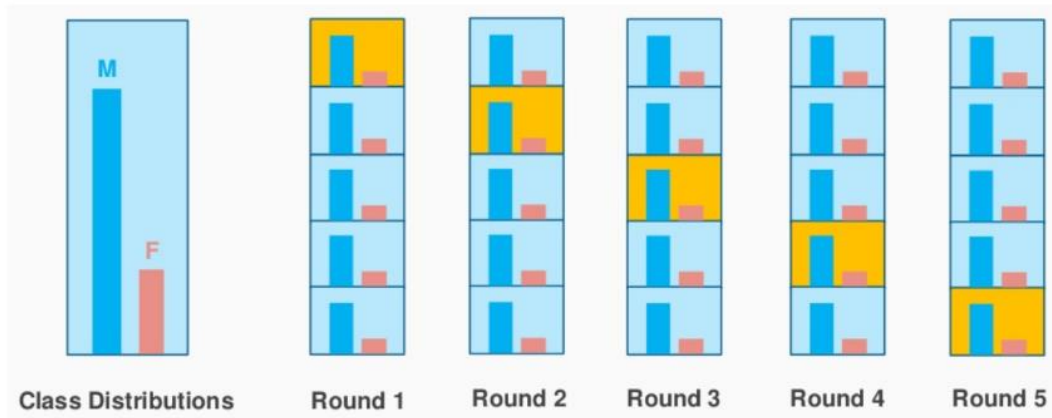
The correctness, accuracy of a model/algorithm developed to predict the class of the chosen problem can be evaluated by various criterion like Cross-Validation, Confusion Matrix, Precision and Recall, F1 Score and the ROC Curve (Receiver Operating Characteristic) (Aurélien Géron )

#### **Cross – Validation:**

This process of deciding whether the numerical results quantifying hypothesized relationships between variables, are acceptable as descriptions of the data, is known as validation. (Prashant Gupta, 2017)

Cross validation aims at resampling the given dataset into multiple sets by splitting so that model use a set for training and another for testing in the basic method called Handout method whereas K fold cross validation divided a given dataset into K subsets. Cross Validation example is shown in the Figure 3.6

Figure 3.6: Cross validation explained



**Confusion Matrix** is the better way of evaluating the classification model, it measures how many times the classifier is confused with the targeted class (Class A predicted as Class B) (Aurélien Géron, ). A confusion matrix is formed of the predicted results compared against the actual results in validation dataset consisting of True Positives (TP), True Negatives (TN) , False Positives (FP) and False Negatives (FN) .

Figure 3.7: Confusion Matrix explained

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

True Positives are the instances that are predicted Class A as Class A

True Negatives are the instances that are predicted Class B as Class B

False Positives are the instances that are predicted Class A as Class B

False Negatives are the instances that are predicted Class B as Class A

**Precision** is the ratio of True Positives (TP) to the sum of True Positives (TP) and False Positives (FP). It addresses the proportion of positive instances

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

**Recall** is the ratio of True Positives (TP) to the sum of True Positives (TP) and False Negatives (FN). It is also called the sensitivity or TPR (True Positive Rate) (Aurélien Géron, ). Recall addresses the proportion of actual positive instances

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

**F score (F1 Score)** or the accuracy score combines the Precision and Recall obtaining a score that ranges from 0 to 1 with higher the score is the most accurate classifier

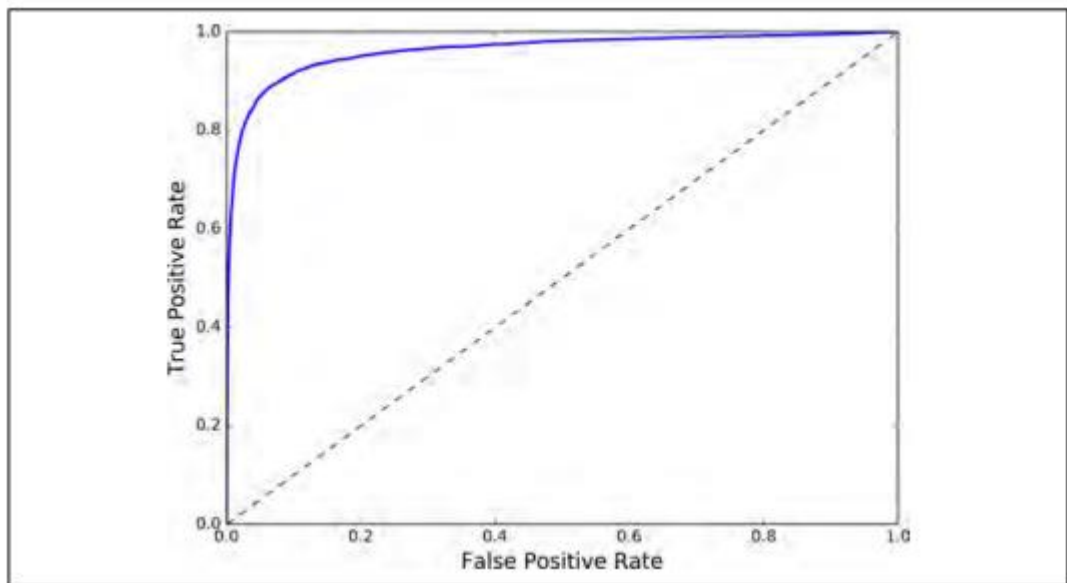
$$\begin{aligned} F_1 &= \frac{2}{\left(\frac{1}{\text{precision}}\right) + \left(\frac{1}{\text{recall}}\right)} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \\ &= TP / \left(TP + \left(\frac{FN + FP}{2}\right)\right) \end{aligned} \quad (4)$$

**The ROC Curve** (receiver operating characteristic) is another important measure for the binary classification. It is a plot drawn on the true positivity rate (TPR) or the recall against the false positivity rate (FPR). The plot is drawn on x axis with false positivity rate and y axis with true positivity rate. True positivity rate is also called sensitivity and True Negativity Rate is called the specificity. FPR can be defined as 1 – specificity hence the ROC curve plot can also called as drawn against sensitivity vs (1 - specificity) (Aurélien Géron, )

$$\text{True Positivity Rate (recall or sensitivity)} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{False Positivity Rate} = \frac{FP}{FP+TN} \quad (7)$$

*Figure 3.8: ROC Curve*



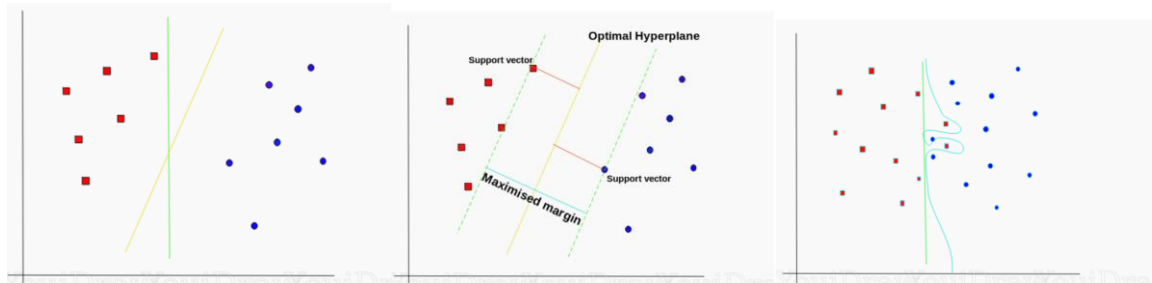
### 3.5 Classifiers used in the current study

#### 3.5.1 Support Vector Machines

Support Vector machines are the simplest machine learning algorithms that are used to solve the classification or regression problems. The SVMs can be of linear and non-linear where the key distinction is the plane (line or hyperplane) created by the algorithm that separates data into various groups or classes

The idea of SVM is to approximate and find a straight line or higher plane between data of two classes and find the closest points to the line/hyperplane. The closest points are called the Support Vectors and the distance between line and vector is calculated and is called the margin. Maximising the margin is the key objective of using the SVM. (Rushikesh Pupale, 2018)

*Figure 3.9: SVM planes explained*



The underlying idea behind the Support Vector Machines are explained in the Figure 3.9.

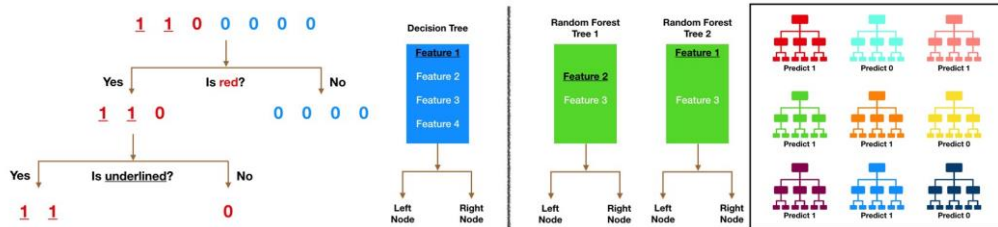
C value and the Gamma are the most important tuning parameters of the Support Vector Machines Classifier. C value helps the fine tuning the boundary and helps the training points classification correctly, the larger the C value, the more training points getting accurately. Gamma defines the reach and influence of a single training instance, the higher the Gamma value, the reach is higher or in otherwards the instances or points are closely connected

### 3.5.2 Random Forest Classifier

Random forest classifier is another machine learning technique to predict the class or regression problem. Like other classifiers, the random forest classifier can predict binary or multi-class problems as well. Decision Trees and grouping of the Trees into an ensemble (Forest) are the underlying ideas behind the random forest classifier. The algorithm offers identification of class

into decision trees and group them based on features as shown the in the Figure 3.10. (Tony Yiu, 2019)

*Figure 3.10: Random forest classification explained*



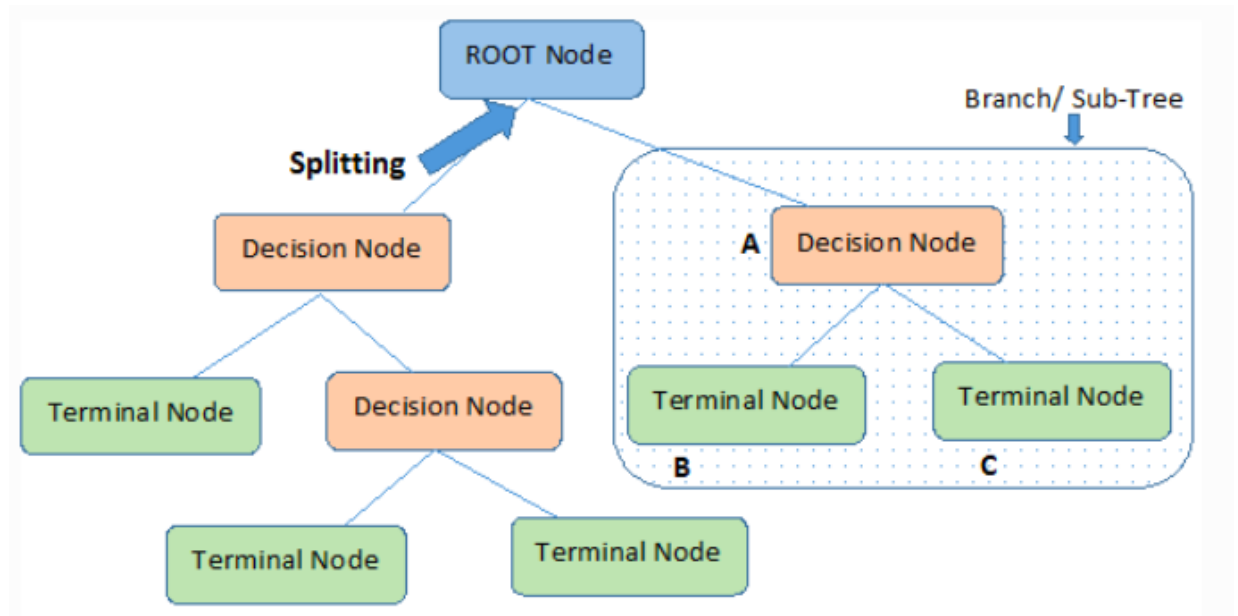
The meta estimator (random forest) that uses the number of classified decision trees with subsets of data on the given dataset by using the average to enhance the accuracy and ensured no over fitting.

### 3.5.3 Decision Tree classifier

Decision Trees is yet another shallow machine learning classification technique that can be used either for classification of categorical variable or prediction of continuous variable. Decision Trees starts with treating entire dataset as root node and divides it into two or more datasets called sub nodes. Each sub node is further divided into sub-nodes, each sub node that have possibility of diving into further nodes is called decision node. A sub node that cannot be divided further is called a leaf or terminal node as shown in the Figure 3.11



Figure 3.11: Decision tree classification explained



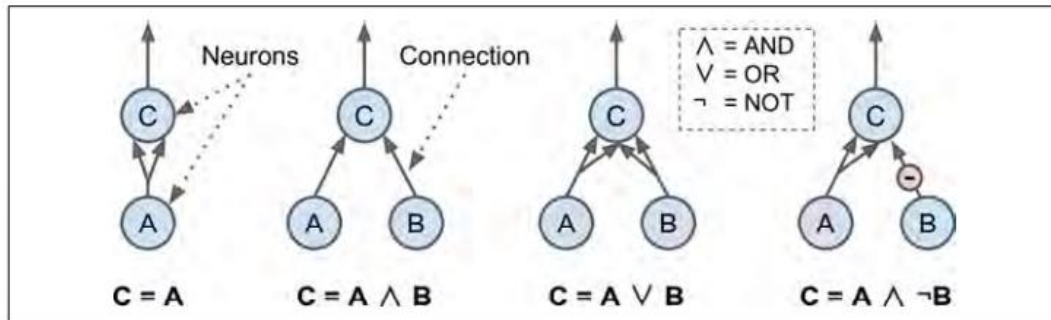
Various key attributes that influence the measure of the Decision Tree algorithm are as follows

- Criterion – measure of the quality of data split (can be of gini, entropy or log\_class)
- Split of each node s
- Depth of the Tree
- No. of features
- Randomness of estimator

### 3.5.4 Multi-Level Perceptron (MLP) – ANN

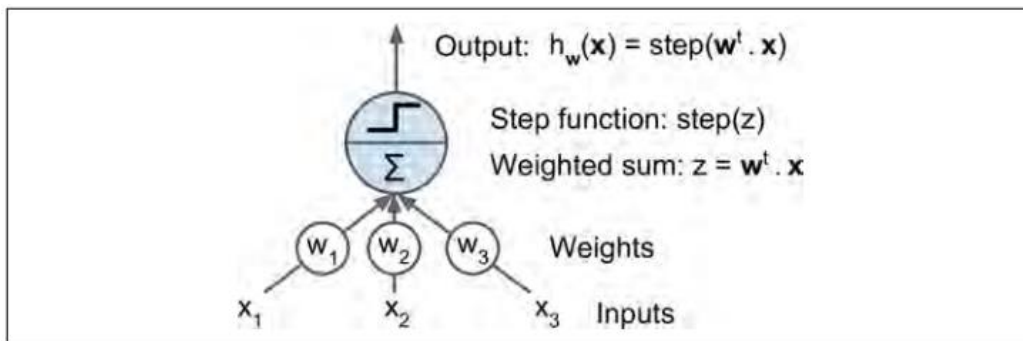
Multi – Level Perceptron also called an MLP is a deep learning algorithm built on Artificial Neural Network architecture. In order explain the MLP model, it needs to know the basic elements of Deep Learning called **artificial neuron and perceptron**. Artificial Neuron or simply neuron is made of one or more binary inputs and generates one binary output when the binary inputs are activated. An artificial neuron can be activated when at least two of its inputs are made ON. The simplest form of a neuron can be seen in the Figure 3.12

Figure 3.12: Artificial neuron



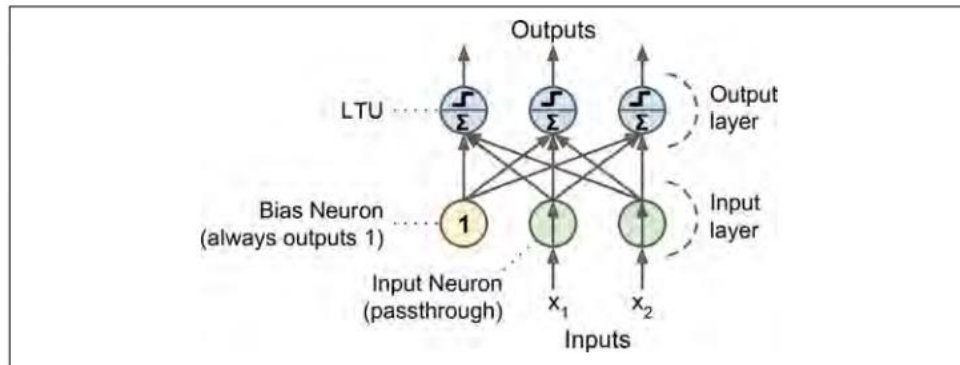
The Perceptron is bit different from artificial neuron and is consisting of LTUs (Linear Threshold Unit). Unlike binary inputs and outputs in the neuron, The perceptron takes numbered inputs called weights. The LTU performs the activity of summing up the input numbers called weighted sum. Upon the sum, the LTU adds a step function to the sum and produces an output as a result as shown in the Figure 3.13.

Figure 3.13: LTU (Linear Threshold Unit)



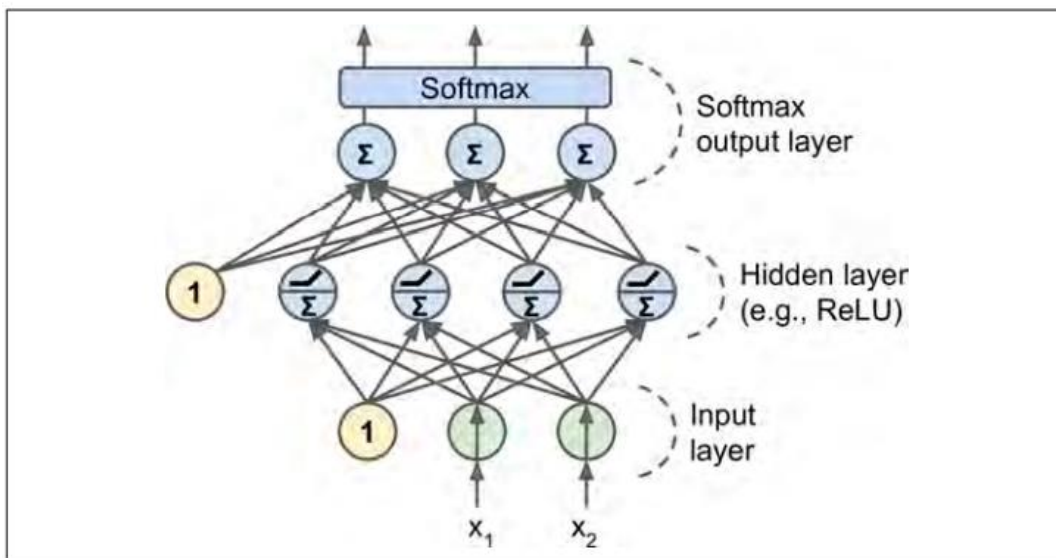
The step functions that are used in the perceptron are Heaviside step function and sign function. A Perceptron is nothing but a composition of single layer of LTU connected with all its inputs by neurons. An example of two inputs and three outputs perceptron is shown in the diagram (Figure 3.14)

Figure 3.14: Perceptron Explained



A multi-layer perceptron (MLP) is comprised of one passthrough input layer, one or more LTUs hidden layers and one output layer (a layer of LTUs). All the layers include a bias neuron except the output layer and are completely connected with the successive layer, when a network has two or more layers of artificial neural network, it is called DNN. A multi-layer perceptron based Artificial Neural Network with step function is shown in the Figure 3.15

Figure 3.15: Multi-Layer Perceptron Explained



An MLP can be used for the prediction of the classification problem where each output denotes a binary class (istumor Yes or No; Spam or Ham etc). In case of multi-class prediction (say classes 0 to 9 image classification) , a shared softmax can be used in place of individual activation function. As the input data flows in one direction, this architecture is also called feed forward neural network (FNN)

Softmax function:

$$p_k = \sigma(S(x))_k = \frac{\exp(s_k(x))}{\sum_{j=1}^K \exp(s_j(x))} \quad (8)$$

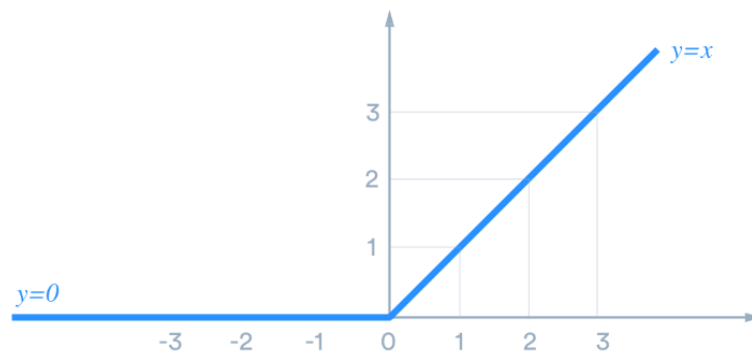
- $k$  is the number of classes.
- $s(x)$  is a vector containing the scores of each class for the instance  $x$ .
- $\sigma(s(x))_k$  is the estimated probability that the instance  $x$  belongs to class  $k$  given the scores of each class for that instance.

ReLU: Rectified Linear Unit is a non-linear or layer based linear activation function. A ReLU calculates the inputs linear function and output the result if it is positive otherwise 0

(Aurélien Géron, ) can be expressed as the Formula and shown in Figure 3.16.

$$f(x) = \max(0, x) \quad (9)$$

*Figure 3.16: ReLU activation function*



## 4Data Analysis

### 4.1 Dataset

The dataset was sourced from BRaTS-Challenge team, that consists of 140 folders of training and 20 folders of validation. Each folder in the training dataset further consists of 4 modality NIfTI files and a landmark csv file for each scan (i.e., 4 NIfTI files for base scan and a landmark csv file and similarly 4 NIfTI files and a landmark file for the follow up scan). Validation dataset each folder will consists of 4 modality NIfTI files for each scan (base and follow-up) and a ground truth landmarks csv file for the follow-up scan

The 4 modalities files for each scan that are also called multi parametric MRIs are respectively 1) native T1-weighted (T1), 2) contrast-enhanced T1 (T1-CE), 2) T2-weighted (T2), and 4) T2 Fluid Attenuated Inversion Recovery (FLAIR) [10]. The base scan NIfTI/csv files can be identified with filename substring notation that contains “\_00\_” and follow-up scan files can be recognized with “\_01\_”, simplified and tabulated in the Table 4.1 and a sample folder filename(s),formats and size(s) can be seen in Figure 4.1

*Table 4.1: Dataset file composition break up table*

Purpose	Scan Type	File Type	FileCnt Per Folder	Total
Training (140)	Base	Modality (NIfTI)	4	560
		landmark (csv)	1	140
	Follow-up	Modality (NIfTI)	4	560
		landmark (csv)	1	140
Validation (20)	Base	Modality (NIfTI)	4	80
		landmark(csv)	0	0
	Follow-up	Modality (NIfTI)	4	80
		landmark(csv)	1	20

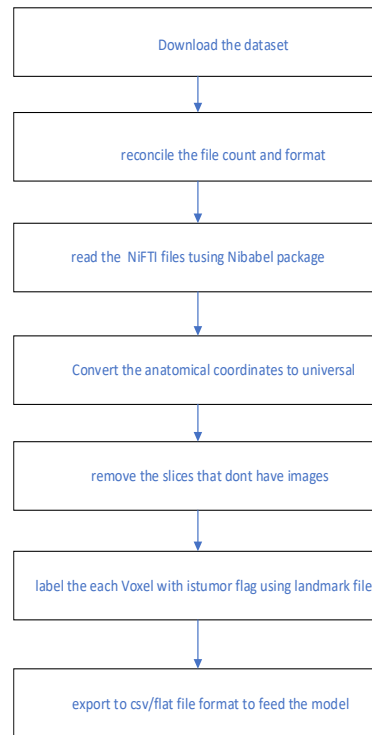
Figure 4.1: Sample folder content for file name(s), type(s) and size(s)

OneDrive - Sheffield Hallam University > OS: Dissertation > BRATS Dataset > BraTSReg_Training_Data_v2 > BraTSReg_001					OneDrive - Sheffield Hallam University > OS: Dissertation > BRATS Dataset > BraTSReg_Validation_Data > BraTSReg_141				
Name	Status	Date modified	Type	Size	Name	Status	Date modified	Type	Size
BraTSReg_001_00_0000_flair.nii.gz	⬆ R	04/01/2022 13:42	GZ File	1,926 KB	BraTSReg_141_00_0000_flair.nii.gz	⬆ R	15/11/2021 21:41	GZ File	2,428 KB
BraTSReg_001_00_0000_t1.nii.gz	⬆ R	04/01/2022 13:42	GZ File	2,076 KB	BraTSReg_141_00_0000_t1.nii.gz	⬆ R	15/11/2021 21:41	GZ File	2,239 KB
BraTSReg_001_00_0000_t1ce.nii.gz	⬆ R	04/01/2022 13:42	GZ File	2,015 KB	BraTSReg_141_00_0000_t1ce.nii.gz	⬆ R	15/11/2021 21:41	GZ File	2,576 KB
BraTSReg_001_00_0000_t2.nii.gz	⬆ R	04/01/2022 13:42	GZ File	1,798 KB	BraTSReg_141_00_0000_t2.nii.gz	⬆ R	15/11/2021 21:41	GZ File	2,367 KB
BraTSReg_001_01_0106_flair.nii.gz	⬆ R	04/01/2022 13:42	GZ File	1,836 KB	BraTSReg_141_01_0505_flair.nii.gz	⬆ R	16/11/2021 14:14	GZ File	5,370 KB
BraTSReg_001_01_0106_t1.nii.gz	⬆ R	04/01/2022 13:42	GZ File	2,114 KB	BraTSReg_141_01_0505_t1.nii.gz	⬆ R	16/11/2021 14:14	GZ File	5,454 KB
BraTSReg_001_01_0106_t1ce.nii.gz	⬆ R	04/01/2022 13:42	GZ File	2,028 KB	BraTSReg_141_01_0505_t1ce.nii.gz	⬆ R	16/11/2021 14:14	GZ File	5,377 KB
BraTSReg_001_01_0106_t2.nii.gz	⬆ R	04/01/2022 13:42	GZ File	1,848 KB	BraTSReg_141_01_0505_t2.nii.gz	⬆ R	16/11/2021 14:14	GZ File	5,461 KB
BraTSReg_001_00_0000_landmarks.csv	⬆ R	04/01/2022 13:42	Microsoft Excel C...	1 KB	BraTSReg_141_01_0505_landmarks.csv	⬆ R	18/01/2022 22:22	Microsoft Excel C...	1 KB
BraTSReg_001_01_0106_landmarks.csv	⬆ R	04/01/2022 13:42	Microsoft Excel C...	1 KB					

## 4.2 Pre-processing

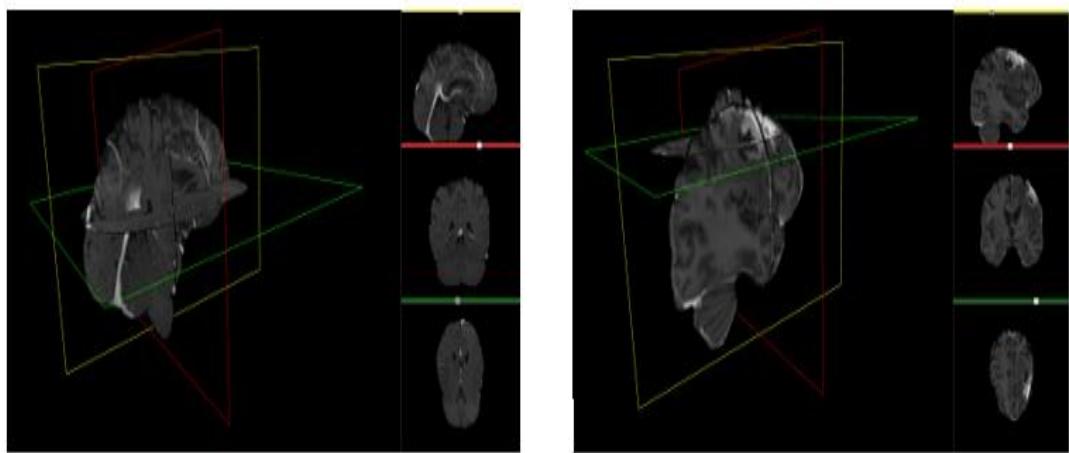
The pre-processing steps involved in transformation of the NIfTI files into model readable format, combining landmark csv files and labelling of each Voxel coordinates is depicted into a flow diagram as shown in Figure 4.2.

Figure 4.2: Data pre-processing flowchart



The dataset preprocessing as various steps involved starts from download the dataset from the BRAITS challenge website from the request (CBICA image processing portal.2022) , analyzing the raw NIFTI files. The 3d NIfTI files can be viewed on all the 3-dimensional planes from any software that is compatible for reading NIFTI files. It's been used two software's for the brain viewer web app (Mri viewer [computer software]) in Figure 4.3 for complete visualization of all the 155 slices (each NIfTI file consists of 155 slices on Z places of each one mm thickness ) (CBICA image processing portal.2022). More technical details like affine, file header was better analyzed with the ImageJ software

*Figure 4.3: Sample 3-dimensional planes analysis for training set (left) and validation set(right) using brain viewer software*



The NIFTI files uses the anatomical coordinate system that was automatically captured as part of MRI scanning devices, whereas the tools [Python] was used to develop the machine learning model normally uses the world coordinate system. In this process it had to convert the anatomical coordinate system to world coordinate system as the provided ground truth landmarks were in anatomical coordinates

To ensure the coordinates conversion was correct, three levels of checks were done 1) calculated the affine 2) requested the BRAITS team to send a sample NIfTI file and landmarks file to process and 3) received a confirmation from BRAITS team on the sample file and landmarks file received upon the request as shown in the Figure 4.4 the conversion in which X,

Y, Z are from the anatomical coordinates and the pY is converted from anatomical to world coordinates

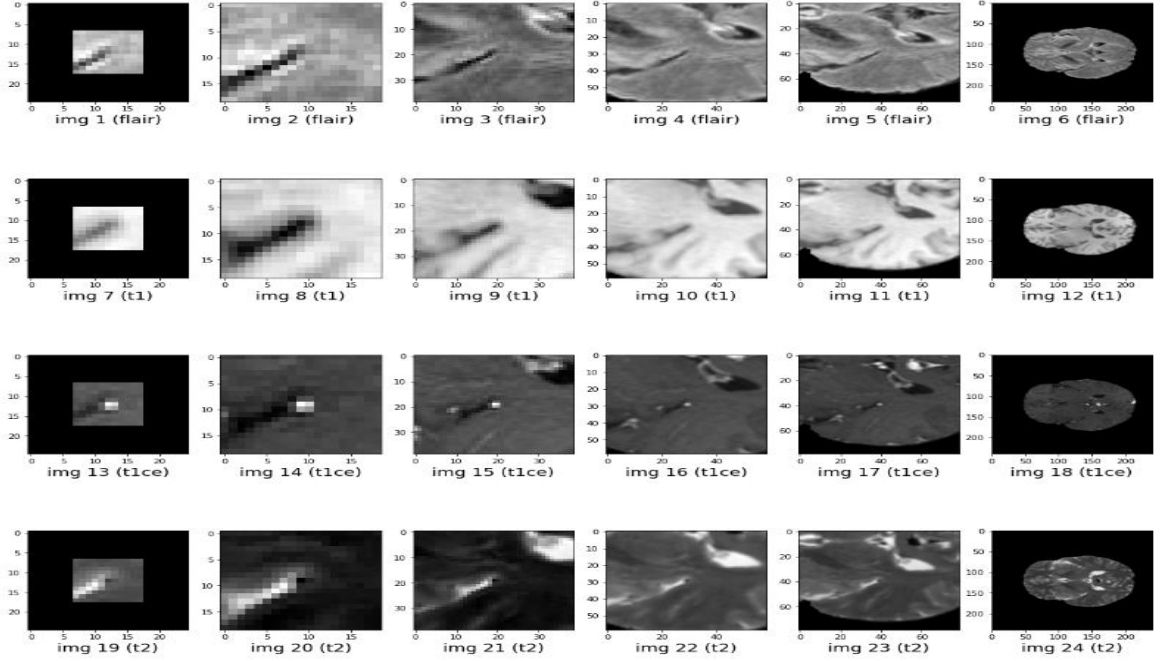
*Figure 4.4: The sample NIFTI file processed mapping landmarks. NIfTI and landmarks both were supplied by BRaTS challenge team and acknowledged on conversion process correctness*

	Index	NiiFileName	X	Y	pY	Z	VoxVal
0	24633	Sample_landmarks.nii.gz	153	-137	102	57	1
1	23439	Sample_landmarks.nii.gz	159	-142	97	63	2
2	26045	Sample_landmarks.nii.gz	125	-131	108	68	4
3	18093	Sample_landmarks.nii.gz	93	-164	75	77	6
4	23400	Sample_landmarks.nii.gz	120	-142	97	77	5
5	21518	Sample_landmarks.nii.gz	158	-150	89	87	3
6	27228	Sample_landmarks.nii.gz	108	-126	113	95	7
7	18574	Sample_landmarks.nii.gz	94	-162	77	123	8

Contrast-enhanced T1 (T1-CE) modality was considered as the best modality for model developments as the modality was proven the best candidate modality by the BRaTS team (Bhakti Baheti et al., 2021), it was also also processed that the landmark of a file with 10mm x 10mm region (X, pY) around the landmark for a slice(Z) shown in Fig 4.5

*Figure 4.5: 4 modalities of a base registration file enlarged from 10mm x10mm to full size and t1ce proved best candidate (3rd row image no.13 that has center pixel is tu-mor marked by experts)*





Once the coordinates conversion formula and modality candidate selection are finalized, the conversion and labelling applied on the actual training dataset (T1-CE modality). All the NIfTI files were converted into flat files with istumor label for each pixel and the pixel values were captured. As the pixel is for 3-dimensional slice of one mm (1mm x 1mm x 1mm) volumetric thickness (Bhakti Baheti et al., 2021), and it was named in the flat file as VoxVal. The VoxVal (the pixel value) is the colour intensity of the cell as shown in Figure 4.6

As part of the exploratory data analysis, the colour intensity value 0 (images where there was no brain matter associated or also called peripheral images) were removed and the final files were saved as csv files.

Figure 4.6: A NIfTI file converted into flat file with istumor label

Unnamed: 0	X	pY	VoxVal	Z	NiiFileName	LandmarksFileName	istumor
45496	136	189	416.500977	58	BraTSReg_141_01_0505_t1ce.nii.gz	BraTSReg_141_01_0505_landmarks.csv	1
28176	96	117	508.371704	66	BraTSReg_141_01_0505_t1ce.nii.gz	BraTSReg_141_01_0505_landmarks.csv	1
25064	104	104	890.701843	71	BraTSReg_141_01_0505_t1ce.nii.gz	BraTSReg_141_01_0505_landmarks.csv	1
25104	144	104	645.254333	72	BraTSReg_141_01_0505_t1ce.nii.gz	BraTSReg_141_01_0505_landmarks.csv	1
28661	101	119	632.807434	77	BraTSReg_141_01_0505_t1ce.nii.gz	BraTSReg_141_01_0505_landmarks.csv	1
36405	165	151	665.296570	80	BraTSReg_141_01_0505_t1ce.nii.gz	BraTSReg_141_01_0505_landmarks.csv	1
28903	103	120	1105.026367	85	BraTSReg_141_01_0505_t1ce.nii.gz	BraTSReg_141_01_0505_landmarks.csv	1
27709	109	115	651.657043	111	BraTSReg_141_01_0505_t1ce.nii.gz	BraTSReg_141_01_0505_landmarks.csv	1
42541	61	177	1798.220947	97	BraTSReg_141_01_0505_t1ce.nii.gz	BraTSReg_141_01_0505_landmarks.csv	0

### 4.3 Verification and Validation

Verification and validation is very important activity in the data pre-processing phase. This activity ensures that the raw dataset or files was successfully processed without any errors or data loss or any unwanted and unknown data fabrication. The activity also ensures that correctness of implemented solution in terms of design decisions that were taken throughout the model development life cycle. As per (Ben H. Thacker et al., 2004), the verification and validation model can make predictions of engineering with a quantified confidence

The validations performed as part of the current study solution development are

- dataset processing without any errors (affinity mapping)
- elimination of null, blank, unknown or unwanted values
- testing of landmark coordinates (that are provided separately) mapping to the processed NIfTI (or nii) files

Affinity mapping was confirmed with the help of BRaTS team, by sharing the processed values with them. In turn BRaTS team has provided a simple NIfTI file to cross check the processed values and they both matched and can be seen in Figure 4.7

Figure 4.7: Sample file verified with BRaTS team

	Index	NiiFileName	X	Y	pY	Z	VoxVal
0	24633	Sample_landmarks.nii.gz	153	-137	102	57	1
1	23439	Sample_landmarks.nii.gz	159	-142	97	63	2
2	26045	Sample_landmarks.nii.gz	125	-131	108	68	4
3	18093	Sample_landmarks.nii.gz	93	-164	75	77	6
4	23400	Sample_landmarks.nii.gz	120	-142	97	77	5
5	21518	Sample_landmarks.nii.gz	158	-150	89	87	3
6	27228	Sample_landmarks.nii.gz	108	-126	113	95	7
7	18574	Sample_landmarks.nii.gz	94	-162	77	123	8

The code snippet in the Figure 4.8 explains (underlined in red row) the elimination of unwanted (precisely the coordinates of Voxel value 0) data from the processed files. (NIfTI files are converted into csv as part of data pre-processing by eliminating unwanted rows and introduced a new flag that denotes whether a coordinate is a tumor or not with the help of landmark files supplied)

Figure 4.8: code snippet of unwanted data removal

```
def get_nii_to_csv_df(self,nii_fn,nii_lm_fn,nii_np): # dont pass empty dataframe nii_df; function to convert nii_numpy into csv
    self.nii_df = pd.DataFrame()
    for i in range(0,155):
        slice = nii_np[:, :, i] # one slice of nii
        slice_df = pd.DataFrame(slice)
        slice_df_melt = pd.DataFrame(slice_df).reset_index().melt('index') # transpose the x y coordinates into two columns
        slice_df_melt.columns = ['X', 'pY', 'VoxVal']
        slice_df_melt['Z'] = i
        slice_df_melt['NiiFileName'] = nii_fn
        slice_df_melt['LandmarksFileName'] = nii_lm_fn
        slice_df_melt = slice_df_melt.drop(slice_df_melt[slice_df_melt.VoxVal == 0].index)
        #print(slice_df_melt.head)
        self.nii_df = self.nii_df.append(slice_df_melt)
    return self.nii_df
```

Another validation performed was the reconciliation of provided landmark coordinates and mapped coordinates to ensure the no. of mappings were correct by randomly validating the processed NIfTI csv files and ground truth landmark csv files provided shown in Figure 4.9

Figure 4.9: validation of MR Image coordinates mapped with ground truth landmarks

	X	Y	Z	niFileName	LandmarksFileName	istumor
321565	133	105	3526	51 BraTSReg_001_00_0000_t1ce.nii.gz	BraTSReg_001_00_0000_landmarks.csv	1
361845	121	96	2543	54 BraTSReg_001_00_0000_t1ce.nii.gz	BraTSReg_001_00_0000_landmarks.csv	1
420056	144	103	2865	58 BraTSReg_001_00_0000_t1ce.nii.gz	BraTSReg_001_00_0000_landmarks.csv	1
563261	163	134	1543	67 BraTSReg_001_00_0000_t1ce.nii.gz	BraTSReg_001_00_0000_landmarks.csv	1
583346	167	98	1596	69 BraTSReg_001_00_0000_t1ce.nii.gz	BraTSReg_001_00_0000_landmarks.csv	1
674883	151	141	2004	70 BraTSReg_001_00_0000_t1ce.nii.gz	BraTSReg_001_00_0000_landmarks.csv	1

Landmark	X	Y	Z
1	121	-143	54
2	167	-141	69
3	163	-115	67
4	151	-98	70
5	133	-134	51
6	144	-136	58

#### 4.4 Data Augmentation

Data augmentation is a technique that allows to synthesize existing data based on the objective of the requirement. As per (Connor Shorten et al., 2021), contains small changes in the existing data that ensures the models prediction objective is variant. For the current study, various data sampling methods were utilised depending on the size of data that was used. E.g., for the SVM, Random Forest and Decision Trees classifiers, istumor = 1 flag data was up sampled to the same number of rows as istumor = 0, whereas for the MLP ANN sequential classifier, minor data (istumor = 1) was up sampled to same number of rows as major data (istumor = 0) with 13% of dataset and up sampling was restricted to 25% with full dataset. A code snippet of data augmentation (resample library) is shown the Figure 4.10

Figure 4.10: Data augmentation code snippet

```
df_no_tumor = b_dataset[(b_dataset['istumor']==0)]
df_is_tumor = b_dataset[(b_dataset['istumor']==1)]
# augment the is_tumor class
df_is_tumor_upsampled = resample(df_is_tumor,
                                replace=True, # augment with replacement
                                n_samples= len(df_no_tumor), # to match no_tumor class
                                random_state=42) # reproducible results
# unify the no_tumor and upsampled is_tumor dataframes
df_augmented = pd.concat([df_is_tumor_upsampled, df_is_tumor])
print("count of class {}".format(df_augmented['istumor'].value_counts()))
```

## 5 Mode Development

The selection of classifiers was made as per the nature of the current study (prediction of landmark coordinates on a dataset that has ground truth landmark coordinates) relevantly and all the selected model's development and outcomes are discussed extensively in the current section.

The selection choice of the models also is as per the (Momina Masood et al., 2021) as it was highlighted in the paper that dataset with the provision of domain expert curated ground truth landmarks can applied shallow machine learning techniques.

### 5.1 Non-linear SVM

Nonlinear Support Vector Machine classifier was selected as the Voxel values trend of istumor class was detected as non-linear nature. Also, since SVM is applied only smaller datasets (JairCervantes, XiaouLi et. al 2008), the given dataset was curtailed so analyse the results pattern. The three-dimensional images of brain for all the 140 samples are curtailed with extraction of only lesion (wound area) of  $5 \text{ mm}^3$  radius across each landmark ground truth and fed to the model. train test split ration was given 70:30 with nu value 0.01 and the model accuracy was obtained with a score of 99.86 as shown in Figure 5.1

*Figure 5.1: SVM implementation code snippet*

```
# Import train_test_split function
from sklearn.model_selection import train_test_split

print(len(b_dataset))
X=b_dataset[['X','pV','VoxVal','Z']]
y=b_dataset[['istumor']]
# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3) # 70% training and 30% test

1729253

#svm model
from sklearn import svm
# fit the model
svm_clf = svm.NuSVC(gamma="auto",nu=0.001)
svm_clf.fit(X_train,y_train)
y_pred=svm_clf.predict(X_test)
# svm model accuracy
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

Accuracy: 0.998605968437664
```

Testing accuracy was experimentally reviewed by analysing the output for a sample file shown in the Figure 5.2. The model has detected almost all the istumor class coordinates as per the ground truth coordinates

*Figure 5.2: testing accuracy of SVM*

	X	pY	VoxVal	Z	NiiFileName	LandmarksFileName	istumor	istumor_Pred
0	87	94	1533.0	34	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
1	105	101	1547.0	45	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
2	105	155	857.0	62	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
3	86	135	1009.0	68	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
4	128	112	1628.0	73	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
5	118	97	875.0	92	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1

However, while analysing the model with the validation dataset, the results were no longer optimistic. The sample results of istumor class ground truth and predicted values shown in Figure 5.3. None of the predicted values are near the realistic range, hence the model was stopped at the early stage

*Figure 5.3: validation accuracy of SVM*

	Unnamed: 0	X	pY	VoxVal	Z	NiiFileName	LandmarksFileName	istumor	istumor_Pred
661268	50736	96	211	2515.424072	68	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
826370	50766	126	211	3580.228027	77	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
917658	50766	126	211	2701.193848	82	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
938661	18576	96	77	2912.270752	84	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1045343	15913	73	66	1281.943604	90	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1045527	16393	73	68	1346.456055	90	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1045621	16633	73	69	1259.177124	90	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1148070	47361	81	197	1454.310303	95	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1200782	15515	155	64	1296.225830	99	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1512043	21990	150	91	1305.942261	122	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1512677	24150	150	100	1225.531006	122	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1512905	24871	151	103	1295.634399	122	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1528112	40918	118	170	5199.113770	123	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1
1539244	24150	150	100	1207.061401	125	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_landmarks.csv	0	1

## 5.2 Random Forest Model

Random forest model is another early choice of classifiers and one of the best suited models for the smaller datasets. The voluminous dataset that was curtailed to extract only the lesion part surrounded by the ground truth landmarks was fed to the random forest model as well and

obtained similar accuracy as SVM with a score of 99.85 as shown in the Figure 5.4 with a test train split of 70:30 ratio which is safest and proven ratio to train the model

*Figure 5.4: Implementation of Random Forest classifier*

```
# Import train_test_split function
from sklearn.model_selection import train_test_split

#X=df_upsampled[['X','pY','VoxVal','Z']]
#y=df_upsampled[['istumor']]
#b_dataset = b_dataset.head( 726731)
print(len(b_dataset))
X=b_dataset[['X','pY','VoxVal','Z']]
y=b_dataset[['istumor']]
# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3) # 70% training and 30% test

1729253

#Import Random Forest Model
from sklearn.ensemble import RandomForestClassifier
rf_clf=RandomForestClassifier(n_estimators=100)
rf_clf.fit(X_train,y_train)
y_pred=rf_clf.predict(X_test)
# rf model accuracy
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

Accuracy: 0.9985581445556464
```

Experimentally analysed the testing accuracy which has performed well with istumor class for the ground truth landmarks were predicted by the model as shown in Figure 5.5

*Figure 5.5: Testing accuracy of Random Forest classifier*

	X	pY	VoxVal	Z	NiiFileName	LandmarksFileName	istumor	istumor_Pred
0	87	94	1533.0	34	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
1	105	101	1547.0	45	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
2	105	155	857.0	62	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
3	86	135	1009.0	68	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
4	128	112	1628.0	73	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1
5	118	97	875.0	92	BraTSReg_010_01_0083_t1ce.nii.gz	BraTSReg_010_01_0083_landmarks.csv	1	1

However, like the SVM, prediction of validation dataset results was not optimistic which has led to the early stopping of the random forest classifier. Istumor ground truth values and predicted values are show in the Figure 5.6

*Figure 5.6: Validation accuracy of Random Forest classifier*

	Unnamed: 0	X	pY	VoxVal	Z	NiiFileName	LandmarksFileName	istumor	istumor_Pred
661268	50736	96	211	2515.424072	68	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
826370	50766	126	211	3580.228027	77	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
917658	50766	126	211	2701.193848	82	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
938661	18576	96	77	2912.270752	84	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1045343	15913	73	66	1281.943604	90	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1045527	16393	73	68	1346.456055	90	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1045621	16633	73	69	1259.177124	90	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1148070	47361	81	197	1454.310303	95	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1200782	15515	155	64	1296.225830	99	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1512043	21990	150	91	1305.942261	122	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1512677	24150	150	100	1225.531006	122	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1512905	24871	151	103	1295.634399	122	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1528112	40918	118	170	5199.113770	123	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1
1539244	24150	150	100	1207.061401	125	BraTSReg_142_01_0154_t1ce.nii.gz	BraTSReg_142_01_0154_Landmarks.csv	0	1

### 5.3 Decision Trees

Another attempt was made with different classifier which is suitable for large datasets Trees (Khaled Alsabti, Sanjay Ranka, & Vineet Singh, ) is Decision Trees Classifier. This time the strategy is switched from the Lesion only Brain images to the complete brain images, however due to the limited system configuration availability 26% of the full dataset was used with data augmentation of minor(istumor = 1) class up sampling to the major class (istumor = 0) with entropy criteria as shown in the Figure 5.7. The model accuracy was obtained with a score of 99.99%

*Figure 5.7: Implementation of Decision Tree classifier*



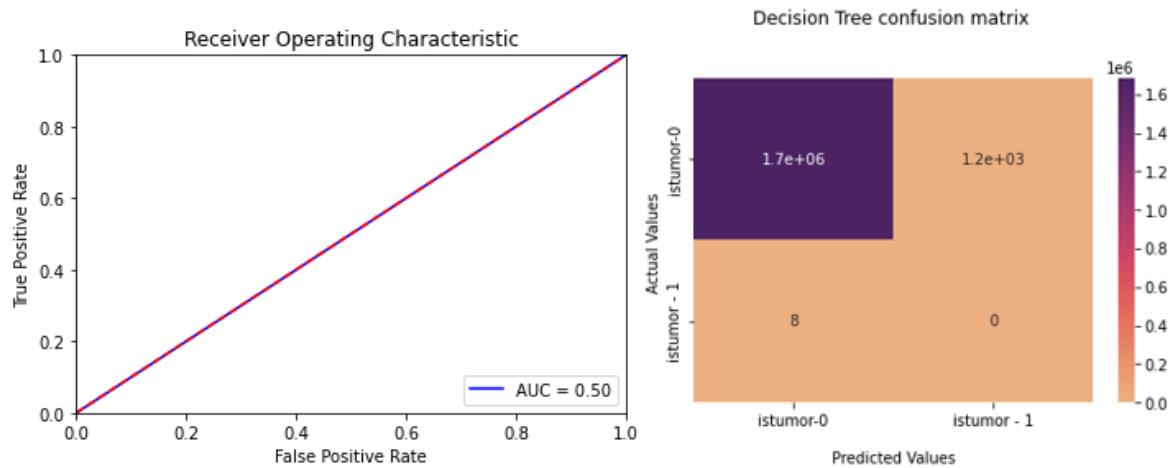
```
# entropy means information gain
classifier = DecisionTreeClassifier(criterion='entropy', random_state=0)

# providing the training dataset
classifier.fit(X_train,y_train)
y_pred = classifier.predict(X_test)
# rf model accuracy
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

Accuracy: 0.9999981166978998

However, while experimentally testing the model with the validation dataset, the results were not optimistic either, hence decision was made to early stoppage of the model rather than going further. However, the model was able to predict the istumor = 0 class to the most, but none of the ground truth landmarks were predicted by the model. The confusion matrix and ROC curves of the decision tree model were shown in the Figure 5.8 where the roc curve is neutral with no improvement.

*Figure 5.8: ROC curve and confusion matrix of Decision Tree Classifier*



#### 5.4 Deep Learning - ANN Multi Level Perceptron

The Multi-Layer Perceptron sequential model, a deep learning Artificial Neural Network technique (Christian Janiesch et al., 2021) for the supervised binary label classification also called multi-layer perceptron has achieved the optimistic results. Initially attempted with 13% of the dataset and data augmentation of minor class equal to major class that obtained optimistic

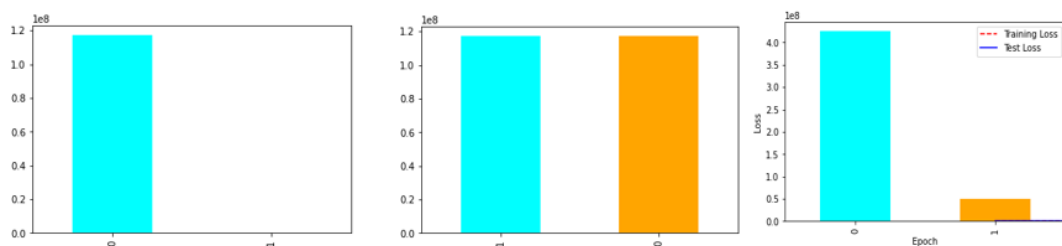
results. Then developed an improved model with full dataset with data augmentation of minor class as 25% of the major class (due to limited system configuration)

Both the models were defined with the input shape of (4,) (i.e., four features and a target label) with “uniform” kernel initializer. The activation method ReLU was used and the dense parameters 12,8 and 1. The model was compiled with the loss='binary\_crossentropy', optimizer='adam', metrics=['accuracy'] Number of Epochs were used 50 and 25 for the a) 13% of dataset, here after called as Model A and b) full dataset with 25% augmentation, here after called as Model B respectively and the batch size taken as 500000. Usually, the no. of epochs will be 3 times the no. of features but due to huge volume of data and to see stabilization in the training accuracy, it was chosen to 50 and 25. The accuracy of model A was 91.89 whereas the Model B has achieved 89.27. The implemented ANN MLP model definition is shown in the Figure 5.9 MLP, and the data augmented was shown in the Figure 5.10

*Figure 5.9: Implementation of MLP keras sequential model*

```
# define the keras model
model = Sequential()
model.add(Dense(12, input_shape=(4,), kernel_initializer = 'uniform', activation='relu'))
model.add(Dense(8, kernel_initializer = 'uniform', activation='relu'))
model.add(Dense(1, kernel_initializer = 'uniform', activation='sigmoid'))
#-----
# compile the keras model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
#-----
# fit the keras model on the dataset
epoch_history = model.fit(X, y, epochs=25, batch_size=500000, validation_split=0.2)
```

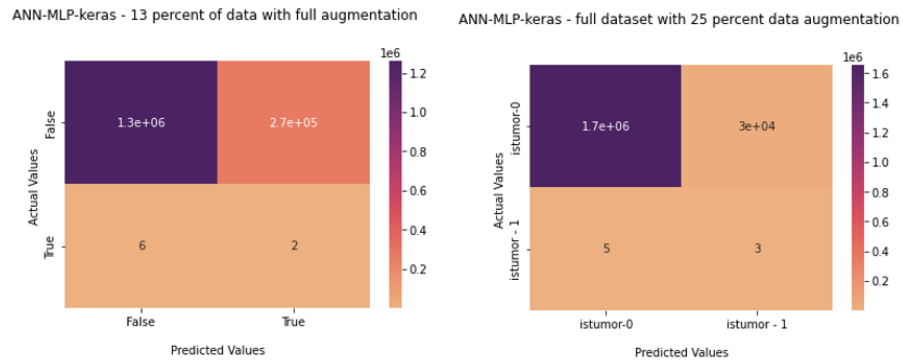
*Figure 5.10: Data augmentation of Model A(centre) and Model B(left)*



The model accuracy can be influenced by not only the overall accuracy score but the other parameters like precision & recall, f1-score and confusion matrix. Also, Receiver Operating Characteristic or the ROC Curve is a plot that provide insights into the diagnostic ability of the binary classifier.

Confusion matrices of both the ANN models were shown in the Figure 5.11, prediction of tumor correspondences on the validation dataset has been improved. Model could predict 2 out of 8 landmarks accurately, Model B could predict 3 out of 8 ground truth landmarks

*Figure 5.11: Confusion matrix of Model A (left) and Model B(right)*



As mentioned in the Research design section, model accuracy measures by the precision and recall and f1 scores as well. While the precision of the moth Models A and B are unchanged, the recall has improved in the Model B with 0.38 ( Model A recall score being the 0.25 ) were shown in Figure 5.12

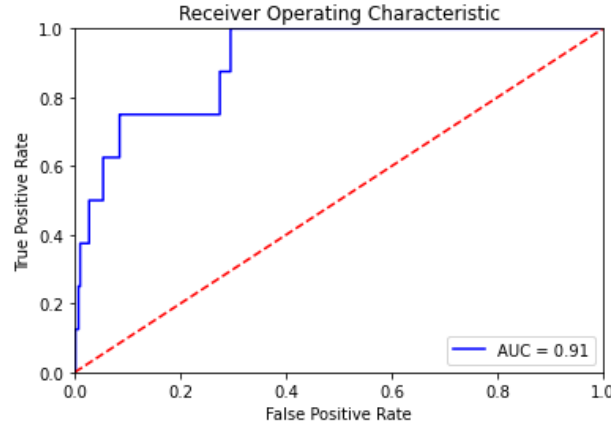
*Figure 5.12: Precision recall f1-score and support of Model A (left) and Model B(right).*

	precision	recall	f1-score	support
0	1.00	0.82	0.90	1532742
1	0.00	0.25	0.00	8
accuracy			0.82	1532750
macro avg	0.50	0.54	0.45	1532750
weighted avg	1.00	0.82	0.90	1532750

	precision	recall	f1-score	support
0	1.00	0.98	0.99	1686825
1	0.00	0.38	0.00	8
accuracy			0.98	1686833
macro avg	0.50	0.68	0.50	1686833
weighted avg	1.00	0.98	0.99	1686833

ROC Curve of Model B was captured to analyse the spread of predicted class of istumor, and all the samples are falling towards the upper left corner side which is an indication of prediction likeliness is towards accurate. The ROC curve for the Model B is shown in the Figure 5.13

*Figure 5.13: ROC Curve for the ANN MLP Keras Sequential Model B*



## 6 Results and Discussion

### 6.1 Major Findings

As the dataset was provided with ground truth landmarks, the appropriate choice would be to apply supervised learning models (Momina Masood et.al., 2021) . The following models were applied on the binary label classification (Istumor yes or no).

- Nonlinear SVM model
- Random Forest algorithm
- Decision trees
- ANN MLP using keras (Sequential)

Due to slow processing limitation on the first two models, the dataset was curtailed so that each NIfTI file will only have 10mm x10mm x 10mm slice around the tumor landmark (5 mm radius). In case of the decision tree models, due to the fact it can handle large datasets, 26 % of full dataset was used. For the sequential model, two models were attempted Model A with dataset of 35 NIfTI files out of 280 available files which is 13% (56259211 rows) of full dataset. The original dataset consists of over 428mn rows (424893869 rows). Of all the models, Sequential MLP Model B has achieved good results in terms of all the accuracy measures

consideration though the model accuracy is less, testing accuracies and confusion matrix. Though the accuracy scores of SVM, Random Forest and Decision Tree classifications were high (shown in Table 6.2.), due to the only lesion part (10mm x 10mm x 10mm) feeding to model, it cannot be relied on those models' accuracy. The sequential model A has achieved 91.89 accuracy with 13% of dataset (35 files) with confusion matrix of recall 0.25 for the istumor 1 and overall weighted average recall 0.82 and f1-score 0.90. Sequential model B has achieved 89.27 accuracy with full data set and 25% data augmentation with confusion matrix of recall 0.38 for the istumor 1 and overall weighted average recall 0.98 and f1-score 0.99 hence the model B was considered for the prediction of spatial coordinates of possible tumor correspondences.

*Figure 6.1: Dataset file composition break up table*

Model	Accuracy score
SVM nonlinear classifier	99.86
Random Forest	99.85
Decision Tree	99.99
Sequential Model A	91.89
Sequential Model B	89.27

## 6.2 Conclusion

A novel attempt had been made on spatial correspondences of tumor registration using 3d MRI medical images with the limited system configuration, cross-sectional time horizon. The pre-processing activity and classification labelling is one of key activity that had been performed along with identifying the keras multi-layer perceptron for binary label as a best model that suits the MR images that has ground truth landmarks provided dataset. Most of the current brain tumor predictions use the CNN that automatically extracts the features and categorize but there is not much evidence of prediction of the spatial tumor correspondences across brain without ground truth landmarks provision (Bhakti Baheti et. al 2021). Further study can be conducted to run the full dataset with thorough data augmentation and use the validation dataset provided by the BRaTS Challenge team provision (Bhakti Baheti et. al 2021) for deeper experimental results

The model results also need to be cross checked with the current best results that are available with the BRaTS team (Bhakti Baheti et. al 2021) as it will help to improve further and conduct future studies

### 6.3 Recommendations

Provide the system to the current novice users (Diagnostic Radiography final year students at SHU) to evaluate the post operative MRI scans files that are in the NIfTI (nii) file format to predict the tumor correspondences

### 6.4 Limitations & Future Scope

Current study has pre-processed and converted the NIfTI files into csv flat files by introducing the istumor class label and mapped the ground truth landmarks provided in the csv format in the dataset. In this regard, options of CNN algorithm can be explored if mapping can be done rather than converting the NIfTI files into csv format

Also, current study has combined the pre-operative and post operative scans and the models were not learnt the distinction between the scans. So, a variant can also be tried to predict the tumor correspondences on post operative scans by introducing the distinction between the scans

As an immediate future study, generate predicted landmark coordinates for the pre-operative scans that are provided as part of the validation dataset and share with BRaTS team to cross check the prediction accuracy with ground truth accuracy (BRaTS team has not provided the ground truth landmarks for the pre-operative scans in the validation set)

## 7 References

Bhakti Baheti, Diana Waldmannstetter, Satrajit Chakrabarty, Hamed Akbari, Michel Bilello, Benedikt Wiestler, Julian Schwarting, Evan Calabrese, Jeffrey Rudie, Syed Abidi, Mina Mousa, Javier Villanueva-Meyer, Daniel S. Marcus, Christos Davatzikos, Aristeidis Sotiras, Bjoern Menze, and Spyridon Bakas. "The brain tumor sequence registration challenge: Establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients" .

Bhakti, B., Diana, W., Hamed, A., & Satrajit, C. (2021). The brain tumor sequence registration challenge: Establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients. Retrieved from [https://www.researchgate.net/publication/357047024\\_The\\_Brain\\_Tumor\\_Sequence\\_Reg-  
istration\\_Challenge\\_Establishing\\_Correspondence\\_between\\_Pre-Operative\\_and\\_Follow-  
up\\_MRI\\_scans\\_of\\_diffuse\\_glioma\\_patients](https://www.researchgate.net/publication/357047024_The_Brain_Tumor_Sequence_Registration_Challenge_Establishing_Correspondence_between_Pre-Operative_and_Follow-up_MRI_scans_of_diffuse_glioma_patients)

Brain tumor sequence registration (brats-reg) challenge. Retrieved from <https://ipp.cbica.upenn.edu/categories/brats22>

Cancer.Net Editorial Board. (2021). Brain tumor: Statistics. Retrieved from [https://www.cancer.net/cancer-types/brain-  
tumor/statistics#:~:text=Brain%20tumors%20account%20for%2085,year%20in%20the%20U  
nited%20States.](https://www.cancer.net/cancer-types/brain-tumor/statistics#:~:text=Brain%20tumors%20account%20for%2085,year%20in%20the%20United%20States.)

CBICA image processing portal. (2022).

Coordinate systems. Retrieved

from [https://www.slicer.org/wiki/Coordinate\\_systems#:~:text=The%20anatomical%20coordinate%20system%20is,whose%20brain%20is%20being%20scanned.](https://www.slicer.org/wiki/Coordinate_systems#:~:text=The%20anatomical%20coordinate%20system%20is,whose%20brain%20is%20being%20scanned.)

Helen Phillips. (2006). Introduction: The human brain. Retrieved

from <https://institutions.newscientist.com/article/dn9969-introduction-the-human-brain/#:~:text=The%20brain%20is%20the%20most,billion%20nerve%20cells%2C%20or%20neurons.>

Larry Abbott. (2020). Beauty and the brain. Retrieved

from <https://www.simonsfoundation.org/2020/02/05/beauty-and-the-brain/#:~:text=The%20brain%20is%20beautiful%20because,more%20slowly%20than%20other%20species.>

Mri viewer [computer software]

R F Kilcoyne, M L Richardson, B A Porter, D O Olson, T K Greenlee, W Lanzer. Magnetic resonance imaging of soft tissue masses. Retrieved

from <https://pubmed.ncbi.nlm.nih.gov/3342555/#:~:text=MRI%20is%20ideally%20suited%20for,image%20directly%20in%20any%20plane.>

Tanya Lewis, Ashley P. Taylor. (2021). Human brain: Facts, functions & anatomy. Retrieved

from <https://www.livescience.com/29365-human-brain.html>

Zawn Villines. (2020). Organs of the body and their functions . Retrieved

from <https://www.medicalnewstoday.com/articles/organs-in-the-body>



- Barbara Zitova. (2018). Mathematical approaches for medical image registration. Retrieved from <https://www.sciencedirect.com/science/article/pii/B9780128012383999902>
- Bhakti Baheti, Diana Waldmannstetter, Satrajit Chakrabarty, Hamed Akbari, Michel Bilello, Benedikt Wiestler, Julian Schwarting, Evan Calabrese, Jeffrey Rudie, Syed Abidi , Mina Mousa, Javier Villanueva-Meyer, Daniel S. Marcus , Christos Davatzikos , Aristeidis Sotiras, Bjoern Menze, and Spyridon Bakas. (2021). The brain tumor sequence registration challenge: Establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients. Retrieved from <https://www.med.upenn.edu/cbica/brats-reg-challenge/>
- Francisco Javier Díaz-Pernas, Mario Martínez-Zarzuela , Míriam Antón-Rodríguez and David González-Ortega. (2021). A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. Retrieved from <https://doi.org/10.3390/healthcare9020153>
- Hannah Snyder. (2019). Journal of business research. *Volume 104*, Pages 333-339.
- Momina Masood , Tahira Nazir , Marriam Nawaz , Awais Mehmood , Junaid Rashid ,Hyuk-Yoon Kwon , Toqeer Mahmood and Amir Hussain. (2021). A novel deep learning method for recognition and classification of brain tumors from MRI images . Retrieved from <https://doi.org/10.3390/diagnostics11050744>
- Nilesh Bhaskarrao Bahadure, Arun Kumar Ray and Har Pal Thethi. (2017). Mage analysis for MRI based brain tumor detection and

feature extraction using biologically inspired BWT and SVM. Retrieved from <https://www.hindawi.com/journals/ijbi/2017/9749108/>

P Gokila Brindha<sup>1</sup>, M Kavinraj , P Manivasakam , P Prasanth. (2021). Brain tumor detection from MRI images using deep learning techniques .(. Sci. Eng. 1055 012115)

David Phair (PhD), Kerry Warren (PhD). (2021). Saunders' research onion: Explained simply. Retrieved from <https://gradcoach.com/saunders-research-onion/>

Mark Saunders, Philip Lewis, & Adrian Thornhill. (2009). *Research methods for business students* Pearson Education Limited.

oxbridgeessays.com. (2021). Writing your dissertation methodology. Retrieved from <https://www.oxbridgeessays.com/blog/writing-dissertation-methodology/>

Sheffield Hallam University. (2020). University research ethics committee research ethics policy and procedures . Retrieved from <https://www.shu.ac.uk/research/excellence/ethics-and-integrity/policies>

thinkinsights.net. (2020). What is the difference between a methodology and a framework? Retrieved from <https://thinkinsights.net/consulting/framework-methodology/>

ukessays.com. (2021). Research onion - explanation of the concept. Retrieved from <https://www.ukessays.com/essays/psychology/explanation-of-the-concept-of-research-onion-psychology-essay.php>

Artificial intelligence (AI) vs machine learning (ML). Retrieved from <https://azure.microsoft.com/en-gb/solutions/ai/artificial-intelligence-vs-machine-learning/#introduction>

Aurélien Géron. *Hands-on machine learning with scikit-learn & TensorFlow* O'REILLY.

Christian Janiesch, Patrick Zschech, Kai Heinrich. (2021). Machine learning and deep learning.

Jason Brownlee. (2020). Classification tasks . Retrieved from <https://machinelearningmastery.com/types-of-classification-in-machine-learning/>

Lakshana, G. V. (2022). 4 ways to evaluate your machine learning model: Cross-validation techniques. Retrieved from <https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-with-python-code/>

Machine learning types. Retrieved from [https://www.researchgate.net/figure/Different-machine-learning-types-and-algorithms\\_fig6\\_330815113](https://www.researchgate.net/figure/Different-machine-learning-types-and-algorithms_fig6_330815113)

Pradeep Kumar. (2021). AI branches. Retrieved from <https://www.h2kinfosys.com/blog/what-are-the-branches-of-artificial-intelligence/>

Prashant Gupta. (2017). Cross-validation in machine learning. Retrieved from <https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f>

Types of artificial intelligence: Details that everyone should know. (2019). Retrieved from <https://www.aiottalk.com/types-of-artificial-intelligence-details-that-everyone-should-know/>

JairCervantes, XiaoouLi, WenYub, & KangLic. (2008). Support vector machine classification for large data sets via minimum enclosing ball clustering.

Khaled Alsabti, Sanjay Ranka, & Vineet Singh. CLOUDS: A decision tree classifier for large datasets. Retrieved from <https://www.aaai.org/Papers/KDD/1998/KDD98-001.pdf>

Nagesh Singh Chauhan. (2022). Decision tree algorithm, explained. Retrieved from <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>

Rushikesh Pupale. (2018). Support vector machines(SVM) — an overview. Retrieved from <https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989#:~:text=SVM%20or%20Support%20Vector%20Machine,separates%20the%20data%20into%20classes.>

Tony Yiu. (2019). Understanding random forest. Retrieved from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

Vivek Praharsha. ReLU (rectified linear unit) activation function. Retrieved from <https://iq.opengenus.org/relu-activation/>

Ben H. Thacker, Scott W. Doebeling, Francois M. Hemez, Mark C. Anderson, Jason E. Pepin, & Edward A. Rodriguez. (2004). Concepts of model verification and validation.

Connor Shorten, Taghi M. Khoshgoftaar, & Borko Furht. (2021). Text data augmentation for deep learning. Retrieved from <https://link.springer.com/article/10.1186/s40537-021-00492-0>

## Appendix A: Research Project Plan

# Development and evaluation of a Machine Learning Model on MRIs segmentation of possible brain tumour affected regions

Raveendrababu Pasumarthi  
c0028603.

## Research question, aims & objectives

The study aims to develop a machine learning model on segmentation of brain tumour affected regions in Magnetic Resonance Images (MRI) of the Brain.

### 1.1 Research question

*How to predict possible tumour affected regions of the brain from the magnetic resonance images(segmentation) using Artificial Intelligence techniques?*

### 1.2 Objectives

The objectives of the study are ...

- To explore and investigate extensively the segmentation algorithms (AI techniques as focus)
- To examine the attributes (features) to predict tumour regions from MRIs.

- To develop and train a machine learning model(s) using a suitable technique based on the examined and explored options using appropriate software tools.
- To identify the best performing model among all

### 1.3 Deliverable

- A machine learning model which can predict tumour effected regions from the MRI of the brain

## 2. Situation – Complication - resolution

### 2.1 Situation

Registration of baseline pre-operative (treatment-naïve) and follow-up brain tumour MRI scans is challenging, yet a clinically important task for a multitude of reasons. Brain tissue shows heavy deformations induced by the apparent tumour (also known as mass effect) that following its resection are relaxed due to the relieving pressure from the resected tissue. Such deformations affect the whole brain (including the lateral ventricles) and are not limited to the vicinity of the tumour. This is particularly important as the relationship of the tumour to the lateral ventricles and the deformations to the rest of the brain tissue are important factors in prognosis and treatment planning. Further changes in the peritumoral edematous/infiltrated tissue, potential tumour recurrence, as well as treatment related changes, also affect the brain tissue elasticity. The resected tissue/tumour also relates to missing correspondences, and inconsistent intensity profiles between the follow up and the baseline pre-operative scans (Bhakti, Diana, Hamed, & Satrajit, 2021).

## 2.2 Complication

Taking all the above into consideration, finding spatial correspondences between two longitudinal scans of brain tumour patients, i.e., the registration between the baseline pre-operative and follow-up MRI scans, can advance mechanistic understanding for these tumours. Specifically, for tumour infiltration and potential recurrence, further contributing in the generation of predictive modelling for related pathophysiological processes, but also in understanding biophysical dynamic and plasticity characteristic of brain tissues, as well as for neurosurgical planning (<https://ipp.cbica.upenn.edu/categories/brats22>, n.d.).

## 2.3 Resolution

- Develop A machine learning model (An Artificial Intelligence System) will be developed using Brain Tumour Magnetic Resonance Images (MRI) that identify the spatial correspondences between two longitudinal MRI scans (baseline registration and follow-up registration)
- Evaluate the system against the existing ground truth results
- and document the feedback for further enhancements
- In this process, datasets will be collected from the website [ipp.cbica.upenn.edu](http://ipp.cbica.upenn.edu). with the consent of (Bhakti, Diana, Hamed, & Satrajit, 2021)

## 2.4 Recommendations

- Provide the system to the current novice users (Diagnostic Radiography final year students at SHU) to evaluate with their current tumour visualisation process

### **Dataset**

The datasets were collected through request approval from the (Bhakti, Diana, Hamed, & Satrajit, 2021) and the download can be found in the SHU cloud at below path

[https://sheffieldhallam-my.sharepoint.com/:f:/g/personal/c0028603\\_hallam\\_shu\\_ac\\_uk/Enf86iYZVudCgU5UOBL5t2UBrydujNVSwvcl9HwyeOhU6A?e=c0MRvR](https://sheffieldhallam-my.sharepoint.com/:f:/g/personal/c0028603_hallam_shu_ac_uk/Enf86iYZVudCgU5UOBL5t2UBrydujNVSwvcl9HwyeOhU6A?e=c0MRvR)

### 3. Research Design

This section talks about the options and the potential choices of the key building blocks (development methods, software tools, test data, environment) of a software project. On high level, the following key building blocks will be evaluated in depth and carefully considered as an alternative.

- Programming Languages for the Machine learning model
- Machine learning algorithms
- User interface design tools
- Software development methodology choices
- Project Management Tools
- Test Data options
- Environment Set-up

#### 3.1 Programming Languages for Machine Learning Model Implementation

The key choices of programming languages to develop machine learning model as these were the languages taught as part of the Big Data Analytics course

1. Python
2. R language

The above mentioned two languages are of open-source type with wider community base and new libraries every day added for the data science and artificial intelligence area. As per the industry, R is more suits for the statistical analysis whereas Python dominates as an undisputed option for the machine learning implementations. The following diagram shows the popularity of Python as an automation and machine learning implementation language. As I have hands on experience and due to its vast open sci-kit libraries and frameworks, Python is shortlisted as a choice. However, the selection will be carefully discussed and chosen as per the supervisor advice and guidance.



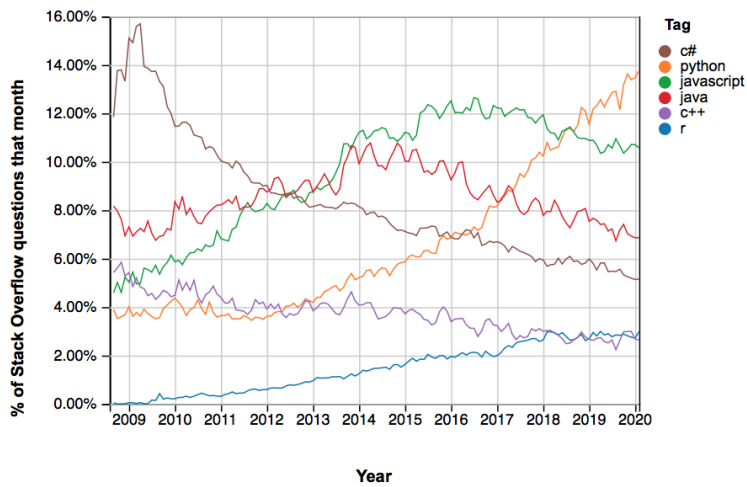


Figure 2. timeline chart of discussion forum questions on programming languages

(source: google images keyword: python popularity)

### 3.2 Software development methodology choices

The dissertation study will adopt the blend of traditional waterfall model and the agile scrum model. The blend model will help to break the major activities into phases as per the waterfall principle, and each phase is time boxed as per the scrum model in the form sprints with each sprint is of two weeks' time. The combination model of software development life cycle ensures the completion of current activity and kick start of new activity as per the planned time schedules. Timeline slippage of any sprint will allow performing multiple activities in next sprint in parallel with new activity by ensuring all the activities at least start as per the plan.

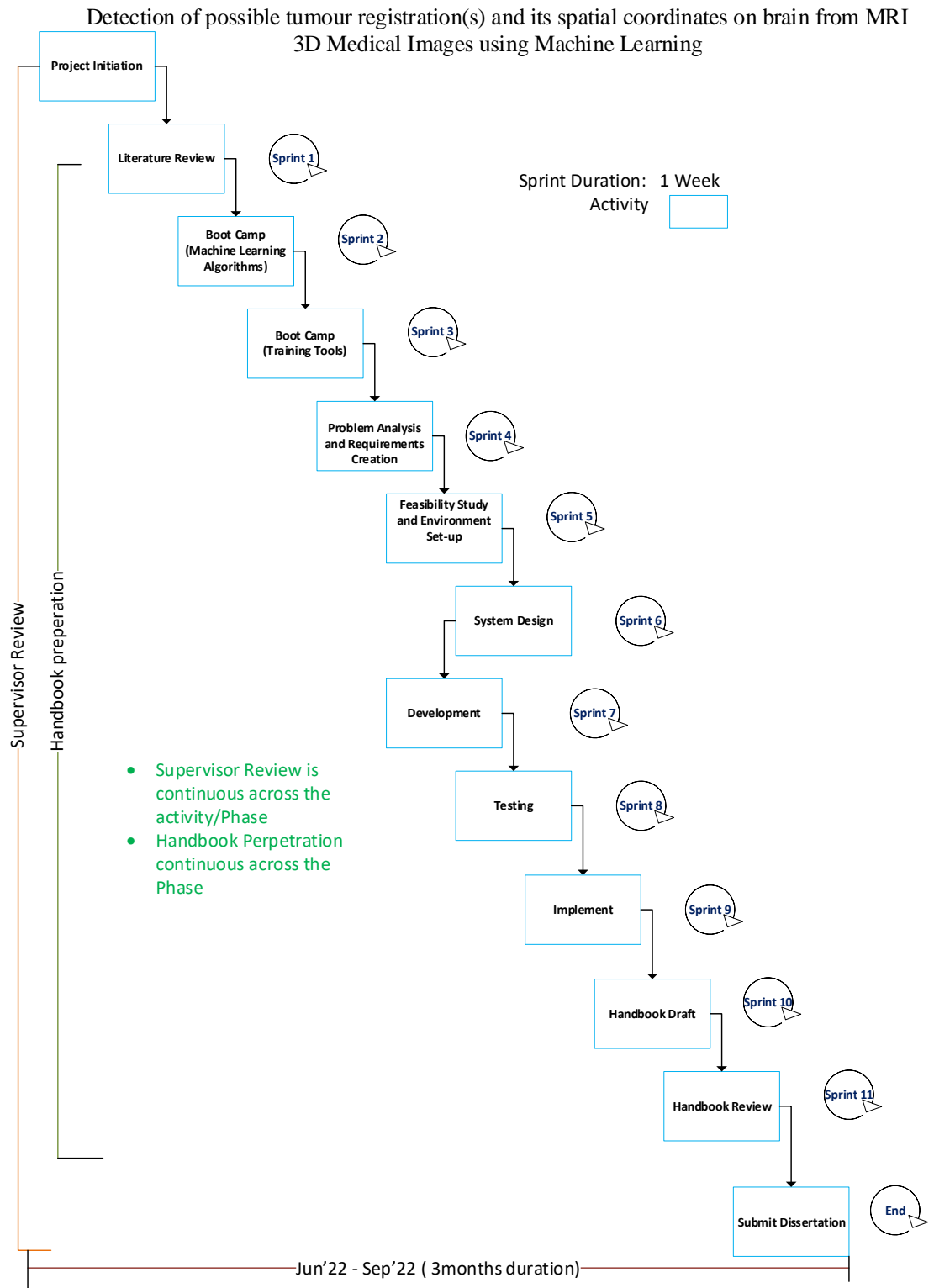


Figure 3. adoption of blended development model (waterfall and scrum)

### 3.3 Project Management Tools

The following list of tools will be used as part of project management throughout the dissertation process.

<b>Project Management Component</b>	<b>Tool</b>	<b>Purpose</b>	<b>Procurement Source</b>
Task and Schedule Management	Microsoft Project Planner	Task creation, tracking, resources, schedules, milestones update and Gantt charts	SHU account Azure portal (software section)
Project Diagrams	Microsoft Visio	Architectures, flow charts, use cases and other business and project diagrams	SHU account Azure portal (software section)
Documents and Presentations	Microsoft Office 365	Content development and presentation	SHU Account Office 365 login
Communications (Audio, Telepresence, Content Sharing)	Zoom	Virtual telephonic and telepresence and screen sharing	SHU domain login (my.shu.ac.uk)
	Collaborate	Virtual telephonic and telepresence and screen sharing alternative	Blackboard
Communications (email)	Email	Official email communication between student and supervisor and other stake holders	SHU Email
Test Management	Microsoft Excel	Test case design, defect tracking, other testing artefacts maintenance	SHU Account Office 360 login

Table 1. Project management tools

### 3.4 Environment set-up

Environment is the collection and combination of hardware resources and software tools where actual execution of development, testing, code deployment will take place. Selection and setting-up of the suitable and affordable environment will be finalised with the consent of supervisor. However, based on the environments that were availed as part of the Big Data Analytics course, the following environment could be used.

- **Azure Labs** – Microsoft Cloud Platform

## 4 Ethics, risks and Issues

### 4. 1 RISKS

<b>Risk</b>	<b>Description</b>	<b>Mitigation</b>
Scope of the Problem & Timelines	Achievability of the chosen problem within the given timeframe	Scope of the problem will be revisited to fit into given timelines
Neural Network/Deep Learning Techniques selection	Some of algorithms need very high-end environment and laboratory facilities	Choice of the algorithm will be narrowed based the achievable environments
Dataset size and volume	HD Images consumes gigantic storage space	Volume of test data will be curtailed to the available storage space
Environment requirements	Performance issues in the Computer Lab hardware and software	Performance bottles will be mitigated with optimisation, reduction in test datasets and suitable algorithms for the given configuration

Table 2. Risk Analysis

### 4. 2 LSEPI ANALYSIS

<b>Concern</b>	<b>Issue</b>	<b>Modality</b>	<b>Mitigation</b>
<b>GDPR</b>	Data privacy breach	Legal	Adhering to the rules.

	directly or indirectly		Anonymise or mask the data
<b>PII</b>	Consuming personal identification information	Legal	Anonymise or mask the consumed data
<b>Code of Conduct</b>	Student availability on time to the supervisor	Professional	Meetings will be booked well in advance

Table 3. LSEPI Analysis

## 5. Time Plan

The duration of the proposed project is of three months and the indicative start date as 1<sup>st</sup> Jun 2022 and completion date as 9<sup>th</sup> Sep 2022. The work break structure divides the tasks of each phase executed in the sprint's manner, with size of sprint as one week (for complex tasks, the size of the sprint will be of two weeks)

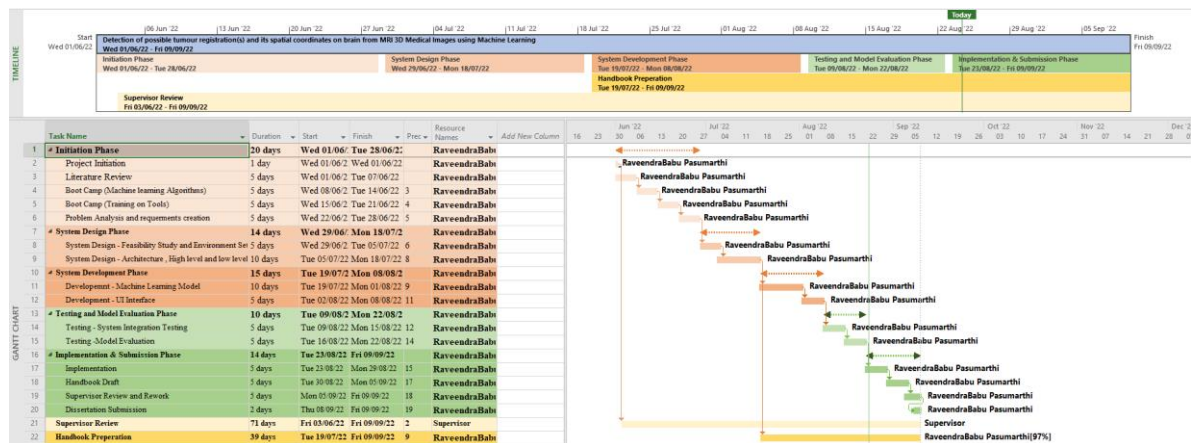


Figure 3. Gantt Chart and Timeline

## REFERENCES

- [1] G., Arnold (2015). 5 Reasons Why Images Speak Louder Than Words. *Published on September 29, 2015*, from <https://www.linkedin.com/pulse/5-reasons-why-images-speak-louder-than-words-gabe-arnold/>
- [2] Shen, C., Kasra, M., Pan, W., Bassett, G. A (G., Arnold, 2015)., Malloch, Y., & O'Brien, J. F. (2019). Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New media & society*, 21(2), 438-463.
- [3] Tariq, S., Lee, S., Kim, H., Shin, Y., & Woo, S. S. (2019, April). Gan is a friend or foe? a framework to detect various fake face images. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (pp. 1296-1303).
- [4] Ghosh, A., Kulharia, V., Namboodiri, V. P., Torr, P. H., & Dokania, P. K. (2018). Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8513-8521).
- [5] Brock, A., Lim, T., Ritchie, J. M., & Weston, N. (2016). Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*.
- [6] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, 1.
- [7] (<https://ipp.cbica.upenn.edu/categories/brats22>, n.d.)
- [8] (Bhakti, Diana, Hamed, & Satrajit, 2021)

## Appendix B: Completed Research ethics form



### RESEARCH ETHICS CHECKLIST FOR STUDENTS (SHUREC7)

This form is designed to help students and their supervisors to complete an ethical scrutiny of proposed research. The SHU [Research Ethics Policy](#) should be consulted before completing the form.

Answering the questions below will help you decide whether your proposed research requires ethical review by a Designated Research Ethics Working Group.

The final responsibility for ensuring that ethical research practices are followed rests with the supervisor for student research.

Note that students and staff are responsible for making suitable arrangements for keeping data secure and, if relevant, for keeping the identity of participants anonymous. They are also responsible for following SHU guidelines about data encryption and research data management.

The form also enables the University and Faculty to keep a record confirming that research conducted has been subjected to ethical scrutiny.

For student projects, the form may be completed by the student and the supervisor and/or module leader (as applicable). In all cases, it should be counter-signed by the supervisor and/or module leader, and kept as a record showing that ethical scrutiny has occurred. Students should retain a copy for inclusion in their research projects, and staff should keep a copy in the student file.

Please note if it may be necessary to conduct a health and safety risk assessment for the proposed research. Further information can be obtained from the Faculty Safety Co-ordinator.

#### General Details

Name of student	Raveendrababu Pasumarthi
SHU email address	<a href="mailto:c0028603@my.shu.ac.uk">c0028603@my.shu.ac.uk</a>
Course or qualification (student)	MSc Big Data Analytics
Name of supervisor	Dr. Hemlata Sharma
email address	<a href="mailto:hs9000@exchange.shu.ac.uk">hs9000@exchange.shu.ac.uk</a>
Title of proposed research	Detection of possible tumour registration(s) and its spatial coordinates on brain from MRI 3D Medical Images using ANN (Multi-Layer Perceptron)

Proposed start date	Jun 2022
Proposed end date	Sep 2022
Brief outline of research to include, rationale & aims (250-500 words)	<p>The study aims to develop a machine learning model on segmentation of brain tumour affected regions in Magnetic Resonance Images (MRI) of the Brain.</p> <p><b>1.1 Research question</b></p> <p><i>How to predict possible tumour affected regions of the brain from the magnetic resonance images(segmentation) using Artificial Intelligence techniques?</i></p> <p><b>1.2 Objectives</b></p> <p>The objectives of the study are ...</p> <ul style="list-style-type: none"> <li>• To explore and investigate extensively the segmentation algorithms (AI techniques as focus)</li> <li>• To examine the attributes (features) to predict tumour regions from MRIs.</li> <li>• To develop and train a machine learning model(s) using a suitable technique based on the examined and explored options using appropriate software tools.</li> <li>• To identify the best performing model among all</li> </ul> <p><b>1.3 outcome</b></p> <p>Prediction, detection of possible tumor affected regions of brain using MR Images using segmentation and AI are being extensively under study and gained the ground in exponential pace thus contributing to medicine domain. Identifying possible spatial correspondences across the longitudinal scans due to the deformation and change of appearance of brain</p>



	<p>tissue is still a challenge. A novel attempt has been made to predict the possible tumor spatial coordinates on highly curated (by domain experts) 3d MRI dataset (from BRaTS Challenge) that has base(pre-operative) and follow up scan's combination for at least 140 patients (total of 280 image files of t1ce modality 3d MRI scans of NIfTI format) that only supplied the ground truth landmark details of tumor for each scan. It was successfully preprocessed the NIfTI format files to flat files by converting anatomical coordinate system to world coordinate system and applied istumor label based on the landmark files provided along with dataset. Results using keras Multi-Layer Perceptron (MLP) for binary label has achieved an accuracy of 91.89 with 13% of dataset (35 files) with confusion matrix of recall 0.25 for the istumor 1 and overall weighted average recall 0.82 and f1-score 0.90</p>
<p>Where data is collected from individuals, outline the nature of data, details of anonymization, storage and disposal procedures if required (250-500 words)</p>	<p>Dataset</p> <ul style="list-style-type: none"> <li>• datasets will be collected from the website <a href="http://ipp.cbica.upenn.edu">ipp.cbica.upenn.edu</a>. with the consent of (Bhakti, Diana, Hamed, &amp; Satrajit, 2021)</li> </ul> <p>The datasets were collected through request approval from the (Bhakti, Diana, Hamed, &amp; Satrajit, 2021) and the download can be found in the SHU cloud at below path  <a href="https://sheffieldhallam-my.sharepoint.com/:f/g/personal/c0028603_hallam_shu_a_c_uk/Enf86iYZVudCgU5UOBL5t2UBrydujNVSwvc19HwyeOhU6A?e=c0MRvR">https://sheffieldhallam-my.sharepoint.com/:f/g/personal/c0028603_hallam_shu_a_c_uk/Enf86iYZVudCgU5UOBL5t2UBrydujNVSwvc19HwyeOhU6A?e=c0MRvR</a></p>

**1. Health Related Research Involving the NHS or Social Care / Community Care or the Criminal Justice Service or with research participants unable to provide informed consent**

Question	Yes/No
----------	--------

1. Does the research involve?	No
<ul style="list-style-type: none"> <li>- Patients recruited because of their past or present use of the NHS or Social Care</li> <li>- Relatives/carers of patients recruited because of their past or present use of the NHS or Social Care</li> <li>- Access to data, organs or other bodily material of past or present NHS patients</li> <li>- Foetal material and IVF involving NHS patients</li> <li>- The recently dead in NHS premises</li> <li>- Prisoners or others within the criminal justice system recruited for health- related research*</li> <li>- Police, court officials, prisoners or others within the criminal justice system*</li> <li>- Participants who are unable to provide informed consent due to their incapacity even if the project is not health related</li> </ul>	

2. Is this a research project as opposed to service evaluation or audit? <i>For NHS definitions please see the following website <a href="http://www.hra.nhs.uk/documents/2013/09/defining-research.pdf">http://www.hra.nhs.uk/documents/2013/09/defining-research.pdf</a></i>	No
---	----

If you have answered **YES** to questions **1 & 2** then you **must** seek the appropriate external approvals from the NHS, Social Care or the National Offender Management Service (NOMS) under their independent Research Governance schemes. Further information is provided below.

NHS <https://www.myresearchproject.org.uk/Signin.aspx>

\* All prison and probation projects also need HM Prison and Probation Service (HMPPS) approval. Further guidance at: <https://www.myresearchproject.org.uk/help/hlphmpps.aspx>

**NB** FRECs provide Independent Scientific Review for NHS or SC research and initial scrutiny for ethics applications as required for university sponsorship of the research. Applicants can use the NHS proforma and submit this initially to their FREC.

## 2. Research with Human Participants

Question	Yes/No
Does the research involve human participants? This includes surveys, questionnaires, observing behaviour etc.	No
Question	Yes/No
1. <i>Note If YES, then please answer questions 2 to 10 - If NO, please go to Section 3</i>	
2. Will any of the participants be vulnerable? <i>Note: 'Vulnerable' people include children and young people, people with learning disabilities, people who may be limited by age or sickness, etc. See definition on website</i>	No
3. Are drugs, placebos or other substances (e.g. food substances, vitamins) to be administered to the study participants or will the study involve invasive, intrusive or potentially harmful procedures of any kind?	No
4. Will tissue samples (including blood) be obtained from participants?	No
5. Is pain or more than mild discomfort likely to result from the study?	No
6. Will the study involve prolonged or repetitive testing?	No
7. Is there any reasonable and foreseeable risk of physical or emotional harm to any of the participants? <i>Note: Harm may be caused by distressing or intrusive interview questions, uncomfortable procedures involving the participant, invasion of privacy, topics relating to highly personal information, topics relating to illegal activity, etc.</i>	No
8. Will anyone be taking part without giving their informed consent?	No
9. Is it covert research? <i>Note: 'Covert research' refers to research that is conducted without the knowledge of participants.</i>	No
10. Will the research output allow identification of any individual who has not given their express consent to be identified?	No

If you answered **YES only** to question **1**, the checklist should be saved and any course procedures for submission followed. If you have answered **YES** to any of the other questions you are **required** to submit a SHUREC8A (or 8B) to the FREC. If you answered **YES** to question **8** and participants cannot provide informed consent due to their incapacity you must obtain the appropriate approvals from the NHS research governance system. Your supervisor will advise.

### 3. Research in Organisations

Question	Yes/No
1. Will the research involve working with/within an organisation (e.g. school, business, charity, museum, government department, international agency, etc)?	No
2. If you answered YES to question 1, do you have granted access to conduct the research? <i>If YES, students please show evidence to your supervisor. PI should retain safely.</i>	No
3. If you answered NO to question 2, is it because: A. you have not yet asked B. you have asked and not yet received an answer C. you have asked and been refused access.  <i>Note: You will only be able to start the research when you have been granted access.</i>	

### 4. Research with Products and Artefacts

Question	Yes/No
1. Will the research involve working with copyrighted documents, films, broadcasts, photographs, artworks, designs, products, programmes, databases, networks, processes, existing datasets or secure data?	Yes
2. If you answered YES to question 1, are the materials you intend to use in the public domain?  <i>Notes: 'In the public domain' does not mean the same thing as 'publicly accessible'.</i> <ul style="list-style-type: none"> <li>Information which is 'in the public domain' is no longer protected by copyright (i.e. copyright has either expired or been waived) and can be used without permission.</li> <li>Information which is 'publicly accessible' (e.g. TV broadcasts, websites, artworks, newspapers) is available for anyone to consult/view. It is still protected by copyright even if there is no copyright notice. In UK law, copyright protection is automatic and does not require a copyright statement, although it is always good practice to provide one. It is necessary to check the terms and conditions of use to find out exactly how the material may be reused etc.</li> </ul> <i>If you answered YES to question 1, be aware that you may need to consider other ethics codes. For example, when conducting Internet research, consult the code of the Association of Internet Researchers; for educational research, consult the Code of Ethics of the British Educational Research Association.</i>	No
3. If you answered NO to question 2, do you have explicit permission to use these materials as data? <i>If YES, please show evidence to your supervisor.</i>	Yes
4. If you answered NO to question 3, is it because: A. you have not yet asked permission B. you have asked and not yet received and answer C. you have asked and been refused access.  <i>Note You will only be able to start the research when you have been granted permission to use the specified material.</i>	

### Adherence to SHU policy and procedures

<b>Personal statement</b>	
I can confirm that:	
YES I have read the Sheffield Hallam University Research Ethics Policy and Procedures	
YES I agree to abide by its principles.	
<b>Student</b>	
Name: Raveendrababu Pasumarthi	Date: 04/09/2022
 Raveendrababu Pasumarthi	
Signature:	
<b>Supervisor or other person giving ethical sign-off</b>	
I can confirm that completion of this form has not identified the need for ethical approval by the FREC or an NHS, Social Care or other external REC. The research will not commence until any approvals required under Sections 3 & 4 have been received.	
Name: Dr. Hemlata Sharma	Date: 04/09/22
Signature: 	

Please ensure the following are included with this form if applicable, tick box to indicate:

	Yes	No	N/A
Research proposal if prepared previously	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Any recruitment materials (e.g. posters, letters, etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Participant information sheet	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Participant consent form	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Details of measures to be used (e.g. questionnaires, etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Outline interview schedule / focus group schedule	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Debriefing materials	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Health and Safety Project Safety Plan for Procedures	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

## Appendix C: Publication Procedure Form





College of Business,  
Technology and  
Engineering

### Research Skills and Dissertation Module (55-706556).

#### PUBLICATION PROCEDURE FORM

In this module, while you create your own research question or topic area, your supervisor makes a significant intellectual contribution to this work as the research progresses. Your supervisor will make the decision on whether your work merits publication based on the quality of the work you have produced. Your supervisor will co-author the paper for publication with you and your supervisor will both be listed as authors. You are required to sign the declaration below to confirm that you understand and will follow this procedure.

Declaration:

I, Raveendrababu Pasumarthi confirm that I understand will comply with the Publication Procedure outlined in the Module Handbook and the Blackboard Site.		
Student:	 Raveendrababu Pasumarthi	Date 04/09/2022
Supervisor:	 Hemlata Sharma	Date 04/09/22

## Appendix D: Dataset Source

[https://sheffieldhallam-my.sharepoint.com/:f:/g/personal/c0028603\\_hallam\\_shu\\_ac\\_uk/Enf86iYZVudCgU5UOBL5t2UBrydujNVSwvcl9HwyeOhU6A?e=1JD1Lk](https://sheffieldhallam-my.sharepoint.com/:f:/g/personal/c0028603_hallam_shu_ac_uk/Enf86iYZVudCgU5UOBL5t2UBrydujNVSwvcl9HwyeOhU6A?e=1JD1Lk)

## **Appendix E: Source Code**

[https://sheffieldhallam-my.sharepoint.com/:f:/g/personal/c0028603\\_hallam\\_shu\\_ac\\_uk/EqcuA7Cq0lZlAn9KvpaQdoBPzb7P41jyYGiuTq5XS8pZg?e=xdE7rG](https://sheffieldhallam-my.sharepoint.com/:f:/g/personal/c0028603_hallam_shu_ac_uk/EqcuA7Cq0lZlAn9KvpaQdoBPzb7P41jyYGiuTq5XS8pZg?e=xdE7rG)