

**RADIOGENOMIC ANALYSIS OF GLIOBLASTOMA WITH DEEP
LEARNING TECHNIQUES**

by

WANG YICHONG

(B.S., National University of Singapore)

A THESIS SUBMITTED FOR THE DEGREE OF

BACHELOR OF ENGINEERING

in

BIOMEDICAL ENGINEERING

in the

GRADUATE DIVISION

of the

NATIONAL UNIVERSITY OF SINGAPORE

2022

Supervisor:

Associate Professor Zhiwei Huang

Examiners:

Associate Professor QIU Anqi

To my delicious friends

Acknowledgments

Firstly, I'm extremely grateful to prof Huang to be my supervisor in NUS to guide me about this project. Let me use this lab space. The second I want to say thank you is my senior Karthik, he give me lots of resources and some really key articles in this area. Though his guide, I learn and write my first machine learning program, U-Net code in Singapore. It is a really break though in my life. Really thank you! Meanwhile, I want to thank you other guys in this lab, especially Shuchi, Bocheng Qiu and Xiyuan Zhang. They helped me a lot when I leave my native place and feel culture shock. Liu Yudi, who from lab next to me, give me a bunch of help about radiogenomics. I cannot ever thank her enough. Also, I cannot forget to thank my family, my parents. Without their money, I can not be here to study. Thanks!

Apart from friends from NUS, I must thank Department of Biomedical Engineering in Zhejiang University. Thanks them to establish this project to let me have a chance come to Singapore to finish my FYP. Finally, I must thank my girl friend in China. Thanks for her staying in a long-distance relationship.

Thanks everyone who help me before!

Contents

Acknowledgments	ii
Abstract	vi
List of Figures	viii
List of Tables	xi
1 Introduction	1
1.1 Glioblastoma	1
1.1.1 Subtypes of Glioblastoma	1
1.1.2 Diagnosis of glioblastoma	6
1.1.3 Therapy of glioblastoma	6
1.2 Radiogenomics	8
1.3 Segmentation Algorithms for Glioblastoma	10
1.3.1 Region-based Approach	10
1.3.2 Edge-based Approach	11
1.3.3 Classification Approaches	11
1.3.4 Atlas-based Approaches	13
1.4 Feature Extraction	13
2 Method	15
2.1 U-Net	15
2.2 Feature Extraction	16
2.2.1 Intensity Based Features	17
2.2.2 Texture Based Features	19
2.2.3 Shape Based Features	22

2.2.4	Pyradiomics	22
2.3	Classifier	23
2.3.1	Support Vector Machines (SVM)	23
2.3.2	K-nearest neighbors Algorithms (KNN)	24
2.3.3	Decision Trees	25
2.3.4	Random Forest	26
2.3.5	Neural Network Classification Models	27
2.4	Mutual Information	27
3	Experiment	30
3.1	Segmentation	30
3.1.1	Data preprocessing	30
3.1.2	Training	32
3.2	Feature Extraction	34
3.3	Classifier	34
3.4	Mutual Information Analysis	35
4	Result	36
4.1	Segmentation	36
4.1.1	T1	37
4.1.2	T2	38
4.1.3	FLAIR	39
4.1.4	Stack	40
4.2	Classification	42
4.3	Mutual Information	48
4.3.1	Graphic User Interference (GUI)	51
5	Discussion	53
5.1	Segmentation	53
5.1.1	Test Part Performance worst than Train Part and Validation Part	53
5.1.2	Other Possible Direction	54
5.2	Feature Extraction	54
5.3	Mutual Information analysis	54

5.4	Significance	55
5.5	Outlook	56
	Bibliography	58
	A Code	64
A.1	U-Net	64
A.2	Read NIfTI file and convert it to PNG file	67
A.3	Feature Extraction	68

Abstract

Radiogenomic Analysis of Glioblastoma with Deep learning Techniques

by

Wang Yichong

Bachelor of Engineering in Biomedical Engineering

National University of Singapore

The most important feature of glioblastoma is intratumoral heterogeneity, that is, a high degree of heterogeneity both in space and time. Even, the current study shows that different subgroups can coexist genetically and indirectly lead to different radiographic manifestations, making glioblastoma lack of personalized treatment plan unable to achieve good results. Even after radiotherapy and adjuvant chemotherapy, residual tumor, spread, and recurrence can occur after gross total resection of the apparent tumor. The emergence of gene expression profiling over the past decade has provided new directions for glioblastoma research, and increasing evidence suggests that the diagnosis of glioma requires comprehensive histological and genotypic characterization to enable more accurate diagnosis and personalized treatment.

The current mainstream personalized treatment plan for glioblastoma is based on the subtype classification of glioblastoma at the molecular level. The mainstream subtype classification method is based on DNA sequencing after the sampling of glioblastoma biological samples. It takes a long time and the cost is high. From the perspective of radiogenomics, this project uses U-Net to segment the lesions of glioblastoma MRI images for radiological feature extraction, and uses machine learning algorithms to classify glioblastoma subtypes based on DNA transcription, with non-invasive, fast and low-cost specificity. The project finally achieved a correct rate of 92% in the context of the three classifications of preneural type, classical type, and mesenchymal type determined by Wang's 2017 study.

The access code is at:

<https://github.com/PinkR1ver/Radiogenomics-on-Ivy-Gap>

<https://github.com/PinkR1ver/GBM-Farsighter>

Keywords: Glioblastoma; Radiogenomics; Radiomics; Medical Image Semantic Segmentation; U-Net

List of Figures

1.1	The process of molecular-based GBM classification: (A) GBM classification timeline and classical subtypes. (B) The relative overlap between subtypes from different classification methods [2]	2
1.2	Characteristics of IDH wild and IDH mutant subtypes [2]	4
1.3	Mainstream classification methods and specific categories of glioblastoma subtypes at the molecular level	5
1.4	: A representation of the proper workflow of a radiomic study, which includes the following steps: 1) image acquisition; 2)image processing, including noise/artifact reduction, intensity and/or orientation standardization, coregistration of themultiparametric MRI scans; 3) ROI definition using manual annotation or (semi-)automatic segmentation; 4) feature extraction basedon human-engineered (conventional radiomics) or deep-learning approaches; and 5) data analysis, involving machine/deep-learningmethods for feature selection, classification, and crossvalidation. Radiogenomics studies should ideally follow the same workflow,with genomics of glioblastoma as their endpoint.3Kazerooni et al.: Review of Glioblastoma Radiogenomics [15]	9
1.5	Conventional pipeline of region-based approach: features of one voxel and their neighbourhoods are analysed. According to similarity criteria pre-defined, voxels are clustered [16].	10
1.6	Conventional pipeline of edge-based approach: a rough delineation drew by the user is used to compute forces from neighbourhood voxel of the first contour. These forces iteratively deform the initial shape under constraints to obtain the final segmentation [16].	11

1.7	Conventional pipeline of learning machine approach: in a first part, the algorithm is trained with several pre-segmented volume and learned features of voxels labelled previously. Then, the initial volume is classified according to rules learned previously by the algorithm [16].	12
1.8	Conventional pipeline of atlas-based approach: an atlas segmented previously is used to find similarities with the studied volume. Voxel are thus labelled in comparison with the atlas to obtain the final segmentation [16].	13
2.1	Original U-net architecture [23]	15
2.2	Classification of typical features that can be extracted from MRI images	17
2.3	Different Parameters to calculate grayscale co-occurrence matrix (GLCM) [25]	20
2.4	Features in <i>Pyradiomics</i>	23
2.5	Decision Trees	25
2.6	Random Forest	26
2.7	Relationship between mutual information and entropy	28
3.1	Workflow of annotation [32]	31
3.2	Confusion Matrix	33
3.3	Feature Extraction	34
4.1	Left is origin MRI image, middle is ground truth, right is prediction . .	37
4.2	T1 series Segmentation Assessment	37
4.3	Left is origin MRI image, middle is ground truth, right is prediction . .	38
4.4	T2 series Segmentation Assessment	38
4.5	Left is origin MRI image, middle is ground truth, right is prediction . .	39
4.6	FLAIR series Segmentation Assessment	39
4.7	Left is origin MRI image, middle is ground truth, right is prediction . .	40
4.8	F1-score tuning curve	41
4.9	Confusion Matrix for classification using different algorithms	43
4.10	Confusion Matrix for classification using different algorithms	45
4.11	Confusion Matrix for classification using different algorithms	47

4.12 Heat map of mutual information analysis of cellular tumor (CT) radiomic features and gene expression intensity. The abscissa is the common genes in the 21 GBM samples, and the ordinate is the exomic features of the extracted MRI images. In the heat map, the closer the color is to orange, the closer the relationship is, and the closer the color is to purple and black, the weaker the relationship.	49
4.13 Heat map of mutual information analysis of infiltrating tumor (IT) radiomic features and gene expression intensity. The abscissa is the common genes in the 21 GBM samples, and the ordinate is the exomic features of the extracted MRI images. In the heat map, the closer the color is to orange, the closer the relationship is, and the closer the color is to purple and black, the weaker the relationship.	50
4.14 Heat map of mutual information analysis of leading edge (LE) radiomic features and gene expression intensity. The abscissa is the common genes in the 21 GBM samples, and the ordinate is the exomic features of the extracted MRI images. In the heat map, the closer the color is to orange, the closer the relationship is, and the closer the color is to purple and black, the weaker the relationship.	51
4.15 GUI Visualization	52

List of Tables

4.1	Best Segmentation Result for T1, in epoch 17	41
4.2	Best Segmentation Result for T2, in epoch 15	41
4.3	Best Segmentation Result for FLAIR, in epoch 16	42
4.4	Best Segmentation Result for Stack, in epoch 16	42
4.5	Decision Trees Assessment	43
4.6	Random Forest Assessment	44
4.7	Ploy SVM Assessment	44
4.8	MLP Assessment	45
4.9	Decision Trees Assessment	46
4.10	Random Forest Assessment	46
4.11	Ploy SVM Assessment	46
4.12	MLP Assessment	47
4.13	Decision Trees Assessment	47
4.14	Random Forest Assessment	48
4.15	Ploy SVM Assessment	48
4.16	MLP Assessment	48

Chapter 1

Introduction

1.1 Glioblastoma

Glioblastoma (GBM) is a very aggressive adult brain tumor and the deadliest brain tumor that originates in glial cells, and in some rare cases intracranial glioblastomas have very small chance of transfer to the spine via cerebrospinal fluid. Gliomas such as glioblastoma have long been thought to originate in glial cells due to the similarities in the immunostaining of glioblastomas and glioblastomas. Recent studies have shown that astrocytes, oligodendrocyte progenitor cells and neural stem cells can also serve as cells of origin. In addition, several studies have shown that specific patterns of genetic alterations shape the clinical features of brain tumors, leading to a rapid increase in the use of these patterns for classification and diagnostic purposes in recent years. Gliomas are characterized by uncontrolled cell proliferation, diffuse infiltration, resistance to apoptosis, and genomic instability. The tumorigenesis mechanism of glioblastoma remains unclear, and many patients relapse due to ineffective treatment options. Clinical data show poor prognosis for glioblastoma, with less than 5% of patients surviving 5 years after diagnosis [1, 2].

1.1.1 Subtypes of Glioblastoma

The World Health Organization (WHO) defines adult diffuse gliomas as grade II and III astrocytomas, II and III oligodendroglomas, and grade IV glioblastomas. Glioblastoma (GBM) is grade IV, the most aggressive and lethal glioma [3]. GBM is defined as primary or secondary GBM based on clinicopathological features.

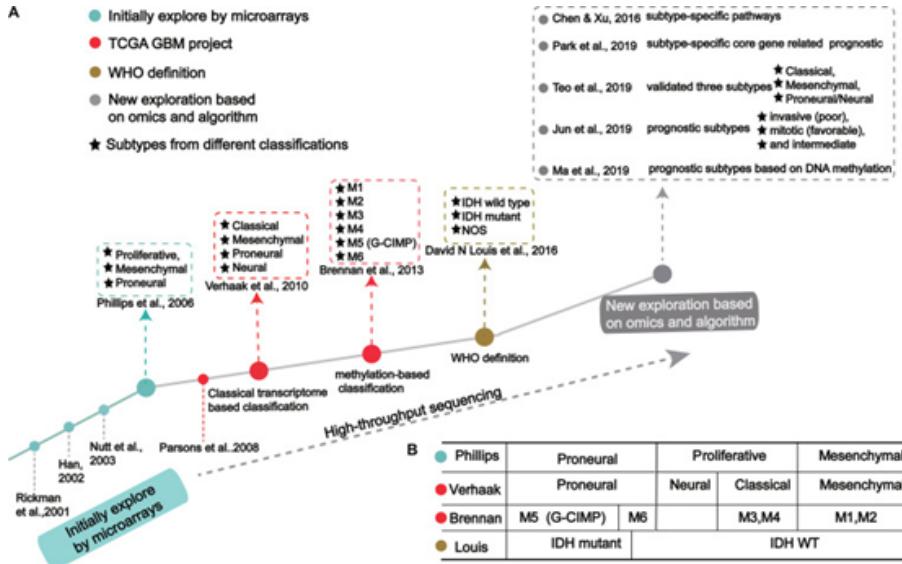


Figure 1.1: The process of molecular-based GBM classification: (A) GBM classification timeline and classical subtypes. (B) The relative overlap between subtypes from different classification methods [2]

Primary GBM starts directly at grade IV, has no signs of lower grades, and is more aggressive and more likely to affect older patients. Secondary GBM develops from astrocytomas (grade II or III gliomas) that initially grow slowly and then gradually become aggressive [4]. Histomorphological ambiguity and tumor heterogeneity pose challenges for the diagnosis, prognosis, and treatment of GBM, limiting the reproducibility of the diagnosis. GBM shows marked histological and genetic inter- and intra-tumor heterogeneity, distinct mutations, and ambiguous phenotypic and epigenetic status leading to different treatment options, and clinical outcomes also indicate genomic instability . Therefore, molecular-level-based diagnosis, patient stratification, and personalized treatment are becoming increasingly important. At present, there are three classification methods based on molecular level classification, which are based on transcription, genetic Alteration, and DNA methylation [2].

1.1.1.1 Transcription-Based Subtypes

Gene expression profiling-based classification of glioblastoma initially used microarray technology, followed by large-scale high-throughput next-generation sequencing. In 2010, the classification method proposed by Verhaak [5] is currently widely used and is divided into four subtypes: Proneural, Neural, Classical, and

CHAPTER 1. INTRODUCTION

Mesenchymal. The classification of glioblastoma based on gene transcription began in 1990. Different scholars used PCR, allele analysis and first-generation sequencing to obtain data to analyze gliomas, but there are many different types and grades of molecular markers, making the classification situation unclear [6].

In 2006, Phillips [7] pioneered the classification of glioblastoma into three subtypes, preneural, proliferative, and mesenchymal, which laid the foundation for Verhaak's classification in 2010. With the advent of large-scale, high-throughput, next-generation sequencing methods, and the advent of machine learning algorithms, complex tumor data has become increasingly accurate. Varhaak's preneural subtype of glioblastoma is more common in younger patients and is characterized by high PDGFRA gene expression and frequent IDH1 mutations. The anterior neural subtype was more likely to have better survival than the other three subtypes, but the anterior neural subtype did not differ significantly in response to chemotherapy and radiation; the neural subtype had gene expression similar to normal brain tissue mode, at the same time, more responsive to radiotherapy and chemotherapy. Neural subtypes often have neural markers such as synaptotagmin-1 (SYT1), solute carrier family 12 member 5(SLC12A5), gamma-aminobutyric acid type A receptor alpha1 subunit (GABRA1), and neurofilament light polypeptide (NEFL); classical subtypes Abnormal changes in the classic subtype of the phenotype: including chromosome 7 amplification, chromosome 10 loss, retinoblastoma-related protein (RB) pathway inactivation, etc. At the same time, Notch signaling pathway and neural precursor and stem cell markers NES is highly expressed in the Classical isoform. Clinically, patients with the classic subtype can significantly establish mortality with aggressive radiotherapy and chemotherapy; the mesenchymal subtype is characterized by extensive necrosis and inflammation, concomitantly, loss of tumor suppressor genes P53, PTEN, and NF1, and tumor High expression of genes in the necrosis factor and NF-B pathways. Despite being responsive to aggressive radiotherapy and chemotherapy, the mesenchymal subtype has the worst prognosis of all subtypes [2, 8].

In 2017, Jiguang Wang et al [9]. further divided glioblastoma into classical, proneural, and mesenchymal. They found that the Neural subtype of Verhaak was the tumor microenvironment. in non-tumor cells. Jiguang Wang's classification is based on tumor cells rather than microenvironment/non-malignant cells in tumor

entities.

1.1.1.2 Genetic Alteration-Based Subtypes

In recent years, large-scale genomic studies have revealed mutations in many tumor suppressor genes and oncogenes, significantly improving our knowledge of GBM. Specifically, mutated IDH, PTEN, and EGFR are associated with patient survival and serve as indicators for patient classification [2].

Among them, the identification of IDH mutations is an important contribution to the molecular pathology of GBM. In 2008, Parsons [10] found that IDH1 gene was mutated in a few glioblastoma samples, and Yan [11] et al. found that the survival rate of glioblastoma patients with IDH1/IDH2 mutations was higher than patients without these mutations. Numerous studies have shown that patients with IDH mutations and patients without IDH mutations have significant differences in molecular and clinical characteristics, as shown in the following Figure 1.2:

	IDH WT	IDH mutant	References
Corresponds to	Primary GBM	Secondary GBM	Louis et al., 2016
Proportion	90%	~10%	Louis et al., 2016
Age	Usually > 60	Younger adults	Louis et al., 2016
CpG methylator	Less frequent	More frequent	Brennan et al., 2013
TERT promoter mutation	~95%	51%	Yan et al., 2009
homologous deletion of CDKN2A/CDKN2B	~45%	Less	Yan et al., 2009
EGFR alterations	~41%	0%	Yan et al., 2009
PTEN mutation/deletion	~25%	0%	Yan et al., 2009
TP53 mutations	~20%	81%	Yan et al., 2009

Figure 1.2: Characteristics of IDH wild and IDH mutant subtypes [2]

IDH wild type refers to the type without IDH mutation, the survival rate is low, and it is mainly differentiated by stellate cells. There are typical diffuse growth patterns, mitotic activity, and microvascular proliferation and necrosis. Multiple studies have shown that IDH mutants have prognostic and predictive value, for example, IDH-mutated patients are more responsive to temozolomide [12] and can be targeted for drug therapy.

1.1.1.3 DNA Methylation-Based Subtypes

Epigenetics is a common hallmark of human cancer, and glioblastoma is no exception. Among them, DNA methylation is a core element of epigenetic changes, an important signaling tool for regulating genome function, and key features for mediating tumorigenesis. In 2013, Brennan [13] et al. GBMs were classified using large-scale methylation sequencing data and classified into six categories according to the expression level of DNA methylation, including Cluster M1 to Cluster M6. In 2019, Ma [14] et al. specific prognostic subtypes and identified 3 GBM methylation clusters (Cluster1, Cluster2 and Cluster3) with significantly different survival curves. Among all clusters, Cluster2 had the best prognosis.

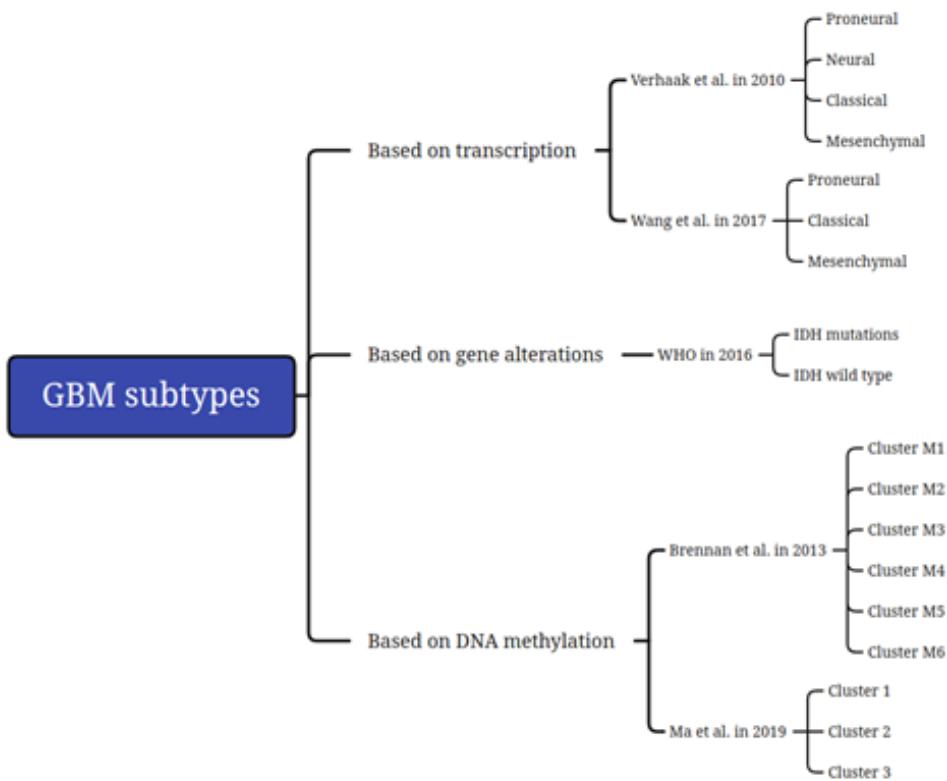


Figure 1.3: Mainstream classification methods and specific categories of glioblastoma subtypes at the molecular level

1.1.2 Diagnosis of glioblastoma

Generally, the diagnosis of glioblastoma involves three main steps. First, a neurological exam is performed to check the patient's vision, hearing balance, and other neurological characteristics. Some abnormal symptoms may provide clues that the brain is affected by a brain tumor. Several sophisticated imaging techniques can then pinpoint the location of the brain tumor or examine the chemical characteristics of the tumor, including computed tomography (CT) and magnetic resonance imaging (MRI), magnetic resonance spectroscopy (MRS), and positron emission tomography (PET). In fact, MRI is the most useful imaging technique for diagnosing glioblastoma, while CT is used less frequently. The diagnostic value of other imaging techniques is currently unclear. The final step in diagnosis is a biopsy. Depending on your condition and the location of the tumor, a biopsy with a needle can be done before or during surgery to remove the glioblastoma. Histological examination and molecular marker assays using samples can determine cell type and level of invasiveness.

1.1.3 Therapy of glioblastoma

Treatment includes surgical resection of glioblastoma, radiotherapy, chemotherapy, targeted drug therapy, etc. The standard approach to treating this type of cancer is to combine surgical resection with chemotherapy and radiation therapy. Chemotherapy has fewer side effects than radiotherapy because of drug interactions that are still unknown for many chemicals, such as systemic effects and cognitive impairment

1.1.3.1 Chemotherapy and Radiotherapy

As always, chemotherapy will be given 4 weeks after a patient's surgery or shortly after radiation therapy. In some cases where the tumor cannot be removed by surgery, chemotherapy will be the main treatment. Chemotherapy works by destroying rapidly dividing cancer cells with powerful drugs. It can be deployed to the patient in a number of ways, such as in the form of a pill for the patient to swallow, injected into a vein, injected into the intrathecal space of the spine, and placed in your brain during surgery. Chemotherapy often needs to be carefully

CHAPTER 1. INTRODUCTION

designed because tumors are often composed of several types of cells that certain drugs can destroy, while some tumor cells can grow unhindered. Some drugs are most commonly used to treat glioblastoma, including temozolomide, carmustine, bevacizumab, lomustine, and others. To determine which drug or combination of drugs is most effective, an assessment must take into account cellular composition and tumor location, patient age, personal preference, and response to prior therapy.

Radiation therapy has been used to treat glioblastoma since the 1940s. In the 1970s, Walker and colleagues discovered a dose relationship in glioblastoma. From the 1979s to the 1980s, two imaging techniques began to be incorporated into radiation therapy, CT and MRI, which allowed better definition of tumor target volumes to aid radiation therapy. The current standard radiotherapy for GBM patients is recommended to be implemented within 1 month after surgery, with a total dose of 54-60 Gy divided into 30 fractions, and a single dose of 1.8-2.0 Gy.

1.1.3.2 Personalized Therapy

Gene Therapy

Gene therapy aims to introduce genetic material into cells to compensate for abnormal genes or to make beneficial proteins. If the mutated gene causes the necessary protein to be lost, gene therapy can introduce normal genes to supplement the protein's function. Gene therapy is the delivery of genes into cells through vectors. Viruses are often used as vectors because they can deliver new genes by infecting cells. These viruses are modified so that they do not cause disease when they are used in humans [2]. Using gene therapy techniques to repair and compensate for tumor suppressor gene mutations in various subtypes of GBM patients, such as PTEN mutations in the classic subtype, has a high possibility of improving the survival time of patients.

Immunotherapy

Immunotherapy holds promise for sustained antitumor immunity that is pathway-independent and has the potential to amplify antigens to enhance immune responses. Tumor-specific antigen vaccines require confirmation that the tumor expresses the targeted antigen [2]. Therefore, immunotherapy limits the scope and population of these vaccines, so specific vaccines can be designed according to the expression of different subtypes of molecules. The most famous of these are dendritic cell-based

vaccines and peptide vaccines.

Targeted therapy

Targeted therapy uses drugs or other substances to stop the growth and spread of cancer cells by targeting specific molecules involved in the growth, development and spread of cancer cells. It is the foundation of precision medicine, a type of medicine that uses personal genetic and protein information to prevent, diagnose and treat disease. Targeted therapy acts on specific molecular targets associated with cancer, which are carefully selected and designed to interact with the target. Among the targeted therapy for glioblastoma, bevacizumab (BEV) targeted therapy is currently more mature. According to research, bevacizumab can prolong the overall survival cycle of proneurotype and IDH wild type, while effectively controlling edema without immunosuppression.

1.2 Radiogenomics

Over the past few decades, the emergence and development of genomic assessment methods and computational approaches has held promise for identifying therapeutic targets that may aid in the treatment of glioblastoma. However, targeted therapy has had little success in curing glioblastoma patients, leaving them with a poor prognosis. Glioblastomas exhibit a high degree of heterogeneity in both space and time. The presence of distinct genetic subgroups in glioblastoma enables this tumor to adapt to environmental forces. As a result, glioblastoma patients do not respond well to prescribed therapy because therapy targets the entire tumor rather than specific genetic subregions. Genomic alterations within tumors produce distinct radiographic phenotypes. In this regard, magnetic resonance imaging plays a key role in characterizing the molecular features of glioblastomas according to the regional variation and phenotypic presentation of the tumor. Radiogenomics has emerged as a new area of research to explore links between genetic alterations and imaging features. Radiogenomics offers many advantages, including non-invasive and holistic assessment of tumors and their response to therapy.

Most existing radiogenomics studies aim to establish the relationship between tumor radiographic features (eg, tumor enhancement volume, degree of necrosis) and gene expression profiles or pathways. These exploratory studies are designed

to lay the groundwork for optimal study design, data collection, and analysis to help formulate relevant hypotheses for future research. Exploratory studies are aimed at finding relevant mutations that may give rise to unique radiological phenotypes. Ideally, radiogenomics studies are designed based on the following systematic approach: Figure 1.4

Typical Workflow of Radiogenomic Studies

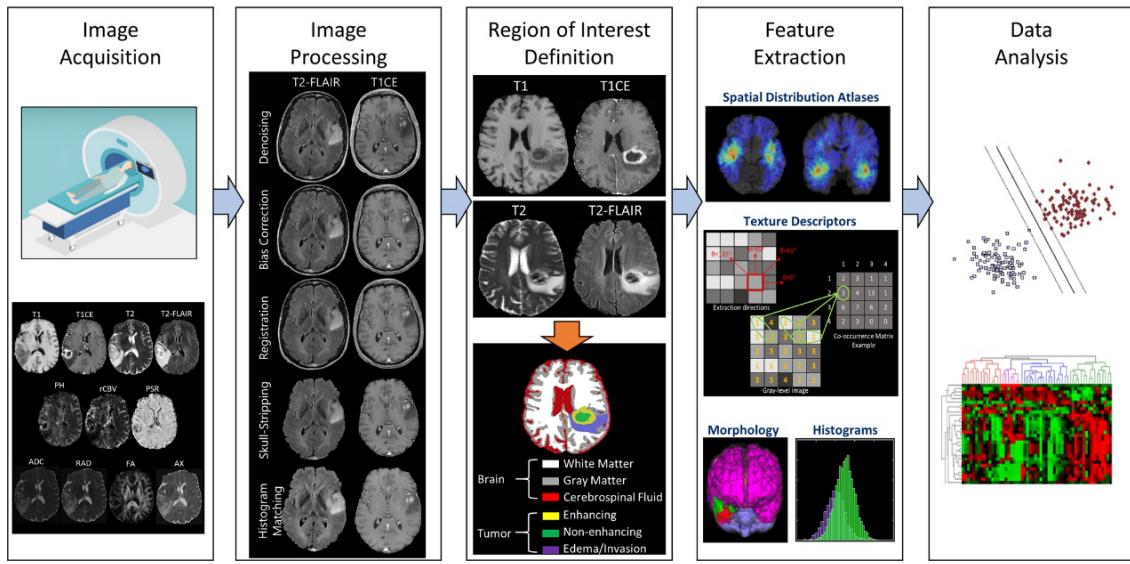


Figure 1.4: : A representation of the proper workflow of a radiomic study, which includes the following steps: 1) image acquisition; 2)image processing, including noise/artifact reduction, intensity and/or orientation standardization, coregistration of themultiparametric MRI scans; 3) ROI definition using manual annotation or (semi-)automatic segmentation; 4) feature extraction basedon human-engineered (conventional radiomics) or deep-learning approaches; and 5) data analysis, involving machine/deep-learningmethods for feature selection, classification, and crossvalidation. Radiogenomics studies should ideally follow the same workflow,with genomics of glioblastoma as their endpoint.3Kazerooni et al.: Review of Glioblastoma Radio-genomics [15]

This final year project is focus on last three step: ROI definition by image segmentation, feature extraction in conventional computer vision, data analysis like subtype prediction and mutual information analysis.

1.3 Segmentation Algorithms for Glioblastoma

1.3.1 Region-based Approach

Region-based methods aim to provide segmentation by finding coherent regions or pixel similarities.

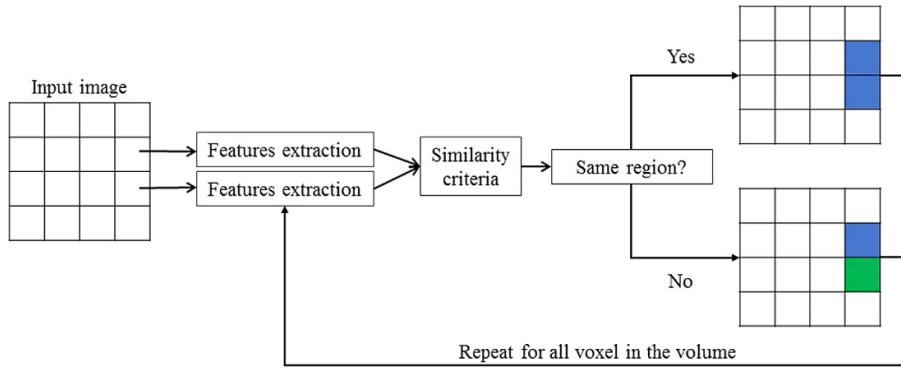


Figure 1.5: Conventional pipeline of region-based approach: features of one voxel and their neighbourhoods are analysed. According to similarity criteria pre-defined, voxels are clustered [16].

The classic algorithm for segmenting glioblastoma lesions based on regions is the watershed segmentation algorithm introduced by Letterboer et al. in 2004 [17], which is applied to tumor segmentation of MRI images. This method is mainly used for preoperative neurosurgical planning. Segmentation of the tumor enables neurosurgeons to quickly distinguish resection boundaries. There is a common problem with applying watershed segmentation - over-segmentation. To overcome this problem, the authors introduced multiscale analysis. In the authors' experiments, they deployed the algorithm on T1GD MRI images of 20 patients. The gold standard was achieved by manual segmentation by three physicians. To assess reproducibility, two automated delineations were applied to each dataset. Results tended to show that watershed segmentation was more reproducible than manual delineation. Discovery accuracy equals human delineation. The computation time for the watershed algorithm is 10 minutes, and the computation time for operator interaction is 1 to 15 minutes. The authors claim that the watershed approach is three times more efficient than manual delineation.

1.3.2 Edge-based Approach

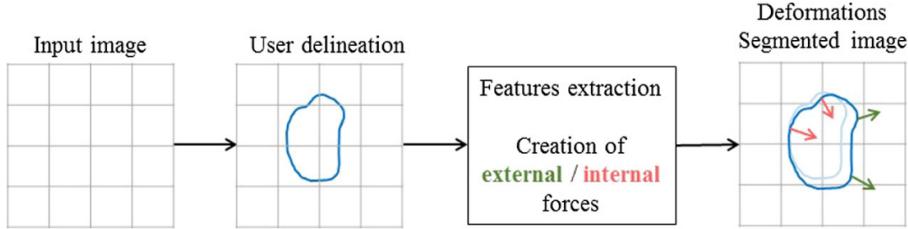


Figure 1.6: Conventional pipeline of edge-based approach: a rough delineation drew by the user is used to compute forces from neighbourhood voxel of the first contour. These forces iteratively deform the initial shape under constraints to obtain the final segmentation [16].

Basically, deformable profiles are 2D or 3D shapes (curves or surfaces) that iteratively deform under the influence of internal and external forces. The main disadvantage of this method is that it depends on user initialization and problems encountered when tumor discontinuities or weak contrasts are present [16]. This method does not work for automatic tasks

1.3.3 Classification Approaches

Classification methods are widely used in image segmentation. They are designed to cluster pixels based on different characteristics of the input vector (intensity, texture, neighbors and spatial distribution in the image) used as a clustering algorithm. Classification methods are often called supervised or unsupervised, such as Fuzzy C-means, a popular unsupervised clustering method, and U-net, a popular supervised method. This automated algorithm is more suitable for the application scenario of my graduation project

In 2008, Corso [18] et al. proposed Bayesian model classification, an unsupervised approach based on two concepts: class models and graph cuts. The aim is to combine the speed of graph cutting and the statistical distribution accuracy of the class model. A graph was created showing similarity to adjacent pixels using features such as intensity contrast, texture differences, and boundary integrity. Then, similar pixels are classified into different groups or nodes. The concept of cut is introduced as all these nodes are linked by affinity and "cut" defines the threshold between two

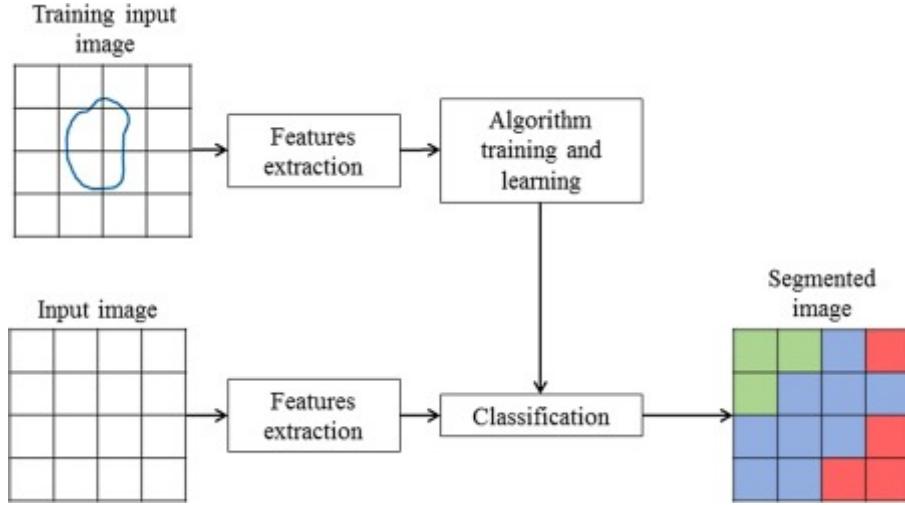


Figure 1.7: Conventional pipeline of learning machine approach: in a first part, the algorithm is trained with several pre-segmented volume and learned features of voxels labelled previously. Then, the initial volume is classified according to rules learned previously by the algorithm [16].

groups. The model was validated to have approximately 70% correlation with expert hand-painting in all T1, T1Gd, T2 and FLAIR sequence MRI images.

In 2013, Jiang [19] et al. proposed a method based on semi-automatic graph cutting. Segmentation of brain tumors using T1, T1Gd, T2, and FLAIR MRI images. This method uses arithmetic operations on imaging modalities such as subtraction between modalities such as T1-T2 or T2-FLAIR to obtain better separation between brain tumor and background. Texture features were also extracted using Gabor filters due to their high discrimination against tumor parts in MRI images. Furthermore, the Real-AdaBoost algorithm draws a probability map from the background to identify the target. Two classifiers (global and local) are trained using the feature cluster samples from the global classifier case and the seed points in the test image of the local classifier case. A cost function consisting of two parts (region term and boundary term) is decorated and defines the final split.

Fuzzy-C mean is a popular unsupervised clustering method. In 2014, Cordova [20] et al. developed a semi-automatic tumor segmentation algorithm using T1 images and fuzzy C-means, which uses only intensity variance as features to cluster voxels.

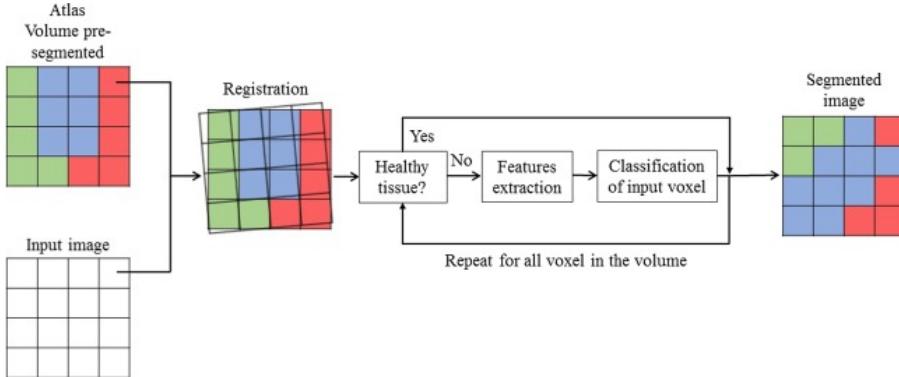


Figure 1.8: Conventional pipeline of atlas-based approach: an atlas segmented previously is used to find similarities with the studied volume. Voxel are thus labelled in comparison with the atlas to obtain the final segmentation [16].

1.3.4 Atlas-based Approaches

Altas is a standard anatomical model in which all internal structures and their associated properties are represented. However, due to the intrinsic properties of glioblastoma, simple segmentation based on brain atlases including glioblastoma cannot be used [16]. Glioblastoma location, shape, tissue heterogeneity, various volumes, and effects on patient topology in the brain make the creation of such altas tricky.

1.4 Feature Extraction

The core of radiomics is the extraction of high-dimensional feature data to quantitatively describe the properties of the volume of interest.

It is worth noting that there is a difference between feature extraction and feature selection. Feature extraction is to find as many features as possible to describe the data. The goal of feature selection is to reduce the number of extracted features to as few as possible, which in turn can be generalized into patterns that can robustly identify concepts hidden in the data while avoiding overfitting the data. Currently, feature selection is the dominant method for selecting and minimizing the number of features [21, 22]. There are various dimensionality reduction methods for selecting and minimizing the number of features, such as covariance matrices or linear discriminant analysis (LDA) such as Kumar et al. The patients described for each analysis have a large number of extractable radiological features, so the

CHAPTER 1. INTRODUCTION

number of features needs to be reduced to avoid overfitting the data. They selected only 39 of the 327 quantitative features that were repeatable and informative rather than redundant.

There are two main approaches to feature extraction: deterministic extraction and non-deterministic extraction [21]. For deterministic feature extraction, the most common approach is to use mathematical formulas to extract features related to imaging features, such as texture, intensity, or shape. Currently, based on literature reviews, texture features are one of the most important imaging features in the field of radiomics.

For feature extraction from MRI images, there are three main categories, Shape-based, Intensity-based, and Texture-based[21, 22].

Chapter 2

Method

2.1 U-Net

U-Net is a convolutional neural network that was developed for biomedical image segmentation. It is a famous representation for encoder-decoder architecture. The architecture is illustrated in Figure 2.1

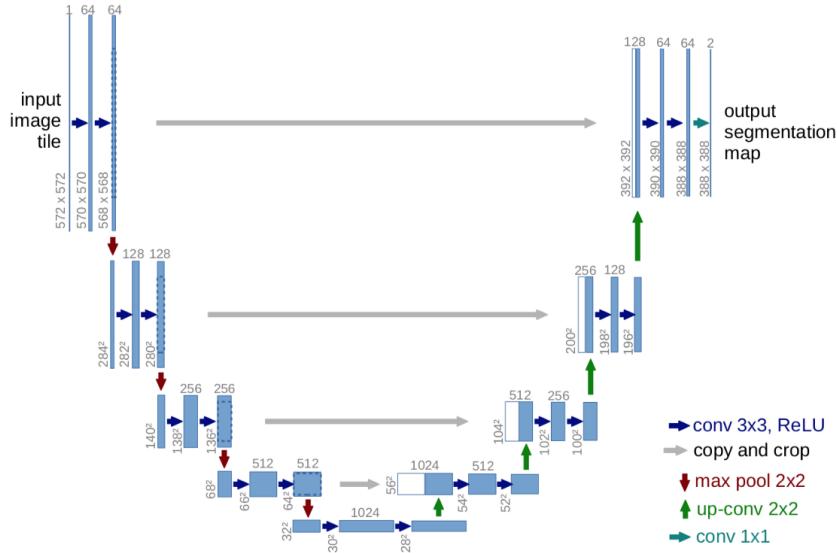


Figure 2.1: Original U-net architecture [23]

The original U-Net encoder network architecture consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling [23]. My network modified the two 3x3 convolutions with 1 reflect padding and

CHAPTER 2. METHOD

used 3x3 convolutions with 1 reflect padding and stride 2 to replace max pooling as downsampling layer. Also, convolution layer followed by a batch normalization layer and dropout layer, then a rectified linear unit (ReLU). At each downsampling step we double the number of feature channels. The code as shown in § A.1

For the decoder part network architecture, every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution. After upsampling, a concatenation with the correspondingly cropped feature map from the contracting path, then two 3x3 convolutions, each followed by a batch normalization layer, a dropout layer and then a ReLU. The upsampling layer code and the whole U-Net architecture code you can see in § A.1

In training, using binary cross entropy as loss function shown in Equation 2.1 and adaptive moment estimation as optimization algorithm shown in Equation 2.2

$$l_n = -w_n[y_n \cdot \log x_n + (1 - y_n) \cdot \log 1 - x_n] \quad (2.1)$$

$$\begin{aligned} v_t &= \beta_1 * v_{t-1} - (1 - \beta_1) * g_t \\ s_t &= \beta_2 * s_{t-1} - (1 - \beta_2) * g_t^2 \\ \Delta w_t &= -\eta \frac{\nu_t}{\sqrt{s_t + \epsilon}} * g_t \\ w_{t+1} &= w_t + \Delta w_t \\ \eta &: \text{Learning rate} \\ g_t &: \text{Gradient at time } t \text{ along } w^j \\ \nu_t &: \text{Exponential average of gradients along } w^j \\ s_t &: \text{Exponential average of squares of gradients along } w^j \\ \beta_1, \beta_2 &: \text{Hyperparameters} \end{aligned} \quad (2.2)$$

2.2 Feature Extraction

As mentioned in introduction, for feature extraction from MRI images, there are three main categories, Shape-based, Intensity-based, and Texture-based.

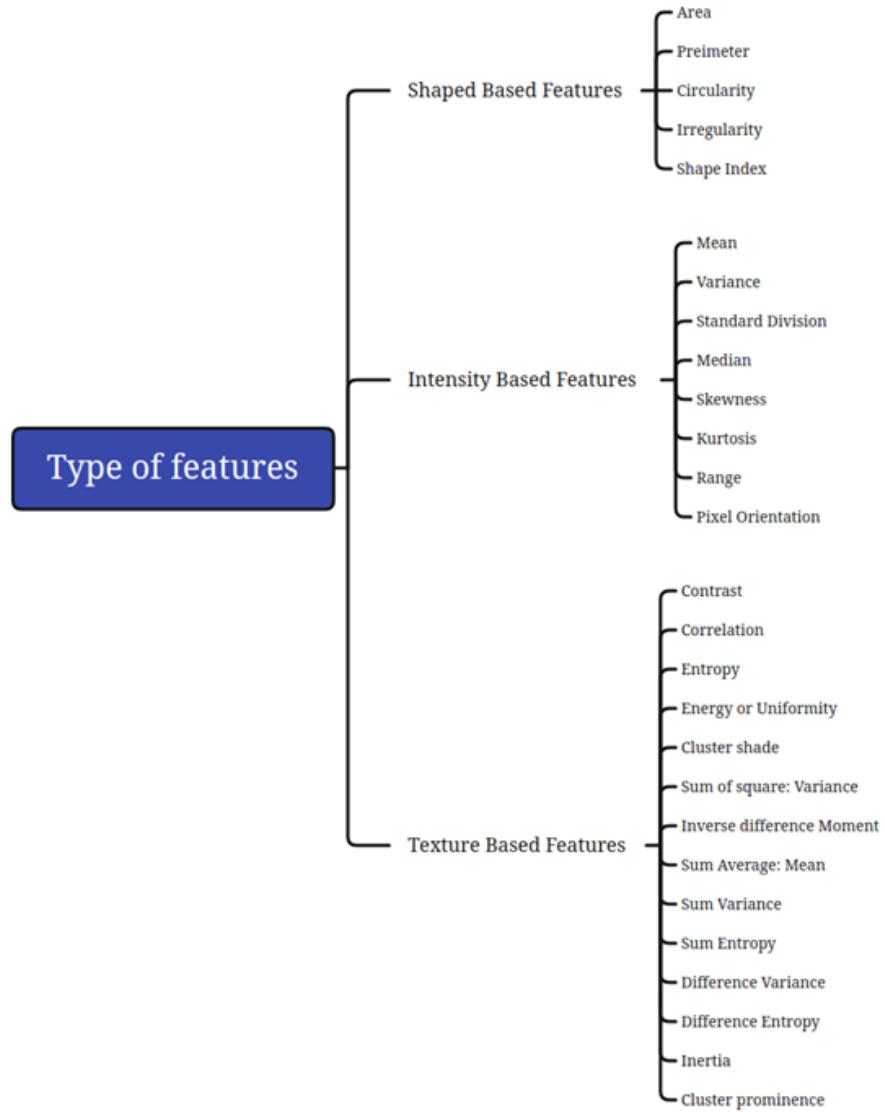


Figure 2.2: Classification of typical features that can be extracted from MRI images

2.2.1 Intensity Based Features

Intensity-based features are one of the most widely used features, such as mean, median, mode, skewness, kurtosis, energy, entropy.

Note that $f(x, y)$ is a two-dimensional function of the image, representing the intensity value of each pixel, which is the grayscale value in a grayscale image; $h(i)$ is the grayscale of the image Level, indicating how many pixels are in the picture for each level of grayscale; Ng indicates the intensity range of this picture, in common int8 type single-channel grayscale pictures, Ng is usually 256; $p(i)$

CHAPTER 2. METHOD

represents the probability distribution, representing the probability distribution of different intensity values in the picture, Equation 2.3, Equation 2.4, Equation 2.5 and Equation 2.6 is their calculation formula [21, 22].

$$f(x, y) = \begin{bmatrix} f(0, 0) & f(0, 1) & f(0, 2) & \dots & f(0, N_y - 1) \\ f(1, 0) & f(1, 1) & f(1, 2) & \dots & f(1, N_y - 1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(N_x - 1, 0) & f(N_x - 1, 1) & f(N_x - 1, 2) & \dots & f(N_x - 1, N_y - 1) \end{bmatrix} \quad (2.3)$$

$$h(i) = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} \delta(f(x, y), i), i = 0, 1, 2, \dots, N - 1 \quad (2.4)$$

$$\delta(i, j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad (2.5)$$

$$p(i) = \frac{h(i)}{N_x N_y}, i = 0, 1, 2, \dots, N - 1 \quad (2.6)$$

Based on this, we can calculate intensity based features.

1) Mean, which defines the average level of image intensity

$$\text{Mean, } \mu = \sum_{i=0}^{N_g-1} i \cdot p(i) \quad (2.7)$$

2) Variance, which defines the variation in intensity around the mean

$$\text{Variance, } \sigma^2 = \sum_{i=0}^{N_g-1} (i - \mu)^2 \cdot p(i) \quad (2.8)$$

3) Standard Deviation:(SD)

$$\text{SD or } \sigma = \sqrt{\sum_{i=0}^{N_g-1} (i - \mu)^2 \cdot p(i)} \quad (2.9)$$

4) Skewness: The skewness defines the symmetry of an image.

$$\mu^3 = \sigma^{-3} \sum_{i=0}^{N_g-1} (i - \mu)^3 \cdot p(i) \quad (2.10)$$

$$\text{Skewness} = \begin{cases} \mu^3 < 0, & \text{Histogram below the mean} \\ \mu^3 = 0, & \text{Histogram is equal to the mean} \\ \mu^3 > 0, & \text{Histogram above the mean} \end{cases} \quad (2.11)$$

5) Kurtosis: The Kurtosis defines the measure of the flatness of the histogram.

$$\text{Kurtosis}, \mu^4 = \sigma^{-4} \sum_{i=0}^{N_g-1} ((i - \mu)^4 \cdot p(i)) - 3 \quad (2.12)$$

6) Energy: Energy defines the measure of the sum of squared elements.

$$\text{Energy}, E = \sum_{i=0}^{N_g-1} [p(i)]^2 \quad (2.13)$$

7) Entropy: Entropy is defined as a measure of uncertainty in a random variable

$$\text{Entropy}, EN = - \sum_{i=0}^{N_g-1} p(i) \cdot \log_2[p(i)] \quad (2.14)$$

2.2.2 Texture Based Features

Texture is a characteristic of an image that provides a high-level description of the image, including information about tonal variations or the spatial distribution of grayscale tones. Texture extraction defines the homogeneity or similarity between image regions. Texture features divide MRI images of the brain into gray matter, white matter, cerebrospinal fluid, and tumor regions. Based on Haralick texture features, there are the following features: energy, entropy, contrast, correlation, Inverse Difference Moment (IDM), etc.

There are many ways to extract texture features, the most common one is texture feature extraction based on Gray Level Co-occurrence Matric (GLCM) design [24].

The grayscale co-occurrence matrix is defined as the probability that starting from a pixel with grayscale i , another pixel at distance (dx, dy) has a grayscale of j , as shown in Equation 2.15

$$P(i, j|d, \theta) = (x, y)|f(x, y) = i, f(x + dx, y + dy) = j; x, y = 0, 1, 2, \dots, N - 1 \quad (2.15)$$

CHAPTER 2. METHOD

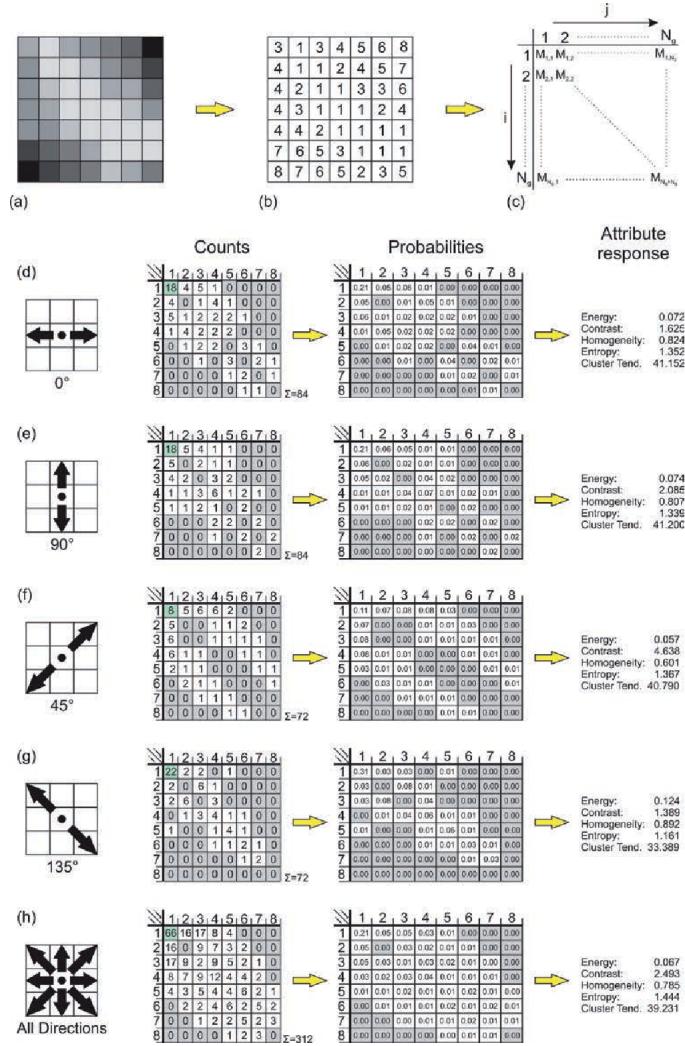


Figure 2.3: Different Parameters to calculate grayscale co-occurrence matrix (GLCM) [25]

It can be seen from the formula that the grayscale co-occurrence matrix has two parameters, d and θ , one represents the distance and the other represents the direction, which is used to represent the number of pixel pairs of a certain shape in the grayscale image that appear in the whole image. $p(i, j)$ is the result of the normalization of the grayscale co-occurrence matrix, indicating the probability of the occurrence of the grayscale value pair (i, j)

From the gray level co-occurrence matrix we can extract the following features[21, 22].

1. Energy(E), it is also known as angular second moment or uniformity, which

CHAPTER 2. METHOD

can measure homogeneity.

$$E = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i, j)^2 \quad (2.16)$$

2. Contrast(Con), it means the intensity variation between the reference pixel and neighbor pixel.

$$Con = \sum_{n=0}^{N_g-1} n^2 \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i, j)^2 \quad (2.17)$$

3. Correlation(Cor), it measures how the reference pixel is related to its neighbor pixel.

$$Cor = \frac{1}{\sigma_x \sigma_y} \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i, j) \cdot p(i, j)^2 - \mu_x \mu_y \quad (2.18)$$

4. Sum of squares: Variance(σ^2), it measures the gray tone variance.

$$\text{Variance}, \sigma^2 = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i - \mu)^2 \cdot p(i, j) \quad (2.19)$$

5. Inverse Difference Moment(IDM), it measures the local homogeneity of an image.

$$IDM = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} \frac{1}{1 + (i - j)^2} p(i, j) \quad (2.20)$$

6. Sum Average(Mean)

$$\text{Mean}, \mu = \sum_{i=0}^{2(N_g-1)} i \cdot p_{x+y}(i) \quad (2.21)$$

7. Sum Variance(SV)

$$\text{Sum Variance}, SV = \sum_{i=0}^{2(N_g-1)} (i - SE)^2 \cdot p_{x+y}(i) \quad (2.22)$$

8. Entropy(En)

$$\text{Entropy}, E = - \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i, j) \cdot \log p(i, j) \quad (2.23)$$

9. Difference Variance(DV)

$$\text{Difference Variance, DV} = \text{Variance of } p_{x-y} \quad (2.24)$$

10. Difference Entropy(DE)

$$\text{Difference Entropy, DE} = - \sum_{i=0}^{N_g-1} p_{x-y}(i) \cdot \log p_{x-y}(i) \quad (2.25)$$

11. Inertia(I)

$$\text{Inertia, } I = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i - j)^2 \cdot p(i, j) \quad (2.26)$$

12. Cluster Shade (CS), shadow size, where the image folds form reflected shadows with the light. The more flat, the smaller the cluster shadow value, on the contrary, the larger the value, the more uneven the picture is.

$$\text{Cluster Shade, CS} = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i + j - \mu_x - \mu_y)^3 \cdot p(i, j) \quad (2.27)$$

13. Cluster Prominence(CP), in the abrupt situation of the object in the picture, the greater the contrast between the texture and the pattern, the greater the cluster protrusion value.

$$CP = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i + j - \mu_x - \mu_y)^4 \cdot p(i, j) \quad (2.28)$$

2.2.3 Shape Based Features

Shape provides geometric information of objects in an image, generally including shape features such as centroid, eccentricity, area, perimeter, roundness, shape index, firmness, orientation, Euler number, etc [21, 26].

Shape based features calculation methods are various, For example, you can calculate area and perimeter by bit-quads or chain-code [26, 27].

2.2.4 Pyradiomics

In the beginning, I designed the image feature by coding by myself, you can see the code here: <https://github.com/PinkR1ver/Image-Feature-Extraction>. But

without parallel computing, the calculation speed is very low. Finally, I use the python module *Pyradiomics* meant to do the feature extraction in the ROI.

There are totally 120 features in *Pyradiomics* module, as shown in Figure 2.4, including intensity based features, shape based features and Texture based features. Texture features occupying the most features in the *Pyradiomics* module. The calculate texture features from gray level matrix(GLCM) and its advanced matrix, such as GLRLM, GLSZM, GLDM and so on.

- First Order Statistics (19 features)
- Shape-based (3D) (16 features)
- Shape-based (2D) (10 features)
- Gray Level Cooccurrence Matrix (24 features)
- Gray Level Run Length Matrix (16 features)
- Gray Level Size Zone Matrix (16 features)
- Neighbouring Gray Tone Difference Matrix (5 features)
- Gray Level Dependence Matrix (14 features)

Figure 2.4: Features in *Pyradiomics*

2.3 Classifier

Faced with the extracted features, after feature selection, using these features to predict the type of glioblastoma requires a classification algorithm. Common classification algorithms based on machine learning or deep learning include support vector machines (SVM), K-nearest neighbors algorithms (KNN), decision trees, random forests and neural network classification models, etc.

2.3.1 Support Vector Machines (SVM)

The goal of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that unambiguously classifies the data points.

There are many possible hyperplanes to choose from in order to separate the two classes of data points. One of our possible options is to find a plane with the largest margin, i.e. the largest distance between the data points of the two classes, which can give us more confidence to separate the data, since such a model has low bias, that is, it has a good effect on the training set.

CHAPTER 2. METHOD

But for support vector machines, the maximum margin plane with low deviation does not have a good model variance, that is, it does not work well for other datasets. Our best bet is the Soft Margin plane, the idea is based on a simple premise: allow the SVM to make a certain number of mistakes and keep the margins as wide as possible so that other points can still be classified correctly, which is a trade off between bias and variance.

The advantage of the support vector machine algorithm is that:

1. Effective in high-dimensional space.
2. It is still valid when the dimension is greater than the number of samples.
3. A subset of training points (called support vectors) is used in the decision function, so it is also memory efficient.
4. Generality: Different kernel functions can be specified for the decision function.
Generic kernels are provided, but custom kernels can also be specified.

The disadvantages are:

1. If the number of features is much larger than the number of samples, the regularization term is crucial when choosing the kernel function to avoid overfitting.
2. SVMs do not provide probability estimates directly, these are computed using expensive five-fold cross-validation.

2.3.2 K-nearest neighbors Algorithms (KNN)

The k-nearest neighbor algorithm (KNN) is a simple and easy-to-implement supervised machine learning algorithm. The basic principle is to classify the similarity by "distance", where k represents the number of neighbors. Algorithmic calculation of k to find the most suitable value of k for the current environment

The main disadvantage of the K-nearest neighbor algorithm is that it becomes significantly slower as the amount of data increases, making it an impractical choice in environments where fast predictions are required.

2.3.3 Decision Trees

A decision tree is a flowchart-like tree structure in which internal nodes represent features (or attributes), branches represent decision rules, and each leaf node represents a result. The topmost node in a decision tree is called the root node, as shown in Figure 2.5

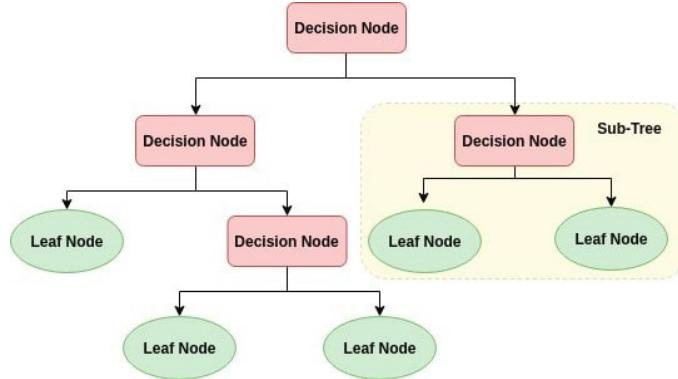


Figure 2.5: Decision Trees

The basic idea behind any decision tree algorithm is as follows:

1. Use Attribute Selection Metric (ASM) to select the best attribute to split records.
2. Make the attribute a decision node and break the dataset into smaller subsets.
3. Start building the tree by recursively repeating this process for each child node until one of the conditions matches:
 - a) All tuples belong to the same attribute value.
 - b) There are no more remaining attributes.
 - c) There are no more instances.

Commonly used measures are Entropy, Gini impure or Gini Index

Compared to the decision tree neural network algorithm, it has faster training time. The time complexity of a decision tree is a function of the number of records and attributes in the given data. A decision tree is a distribution-free or nonparametric method that does not rely on probability distribution assumptions. Decision trees work well with high-dimensional data. Because decision trees are nonparametric

methods, decision trees are also easy to interpret and visualize. Because the decision tree does not use the idea of linear recursion, the decision tree also has a certain adaptability to nonlinear relationships. The biggest problem of decision tree is that it is too sensitive to noise data, and it is very easy to overfit noisy data [28].

2.3.4 Random Forest

A random forest is composed of a large number of independent decision trees that operate as a whole. Each tree in the random forest outputs a class prediction, and the class with the most votes becomes our model's prediction, as shown in Figure 2.6.

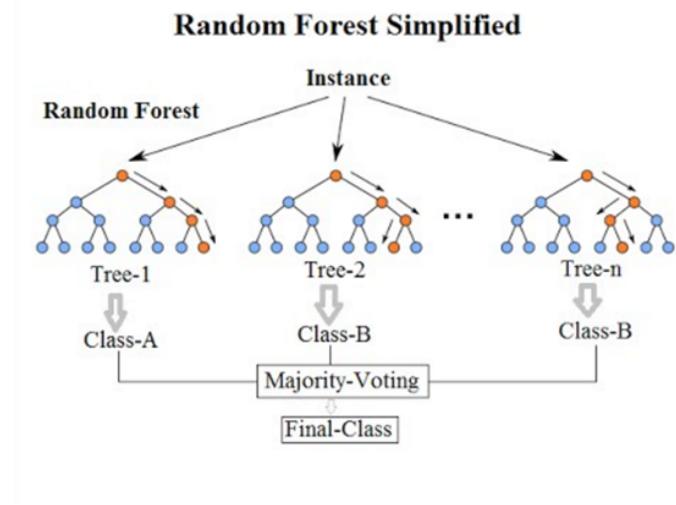


Figure 2.6: Random Forest

Random forests take a bootstrap aggregation approach to compensate for the overfitting of decision trees, which are very sensitive to the data they are trained on - small changes to the training set can result in significantly different tree structures. Random forests take advantage of this by allowing each tree to be randomly sampled from the dataset and replaced, resulting in a different tree. This process is called bootstrap aggregation. The random forest algorithm is very popular and practical in the field of finance and investment.

2.3.5 Neural Network Classification Models

Neural networks perform well in both binary classification and multi-classification, and different neural network structures can be adapted to different scenarios.

The most basic operating principle of neural network for classification is to reduce the loss function by optimizing the weight and bias, so that the model can extract the features of each category.

2.4 Mutual Information

Mutual information theory provides an intuitive tool to measure the uncertainty of random variables and the information they share, where entropy and mutual information are two key concepts [29, 30].

For two discrete variables X and Y, whose joint probability distribution is $P_{XY}(x, y)$, the mutual information between them, denoted $I(X; Y)$, is given Equation 2.29, Equation 2.30 and Equation 2.31

$$I(X; Y) = \sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)} = E_{P_{XY}} \log \frac{P_{XY}}{P_X P_Y} \quad (2.29)$$

$$P_X(x) = \sum_y P_{XY}(x, y), \quad P_Y(y) = \sum_x P_{XY}(x, y) \quad (2.30)$$

$$E_{P_{XY}} = \sum_{x,y} P_{XY}(x, y), \text{ E denotes Expect Value} \quad (2.31)$$

To understand what $I(X; Y)$ actually means, one must understand the definitions of entropy and conditional entropy, as shown in Equation 2.32

$$\text{Entropy, } H(X) = - \sum_x P_X(x) \log P_X(x) = -E_{P_X} \log P_X \quad (2.32)$$

Qualitatively, entropy is a measure of uncertainty – the higher the entropy, the more uncertain one is about a random variable.

The conditional entropy is the average uncertainty about X after observing a second random variable Y, and is given by Equation 2.33

$$H(X|Y) = \sum_y P_Y(y) \left[- \sum_x P_{X|Y}(x|y) \log(P_{X|Y}(x|y)) \right] = E_{P_Y}[-E_{P_{X|Y}} \log P_{X|Y}] \quad (2.33)$$

With the definitions of $H(X)$ and $H(X|Y)$, $I(X; Y)$ can be written Equation 2.34

$$I(X; Y) = H(X) - H(X|Y) \quad (2.34)$$

The relationship between mutual information and entropy is like the Figure 2.7

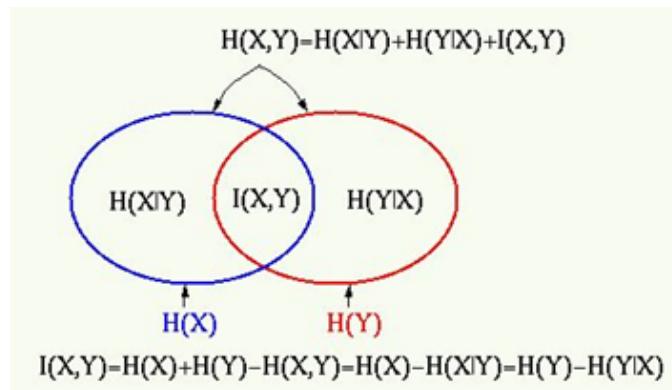


Figure 2.7: Relationship between mutual information and entropy

This gives the mutual information between Y and X meaning that if X and Y are independent the mutual information will be zero and greater than zero if they are dependent. Also, mutual information (MI) can be defined as a distance measure given by Equation 2.35

$$K(f, g) = \int f(y) \log\left(\frac{f(y)}{g(y)}\right) \quad (2.35)$$

The measure K in is the Kullback-Leibler divergence between two densities which can also be used as a measure of MI.

From the above equation, it's necessary to know the probability density function of the variables to calculate MI. Since the data obtained is only finite samples, the probability density function cannot be calculated accurately. Several methods have been developed for estimating the MI. Once a particular method is chosen for calculating MI then one of the simplest methods for feature selection is to find the MI between each feature and the output class labels and rank them based on this

CHAPTER 2. METHOD

value. This is a simple method and the results can be poor since inter-feature MI is not taken into account.

Chapter 3

Experiment

3.1 Segmentation

3.1.1 Data preprocessing

3.1.1.1 Dataset Description

The data I use in this project is from public dataset Ivy Glioblastoma Atlas Project (Ivy GAP). This data collection consists of MRI/CT scan data for brain tumor patients that form the cohort for the resource Ivy Glioblastoma Atlas Project (Ivy GAP). There are 390 studies for 39 patients that include pre-surgery, post-surgery and follow up scans [31]. Considering this dataset don't have annotation for segmentation, I use an advanced dataset based on this dataset, which comprises two paired sets of expert segmentation labels for tumor sub-compartments of the pre-operative multi-institutional scans of the Ivy Glioblastoma Atlas Project (Ivy GAP) collection of The Cancer Imaging Archive (TCIA). These labels have been approved by independent expert board-certified neuroradiologists at the Hospital of the University of Pennsylvania and at Case Western Reserve University [32].

Furthermore, for each of the paired sets of approved labels, a diverse comprehensive panel of radiomic features is provided, along with their corresponding skull-stripped and co-registered multi-parametric magnetic resonance imaging (mpMRI) volumes (i.e. native (T1) and post-contrast T1-weighted (T1-Gd), T2, T2-FLAIR), in NIfTI format, but I don't use the radiomics features they provided. I extract radiomic features on my own in next step.

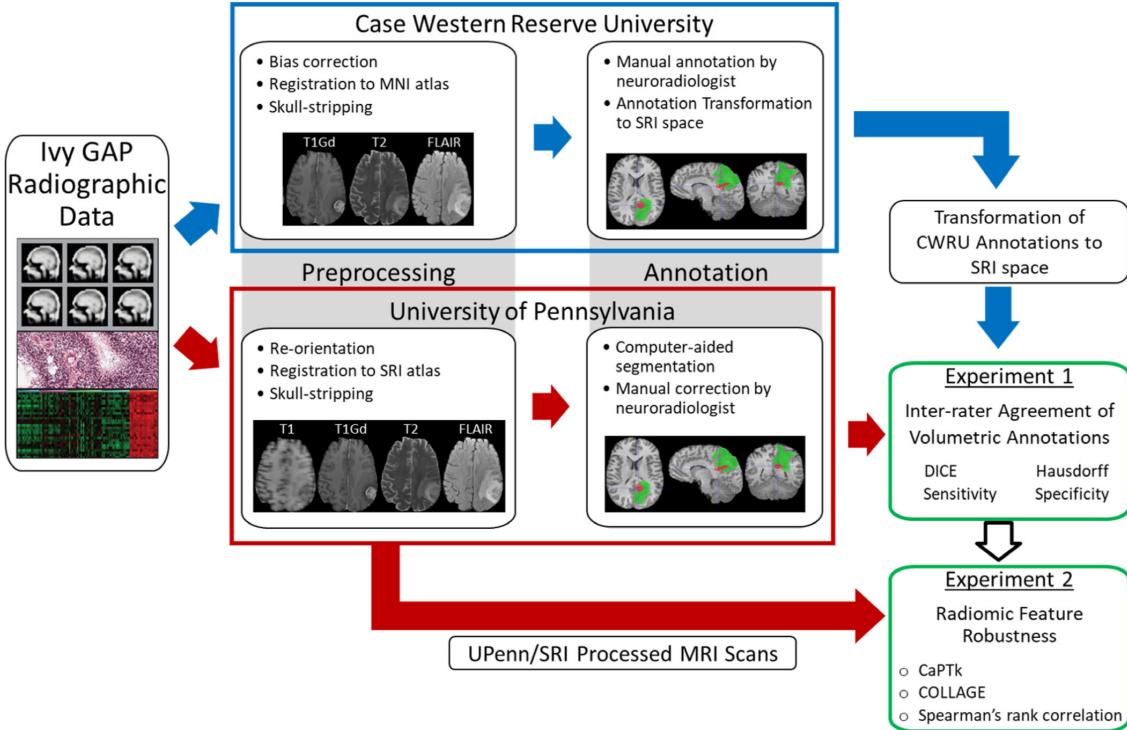


Figure 3.1: Workflow of annotation [32]

As shown in Figure 3.1, the annotation is computer-aided segmentation, manual correction by neuroradiologist. In my final year project, I choose the annotation from Case Western Reserve University (CRWU), in MNI space.

3.1.1.2 Data Format Convert and Normalization

The data format in dataset is Neuroimaing Informatics Technology Initiative (NIfTI), with .nii as file extension. It contains the metadata of the MRI, such as MRI series, description, MRI machine sampling parameter and so on. It also contains the gray scale of MRI image in 3 direction. Considering the U-Net need 2D image to input, I need to convert the dataset file format from NIfTI to uint8 image format, such as PNG or JPG, code as shown in § A.2.

In this processing, I use python module *nibabel* to read NIfTI file information and use python module *imageio* to convert every slice into uint8 image to store. In this processing, I do normalization to every brain to make them gray scale intensity as 0-255. Furthermore, before put the image into U-Net, we do normalization to the whole dataset with python deep learning framework *pytorch* function *ToTensor()*.

It will convert (0,255) to (0,1).

3.1.2 Training

In the training processing, we use U-Net as shown in § A.1. I divide the 39 patient images into 3 groups, 32 patients four out of five into train group, one out of five into validation group. 6 patients into test group. For every patient, considering their 0-20 slices are almost black and 150-192 slices are also almost black, so I choose 21-149 slices as input images. Meanwhile, I create three list to store every epoch's assessment, such as accuracy, sensitivity, specificity, IoU, f1-score and so on. Meanwhile, I store the original image, ground truth, segmentation result concat in horizontal level for every image in training processing. Every 10 epochs, I will calculate the best threshold right now for segmentation and dynamically adjust the threshold for segmentation. Every 25 epochs, I will store a model individually, meanwhile, I will update model every epoch. The whole training code you can find in <https://github.com/PinkR1ver/Radiogenomics-on-Ivy-Gap/blob/train/train.py>.

3.1.2.1 Assessment

In the training, I use some index to assess segmentation. This assessment using confusion matrix, in which each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, as shown in Figure 3.2. TN represent true negative. FN represent false negative. TP represent true positive. FP represent false positive.

With confusion matrix, we can calculate assessment index to assess the segmentation algorithm in processing.

1. Accuracy

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \quad (3.1)$$

2. Balance Accuracy

$$\begin{aligned} \text{Balance Accuracy} &= \frac{\text{TPR} \text{ (True Positive Rate)}}{2} + \frac{\text{TNR} \text{ (True Negative Rate)}}{2} \\ &= \frac{\text{TP}}{2(\text{TP} + \text{FN})} + \frac{\text{TN}}{2(\text{TN} + \text{FP})} \end{aligned} \quad (3.2)$$

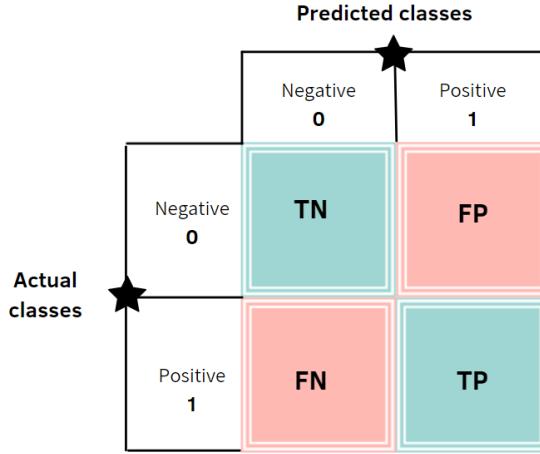


Figure 3.2: Confusion Matrix

3. Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.3)$$

4. Recall, Sensitivity, TPR

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.4)$$

5. Specificity, TNR

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3.5)$$

6. F1-score, Dice coefficient

$$\begin{aligned} \text{F1-score} &= 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \\ &= \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \end{aligned} \quad (3.6)$$

7. IoU

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \quad (3.7)$$

3.1.2.2 Threshold

As code shown in § A.1, the U-Net last layer is sigmoid function. It will output the number between (0-1), but the final result will be binary classification, meaning normal brain and tumor. We need to set a threshold to distinguish them. As usual, the threshold will be 0.5. But our data is very imbalance with so many

background and normal brain area and the tumor area is very small, we need to find the best threshold to distinguish the normal brain and tumor part. Normally, we can use ROC (Receiver operating characteristic) curve with G-mean to find the best threshold for data imbalance situation. But for medical image segmentation, we care tumor part more, so I use f1-score tuning curve with the biggest f1-score as the best threshold [33].

3.2 Feature Extraction

As methods shows that, we use python module *pyradiomics* to extract features in ROI, as code shown in § A.3. Finally, we print every slices' features in ROI into a .csv file, every row is a vector of features represent a slice, as shown in Figure 3.3

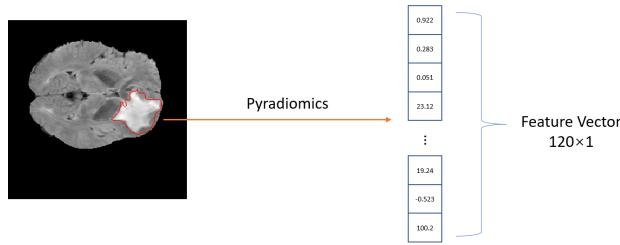


Figure 3.3: Feature Extraction

3.3 Classifier

In this project, I use python module *sci-kit learning (sklearn)* to deploy classifier to do classification to every feature vector as code shown in <https://github.com/PinkR1ver/Radiogenomics-on-Ivy-Gap/blob/master/classification.py>.

In this project, I use four classification algorithms, SVM, decision trees, random forest and Multilayer Perceptron (MLP). Among them, SVM need to do normalization and Principal Component Analysis (PCA) to data as preprocessing, cause the features here are too much, without PCA, the computer cannot handle the calculation. For MLP, we need to do normalization in the beginning to scale input variables.

3.4 Mutual Information Analysis

In this project, I focused on the gene expression of the three regions of GBM: Leading Edge (LE), Infiltrating Tumor (IT), and Cellular Tumor (CT), and focused on genes with strong gene expression such as CD44, OLIG2, PDPN, etc. The gene expression of each brain was expressed as the mean of the gene expression intensity values obtained by sampling and sequencing. A total of 32 brains meet our research conditions, and finally a 32×21 matrix is obtained, where 32 is the number of samples and 21 is the number of genes we focus on. At the same time, we also used the average value of the radiomic features of all slices in each brain to represent the radiomic features of this brain, and obtained a 32×94 matrix, 32 is the number of samples, and 94 is the radiomics features we extracted. number. Then, using these two matrices, we performed mutual information analysis on gene expression values and radiomic features, and obtained the relationship between radiomic features and gene expression intensity. See code in https://github.com/PinkR1ver/Radiogenomics-on-Ivy-Gap/blob/master/mutual_information.py.

Chapter 4

Result

4.1 Segmentation

I do segmentation for different MRI series, including T1, T2 and FLAIR. Finally, also do a segmentation for all MRI series images mixed up together.

4.1.1 T1

4.1.1.1 Partial Result Visualization

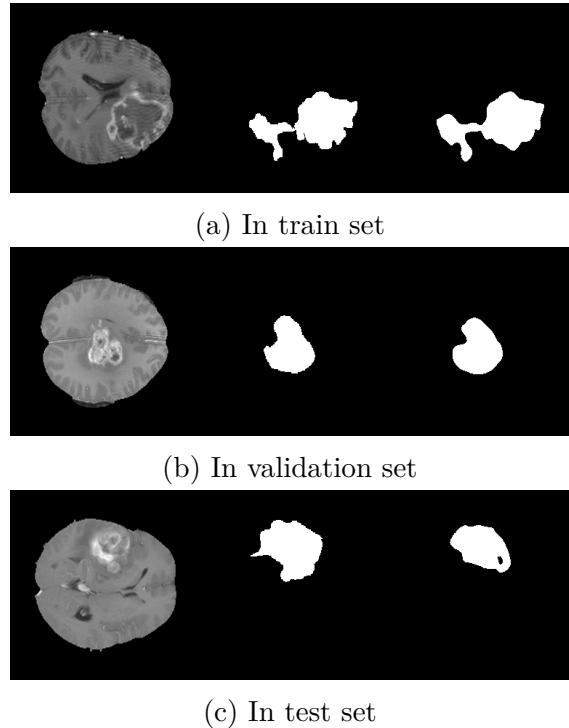


Figure 4.1: Left is origin MRI image, middle is ground truth, right is prediction

4.1.1.2 Assessment

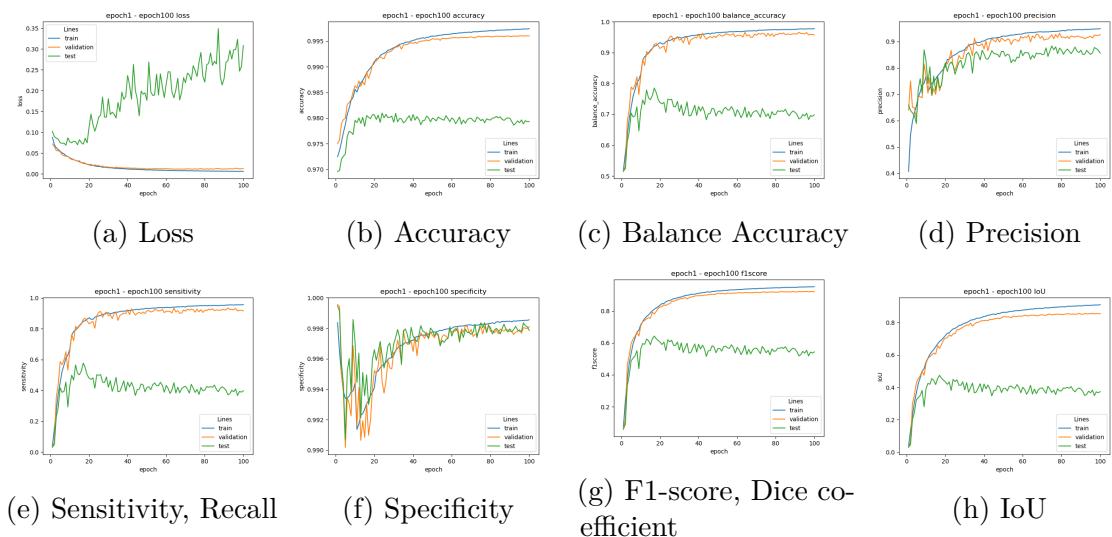


Figure 4.2: T1 series Segmentation Assessment

4.1.2 T2

4.1.2.1 Partial Result Visualization

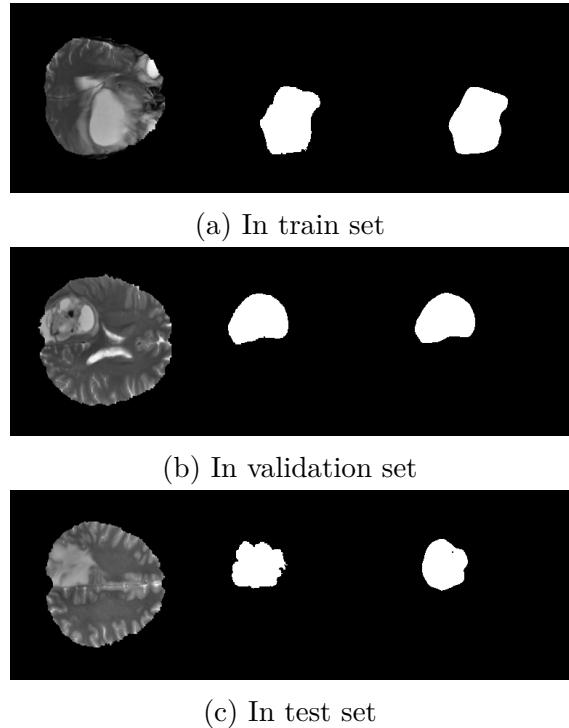


Figure 4.3: Left is origin MRI image, middle is ground truth, right is prediction

4.1.2.2 Assessment

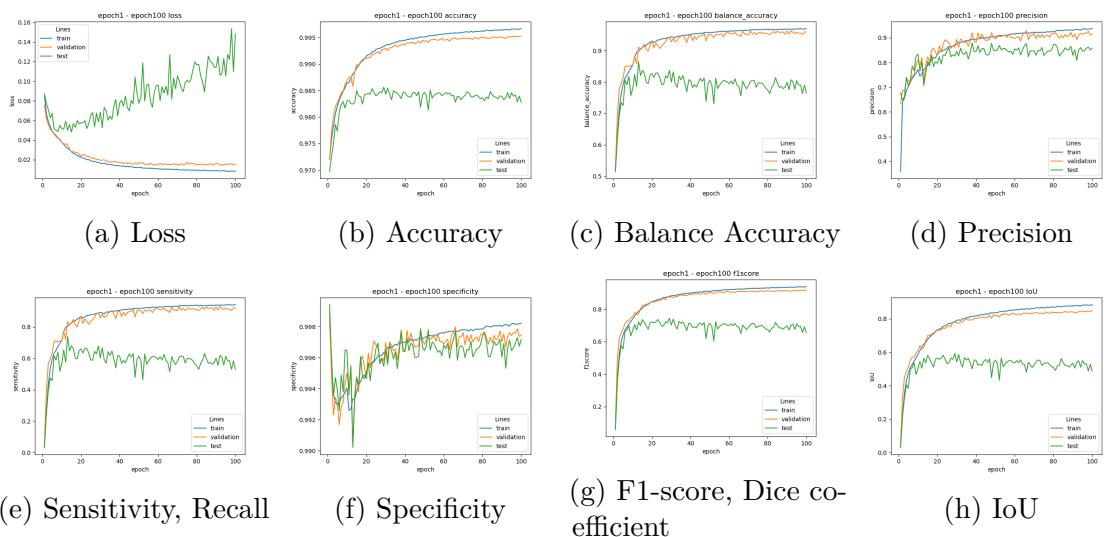


Figure 4.4: T2 series Segmentation Assessment

4.1.3 FLAIR

4.1.3.1 Partial Result Visualization

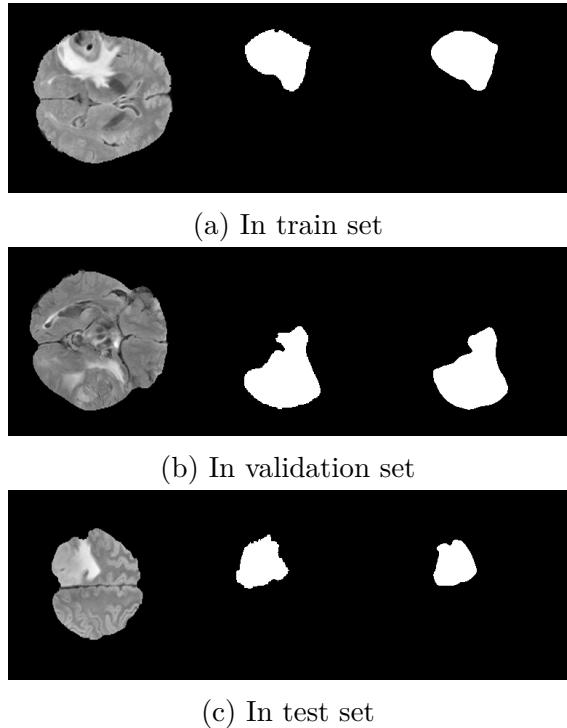


Figure 4.5: Left is origin MRI image, middle is ground truth, right is prediction

4.1.3.2 Assessment

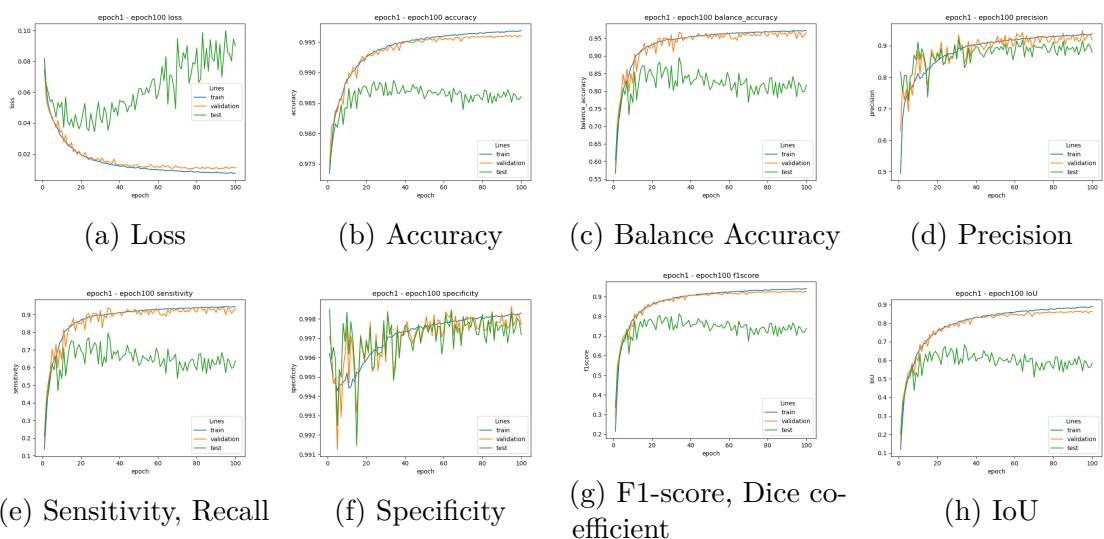


Figure 4.6: FLAIR series Segmentation Assessment

4.1.4 Stack

4.1.4.1 Partial Result Visualization

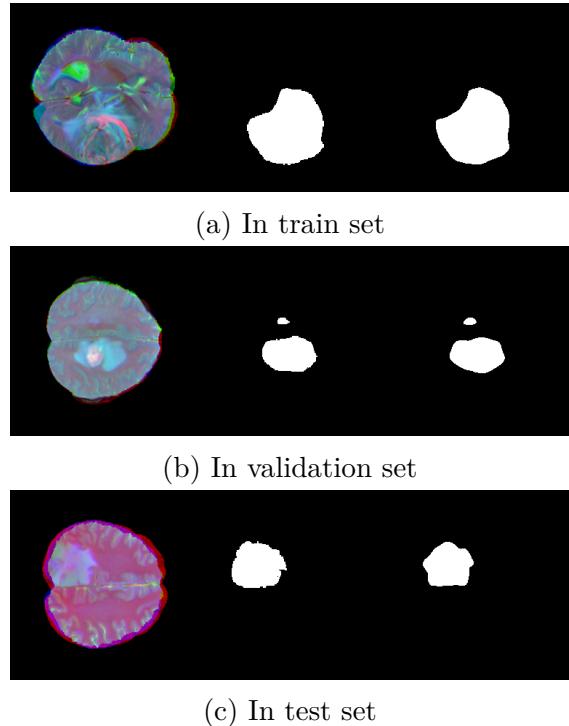


Figure 4.7: Left is origin MRI image, middle is ground truth, right is prediction

As Figure 4.2, Figure 4.4 and Figure 4.6 shown that, the validation part and train part entangle together. Their loss, accuracy, balance accuracy, precision, recall, specificity, F1-score and IoU almost have the same assessment, but test part has a lower level than them.

Depending on loss function curve, around 20 epochs is the best trade off. Also, depending the F1-score tuning curve shown in Figure 4.8, different MRI series have different best threshold for the segmentation when training epoch is 20, around 0.45.

Then we find the best trade off point based on the lowest test loss, the result as shown in Figure 4.2, Table 4.2 and Table 4.4.

CHAPTER 4. RESULT

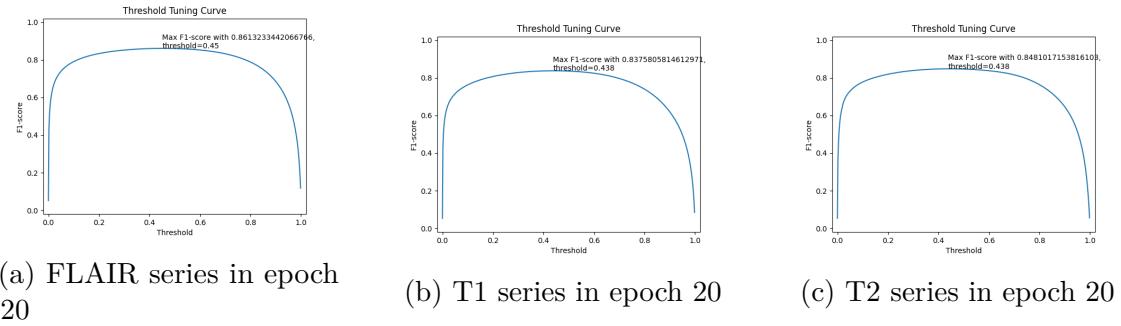


Figure 4.8: F1-score tuning curve

Table 4.1: Best Segmentation Result for T1, in epoch 17

Best Segmentation Result for T1, in epoch 17			
	train	validation	test
Accuracy	0.989330	0.987710	0.980175
Balance Accuracy	0.920806	0.926565	0.785334
Precision	0.777257	0.712830	0.724325
Sensitivity	0.848366	0.862156	0.577671
Specificity	0.993246	0.990973	0.992996
F1-score	0.811257	0.780414	0.642738
IoU	0.682449	0.639901	0.473555

Table 4.2: Best Segmentation Result for T2, in epoch 15

Best Segmentation Result for T1, in epoch 15			
	train	validation	test
Accuracy	0.989264	0.989870	0.984979
Balance Accuracy	0.910587	0.909416	0.836919
Precision	0.789571	0.823274	0.803875
Sensitivity	0.827365	0.824060	0.679117
Specificity	0.993809	0.994771	0.994722
F1-score	0.808027	0.823667	0.736248
IoU	0.677890	0.700199	0.582589

Table 4.3: Best Segmentation Result for FLAIR, in epoch 16

Best Segmentation Result for FLAIR, in epoch 16			
	train	validation	test
Accuracy	0.992049	0.992683	0.987521
Balance Accuracy	0.927457	0.925201	0.852725
Precision	0.789571	0.823274	0.803875
Sensitivity	0.859252	0.853861	0.709059
Specificity	0.995663	0.996541	0.996391
F1-score	0.851310	0.863222	0.778177
IoU	0.741113	0.759359	0.636898

Table 4.4: Best Segmentation Result for Stack, in epoch 16

Best Segmentation Result for Stack, in epoch 16			
	train	validation	test
Accuracy	0.993412	0.993930	0.989395
Balance Accuracy	0.946541	0.944896	0.797306
Precision	0.871109	0.881431	0.858787
Sensitivity	0.896896	0.893092	0.785677
Specificity	0.996186	0.996700	0.995885
F1-score	0.883814	0.887223	0.820607
IoU	0.791817	0.797306	0.695788

Based on Figure 4.2, Table 4.2 and Table 4.4, the FLAIR have the best performance in segmentation and T2 have a better performance than T1.

4.2 Classification

Because glioblastomas exhibit a high degree of heterogeneity in space, a patient will have a chance to have two different subtypes in different area of his brain. I make this special subtype as a tuple, such as (classical, Neural). I split the whole feature vectors, which are extracted from ROI, into 4:1 as train:test. Based on this, we can see the confusion matrix and assessment using different classification algorithms in test part, as shown in Figure 4.9, Table 4.5, Table 4.6, Table 4.7 and Table 4.8.

CHAPTER 4. RESULT

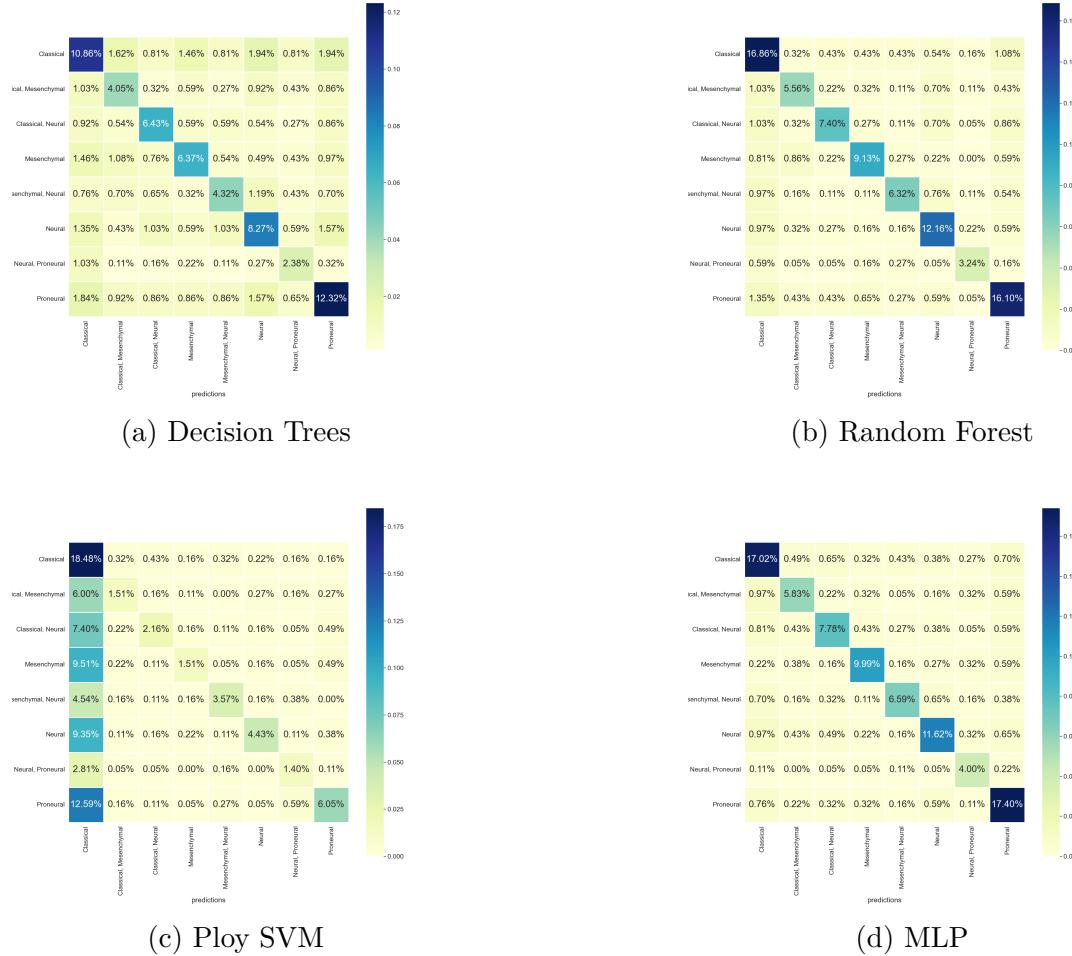


Figure 4.9: Confusion Matrix for classification using different algorithms

Table 4.5: Decision Trees Assessment

	precision	recall	f1-score	support
Classical	0.59	0.55	0.57	375
Classical, Mesenchymal	0.41	0.45	0.43	157
Classical, Neural	0.56	0.61	0.59	199
Mesenchymal	0.56	0.50	0.53	224
Mesenchymal, Neural	0.49	0.49	0.49	168
Neural	0.56	0.56	0.56	275
Neural, Proneural	0.46	0.53	0.49	85
Proneural	0.62	0.62	0.62	368
accuracy			0.55	1851
macro avg	0.53	0.54	0.54	1851
weighted avg	0.56	0.55	0.55	1851

CHAPTER 4. RESULT

Table 4.6: Random Forest Assessment

	precision	recall	f1-score	support
Classical	0.72	0.82	0.76	375
Classical, Mesenchymal	0.69	0.66	0.67	157
Classical, Neural	0.82	0.69	0.75	199
Mesenchymal	0.80	0.75	0.78	224
Mesenchymal, Neural	0.83	0.70	0.76	168
Neural	0.78	0.81	0.79	275
Neural, Proneural	0.86	0.71	0.77	85
Proneural	0.77	0.83	0.80	368
accuracy			0.77	1851
macro avg	0.78	0.75	0.76	1851
weighted avg	0.77	0.77	0.77	1851

Table 4.7: Ploy SVM Assessment

	precision	recall	f1-score	support
Classical	0.26	0.91	0.41	375
Classical, Mesenchymal	0.55	0.18	0.27	157
Classical, Neural	0.66	0.20	0.31	199
Mesenchymal	0.64	0.12	0.21	224
Mesenchymal, Neural	0.78	0.39	0.52	168
Neural	0.81	0.30	0.44	275
Neural, Proneural	0.48	0.31	0.37	85
Proneural	0.76	0.30	0.43	368
accuracy			0.39	1851
macro avg	0.62	0.34	0.37	1851
weighted avg	0.61	0.39	0.38	1851

CHAPTER 4. RESULT

Table 4.8: MLP Assessment

	precision	recall	f1-score	support
Classical	0.81	0.83	0.82	375
Classical, Mesenchymal	0.69	0.66	0.67	157
Classical, Neural	0.73	0.80	0.76	199
Mesenchymal	0.88	0.77	0.82	224
Mesenchymal, Neural	0.83	0.68	0.75	168
Neural	0.76	0.84	0.80	275
Neural, Proneural	0.85	0.71	0.77	85
Proneural	0.81	0.85	0.83	368
accuracy			0.79	1851
macro avg	0.79	0.77	0.78	1851
weighted avg	0.80	0.79	0.79	1851

Meanwhile, I split patients with two subtypes in brain out, just considering the simplified situation. The result shown in Figure 4.10, Table 4.9, Table 4.10, Table 4.11 and Table 4.12

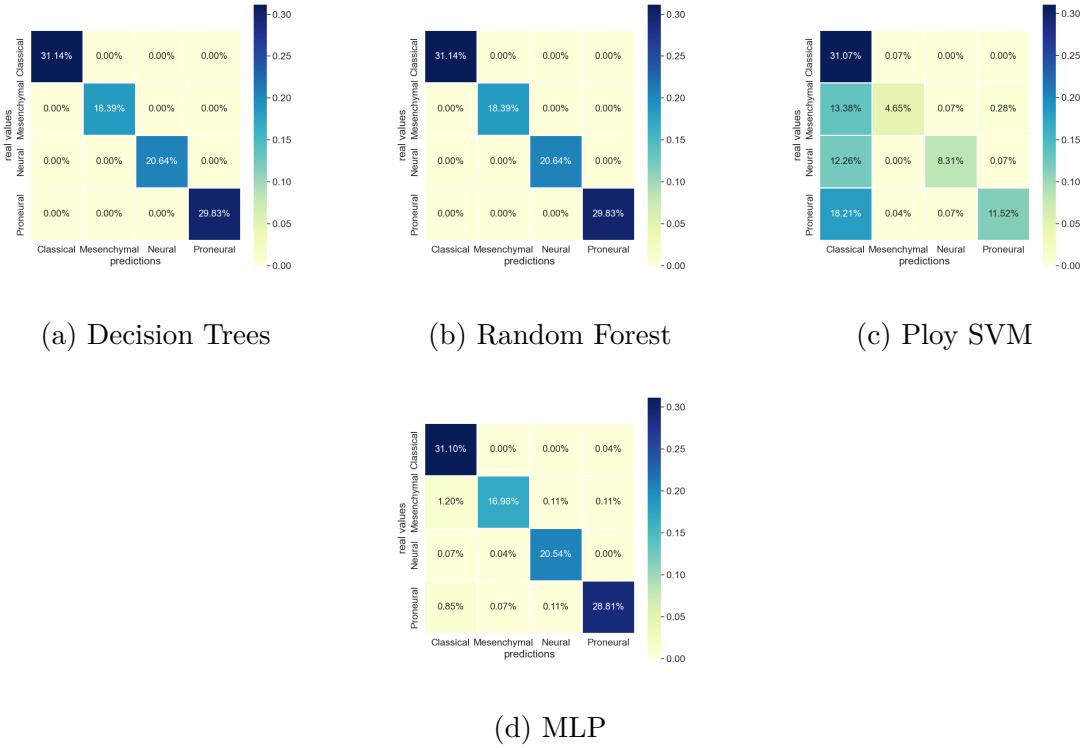


Figure 4.10: Confusion Matrix for classification using different algorithms

CHAPTER 4. RESULT

Table 4.9: Decision Trees Assessment

	precision	recall	f1-score	support
Classical	0.68	0.65	0.66	361
Mesenchymal	0.62	0.68	0.65	216
Neural	0.66	0.67	0.66	251
Proneural	0.72	0.71	0.71	390
accuracy			0.68	1218
macro avg	0.67	0.67	0.67	1218
weighted avg	0.68	0.68	0.68	1218

Table 4.10: Random Forest Assessment

	precision	recall	f1-score	support
Classical	0.83	0.84	0.83	361
Mesenchymal	0.89	0.87	0.88	216
Neural	0.83	0.83	0.83	251
Proneural	0.85	0.85	0.85	390
accuracy			0.84	1218
macro avg	0.85	0.85	0.85	1218
weighted avg	0.85	0.84	0.84	1218

Table 4.11: Ploy SVM Assessment

	precision	recall	f1-score	support
Classical	0.36	0.95	0.53	361
Mesenchymal	0.54	0.12	0.20	216
Neural	0.82	0.29	0.43	251
Proneural	0.84	0.30	0.44	390
accuracy			0.46	1218
macro avg	0.64	0.42	0.40	1218
weighted avg	0.64	0.46	0.42	1218

CHAPTER 4. RESULT

Table 4.12: MLP Assessment

	precision	recall	f1-score	support
Classical	0.87	0.88	0.88	361
Mesenchymal	0.88	0.88	0.88	216
Neural	0.85	0.85	0.85	251
Proneural	0.91	0.89	0.90	390
accuracy			0.88	1218
macro avg	0.88	0.88	0.88	1218
weighted avg	0.88	0.88	0.88	1218

In 2017, Jiguang Wang et al [9]. further divided glioblastoma into classical, proneural, and mesenchymal. They found that the Neural subtype of Verhaak was the tumor microenvironment. in non-tumor cells and Wang's subtype classification method is more common now, so I think neural subtype as proneural subtype to do three subtype classification. Result shown in Table 4.13, Table 4.14, Table 4.15 and Table 4.16

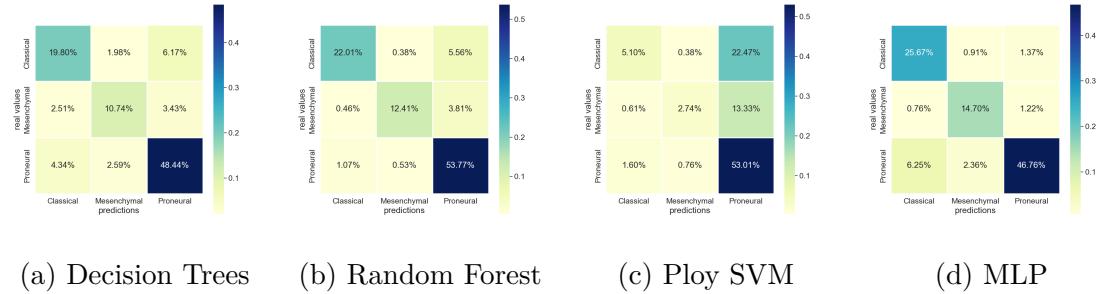


Figure 4.11: Confusion Matrix for classification using different algorithms

Table 4.13: Decision Trees Assessment

	precision	recall	f1-score	support
Classical	0.77	0.71	0.74	367
Mesenchymal	0.66	0.66	0.66	219
Proneural	0.84	0.87	0.86	727
accuracy			0.79	1313
macro avg	0.76	0.75	0.75	1313
weighted avg	0.79	0.79	0.79	1313

CHAPTER 4. RESULT

Table 4.14: Random Forest Assessment

	precision	recall	f1-score	support
Classical	0.94	0.80	0.86	367
Mesenchymal	0.93	0.74	0.82	219
Proneural	0.86	0.97	0.91	727
accuracy			0.88	1313
macro avg	0.91	0.84	0.86	1313
weighted avg	0.89	0.88	0.88	1313

Table 4.15: Ploy SVM Assessment

	precision	recall	f1-score	support
Classical	0.70	0.18	0.29	367
Mesenchymal	0.71	0.16	0.27	219
Proneural	0.60	0.96	0.74	727
accuracy			0.61	1313
macro avg	0.67	0.43	0.43	1313
weighted avg	0.64	0.61	0.53	1313

Table 4.16: MLP Assessment

	precision	recall	f1-score	support
Classical	0.90	0.90	0.90	367
Mesenchymal	0.91	0.88	0.89	219
Proneural	0.94	0.95	0.94	727
accuracy			0.92	1313
macro avg	0.92	0.91	0.91	1313
weighted avg	0.92	0.92	0.92	1313

4.3 Mutual Information

In the FYP, a mutual information analysis was carried out on the radiomic features and gene expression intensity of the tumor cell area, the infiltrating area and the tumor margin area, and the results such as Figure 4.12, Figure 4.13 and Figure 4.14. The mutual information graph is huge, you can go to the Github repository where the project is located view the original image in the library. This

CHAPTER 4. RESULT

image is a 3000×3000 png format image generated by the python library *seaborn*, and the information will not be lost after scaling.

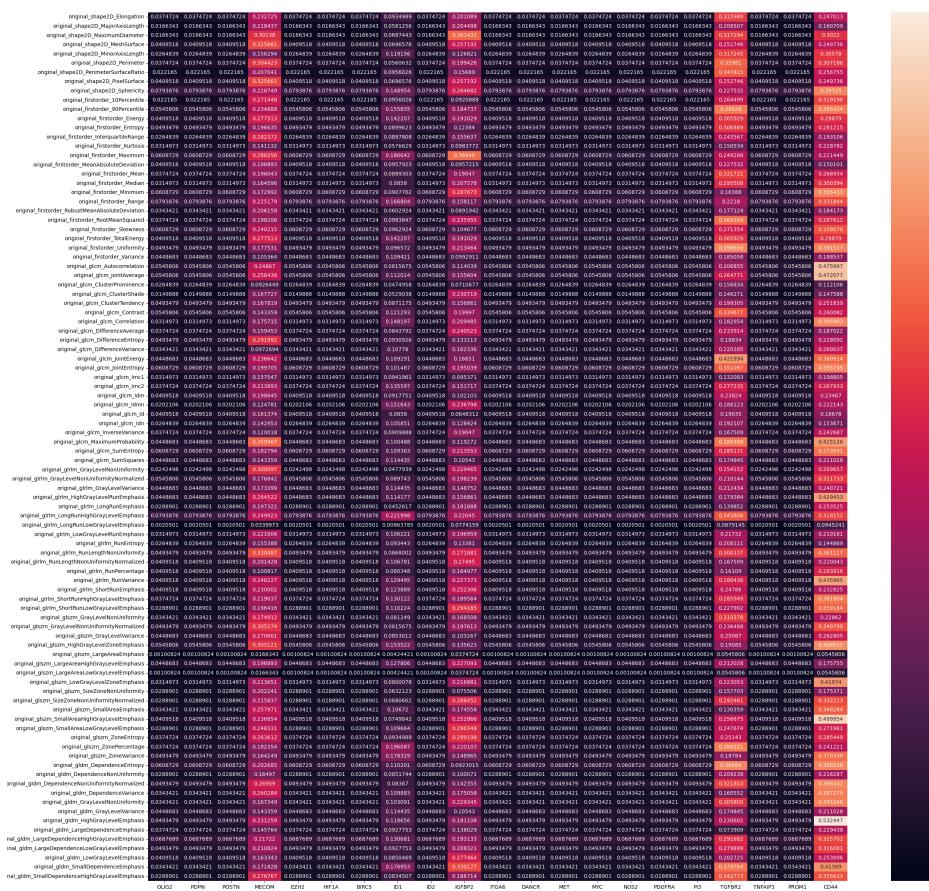


Figure 4.12: Heat map of mutual information analysis of cellular tumor (CT) radiomic features and gene expression intensity. The abscissa is the common genes in the 21 GBM samples, and the ordinate is the exomic features of the extracted MRI images. In the heat map, the closer the color is to orange, the closer the relationship is, and the closer the color is to purple and black, the weaker the relationship.

CHAPTER 4. RESULT

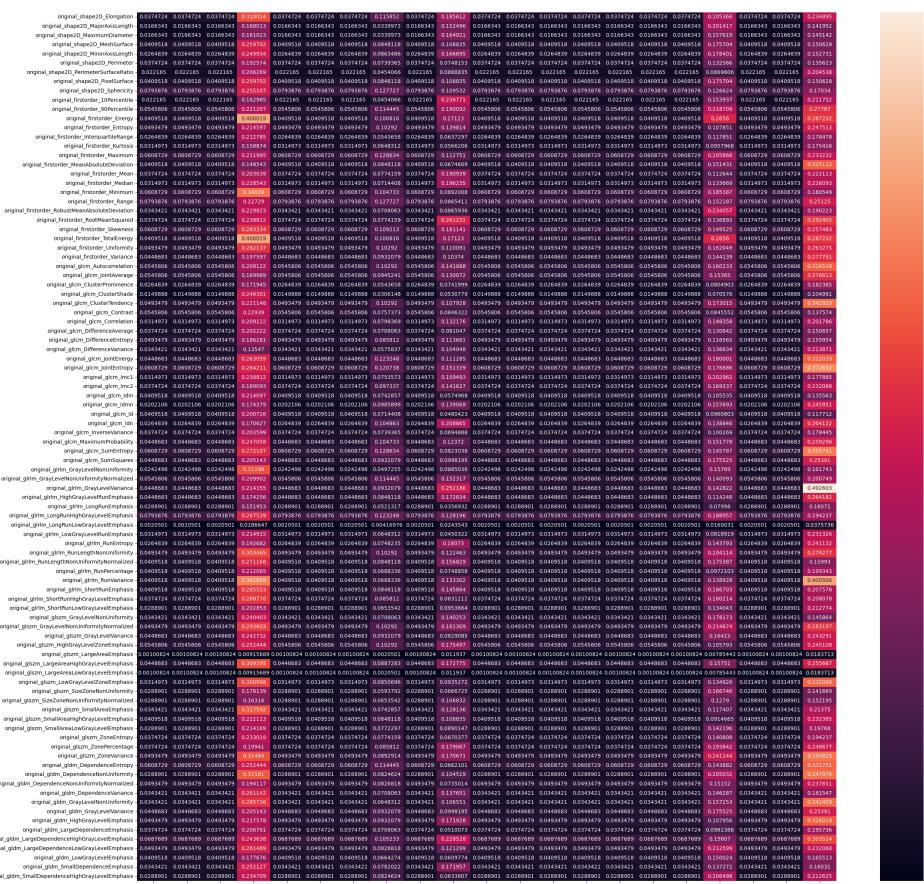


Figure 4.13: Heat map of mutual information analysis of infiltrating tumor (IT) radiomic features and gene expression intensity. The abscissa is the common genes in the 21 GBM samples, and the ordinate is the exomic features of the extracted MRI images. In the heat map, the closer the color is to orange, the closer the relationship is, and the closer the color is to purple and black, the weaker the relationship.

CHAPTER 4. RESULT

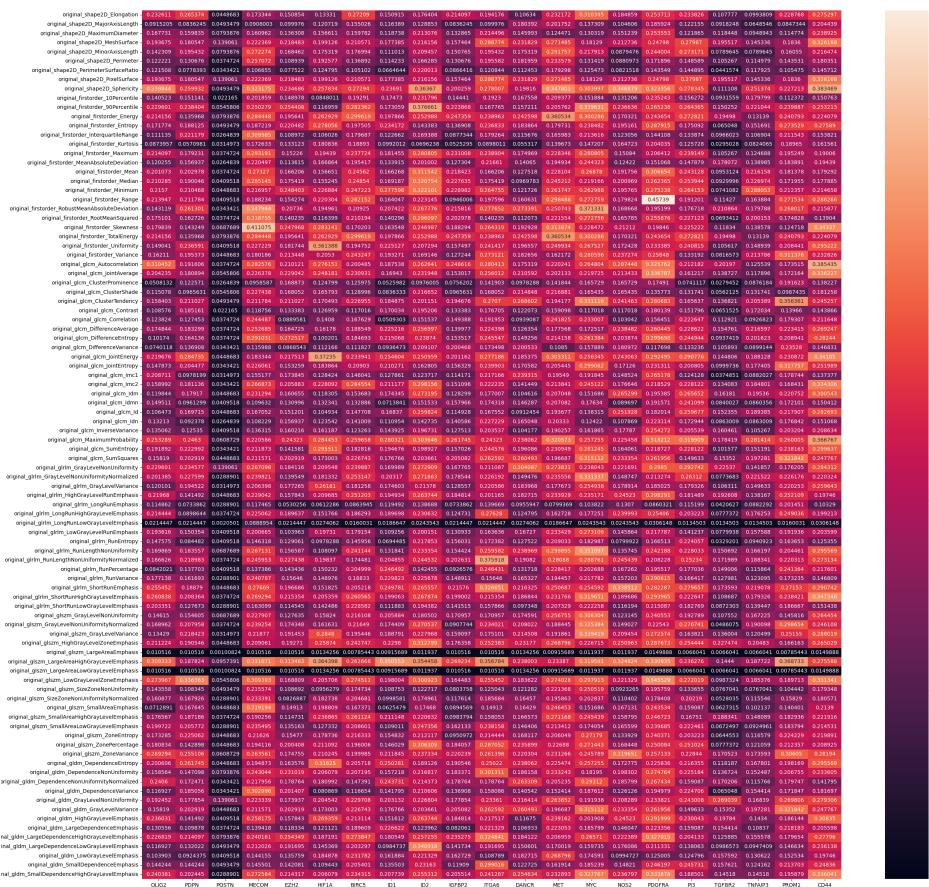


Figure 4.14: Heat map of mutual information analysis of leading edge (LE) radiomic features and gene expression intensity. The abscissa is the common genes in the 21 GBM samples, and the ordinate is the exomic features of the extracted MRI images. In the heat map, the closer the color is to orange, the closer the relationship is, and the closer the color is to purple and black, the weaker the relationship.

4.3.1 Graphic User Interference (GUI)

At the end of this project, a graphical interactive interface was created for the process of segmenting and judging GBM subtypes. The project uses *PyQt* basic components. The final result is shown in Figure 4.15, you can visit the Github repository to view the code and effects.

CHAPTER 4. RESULT

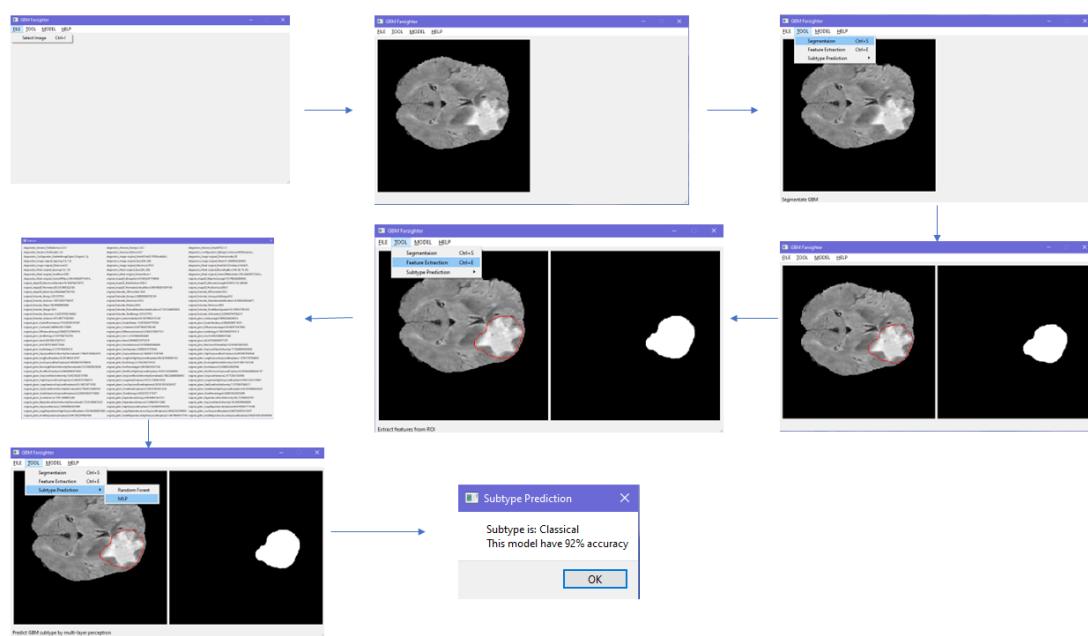


Figure 4.15: GUI Visualization

Chapter 5

Discussion

5.1 Segmentation

5.1.1 Test Part Performance worst than Train Part and Validation Part

As shown in chapter result, the validation part performance is almost the same as train part performance. It means that doctors can do some annotation to some slice, with training, computer can auto segment leftover slice image.

The test part is worst than train and validation, because it use totally different patient slice image. In my opinion, the reason why test part have worse performance are:

1. GBM heterogeneity

Glioblastomas exhibit a high degree of heterogeneity in both space and time. So different patients' MRI images will be a huge difference that U-Net model cannot learn because the patient sample number is so small. only 39 patients in total and only 32 patients for training.

2. Some MRI series cannot show the tumor architecture.

With experiment, we find that T1 series MRI image is not clear to show GBM tumor area, especially for some special subtypes in guess.

5.1.2 Other Possible Direction

TransUnet [24] architectures is the brand new architectures published in 2021.9, the first medical image segmentation framework, which establishes self-attention mechanisms from the perspective of sequence-to-sequence prediction. It generally yield weak performances especially for target structures that show large inter-patient variation in terms of texture, shape and size. It may help us to solve the GBM heterogeneity problems.

Meanwhile, it is undeniable that with the advent of deep learning methods, the demand for rich and diverse data is growing, but such a large amount and diversity cannot be found in a single institution, and can currently only be pooled across multiple institutions. found in a retrospective dataset. However, standard clinical practice data is often heterogeneous in its acquisition protocol (in terms of the type of modalities used and its solutions) and the treatment course followed, further reducing the amount of data that meets the criteria, and at the same time, imaging Data privacy issues also make it difficult to disclose data. Considering the data situation, using GAN to generate glioblastoma data to do data augmentation for training will definitely be a research trend. Based on some test patient have a great performance, I think with more data, the GBM segmentation network will be better without doubts.

5.2 Feature Extraction

In this project, we use features extracted by python module *pyradiomics*, which are designed by specific math equation. With development of deep learning, using deep learning method to get feature map from image for specific task will be more useful way to do classification task.

5.3 Mutual Information analysis

Through mutual information analysis, we can see that image features are closely related to genes CD44, TGFBR2, IGFBP2, and MECOM, both in the core area of the tumor and in the infiltrated area. Other gene expression is almost irrelevant. Perhaps this conclusion can reveal more key genes for transcription-based molecular

CHAPTER 5. DISCUSSION

subtype classification. At present, strong evidence cannot be drawn only through the mutual information analysis of radiomic characteristics and gene strength, and it needs to be combined at the biological level to explore.

5.4 Significance

At present, the significance of gene expression profiling and molecular signature analysis for the diagnosis, detection and treatment of glioblastoma is very clear. Biopsy and RNA sequencing are the traditional methods to obtain the genetic profile and molecular characteristics of glioblastoma, but biopsy usually Prone to sampling errors and unable to provide comprehensive molecular profiling within tumors, radiogenomics is a relatively new paradigm for noninvasively obtaining tumor molecular characterizations through radiophenotypic characterization of tumors.

In specific,

1. Noninvasive

Tumor molecular characterization can be obtained non-invasively by radiophenotypic characterization of tumors

2. Fast with low cost

Molecular characterization of tumors by radiophenotypic characterization can offset the expensive cost and time cost of RNA sequencing

3. Macroscopic

Biopsy is often prone to sampling errors, and cannot provide comprehensive molecular features within the tumor. It can only focus on the molecular features of glioblastoma in a certain area, and cannot predict and guide the overall trend.

4. The guiding significance of preoperative guidance and postoperative treatment plan selection

The segmentation of the target region can guide the location of the packing and the size of the tumor to be cut, and the prediction and molecular characterization of the obtained gene expression profiles can provide guidance for

CHAPTER 5. DISCUSSION

the subsequent radiotherapy and chemotherapy, and greatly help to monitor the tumor response to treatment and adapt to the longitudinal direction of the treatment strategy.

5.5 Outlook

For glioblastoma, we should not only do precision medicine. A single patient often has many different molecular-level gene mutations, which often cannot be accurately classified. Future radiogenomics research directions for glioblastoma need to focus more on the concept of personalized and adaptive medicine, which will allow the potential determination of baseline, both radiophenotypes developed in the primary tumor and subsequent scans Retention, regardless of treatment, identifies the historical molecular profile of recurrent tumors, or the presence or absence of radiophenotypic variation over time, allowing noninvasive longitudinal monitoring of mutational status and thus assessment of treatment response.

At the same time, a distributed approach, that is, cultivating machine learning-based models across multiple institutions without sharing patient data, such a distributed machine learning model building approach can be further investigated in radiogenomics research, thereby solving the problem of identifying rich and diverse data collection issues, while overcoming various data ownership issues and facilitating the exploitation of the full potential of homogenized clinical trial data.

Radiogenomics research is beginning to expand naturally to the synergistic analysis of radiological, histopathological, genetic and clinical information. This radiopathological genomic analysis should accelerate the progress of scientific discovery and provide quantitative comprehensive evaluation of patient data at multiple scales, aiming to contribute to improving personalized and precision medicine. Radiogenomics should not be limited to the field of brain tumors, but should expand to the research direction of various tumors.

More than tumor, radiogenomics or radiomics is a think of mapping two things. In this project, we map MRI images with subtypes, a gather of specific gene information to avoid invasive sampling from brain. As same, we can also map different thing in different research to make prediction from one thing to other. This thinking is

CHAPTER 5. DISCUSSION

instructive for research.

Bibliography

- [1] R. Batash, N. Asna, P. Schaffer, N. Francis, and M. Schaffer, “Glioblastoma multiforme, diagnosis and treatment recent literature review”, *Current Medicinal Chemistry*, vol. 24, no. 27, Sep. 2017. [Online]. Available: <https://doi.org/10.2174/0929867324666170516123206>.
- [2] P. Zhang, Q. Xia, L. Liu, S. Li, and L. Dong, “Current opinion on molecular characterization for GBM classification in guiding clinical diagnosis, prognosis, and therapy”, *Frontiers in Molecular Biosciences*, vol. 7, Sep. 2020. [Online]. Available: <https://doi.org/10.3389/fmolb.2020.562798>.
- [3] D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison, “The 2016 world health organization classification of tumors of the central nervous system: A summary”, *Acta Neuropathologica*, vol. 131, no. 6, pp. 803–820, May 2016. [Online]. Available: <https://doi.org/10.1007/s00401-016-1545-1>.
- [4] H. Ohgaki and P. Kleihues, “Genetic pathways to primary and secondary glioblastoma”, *The American Journal of Pathology*, vol. 170, no. 5, pp. 1445–1453, May 2007. [Online]. Available: <https://doi.org/10.2353/ajpath.2007.070011>.
- [5] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O’Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, and D. N. Hayes, “Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1”, *Cancer Cell*, vol. 17, no. 1, pp. 98–110, Jan. 2010. [Online]. Available: <https://doi.org/10.1016/j.ccr.2009.12.020>.

BIBLIOGRAPHY

- [6] A. Sehgal, “Molecular changes during the genesis of human gliomas”, *Seminars in Surgical Oncology*, vol. 14, no. 1, pp. 3–12, Jan. 1998. [Online]. Available: [https://doi.org/10.1002/\(sici\)1098-2388\(199801/02\)14:1%3C3::aid-ssu2%3E3.0.co;2-f](https://doi.org/10.1002/(sici)1098-2388(199801/02)14:1%3C3::aid-ssu2%3E3.0.co;2-f).
- [7] H. S. Phillips, S. Kharbanda, R. Chen, W. F. Forrest, R. H. Soriano, T. D. Wu, A. Misra, J. M. Nigro, H. Colman, L. Soroceanu, P. M. Williams, Z. Modrusan, B. G. Feuerstein, and K. Aldape, “Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis”, *Cancer Cell*, vol. 9, no. 3, pp. 157–173, Mar. 2006. [Online]. Available: <https://doi.org/10.1016/j.ccr.2006.02.019>.
- [8] H. Colman, L. Zhang, E. P. Sulman, J. M. McDonald, N. L. Shooshtari, A. Rivera, S. Popoff, C. L. Nutt, D. N. Louis, J. G. Cairncross, M. R. Gilbert, H. S. Phillips, M. P. Mehta, A. Chakravarti, C. E. Pelloski, K. Bhat, B. G. Feuerstein, R. B. Jenkins, and K. Aldape, “A multigene predictor of outcome in glioblastoma”, *Neuro-Oncology*, vol. 12, no. 1, pp. 49–57, Oct. 2009. [Online]. Available: <https://doi.org/10.1093/neuonc/nop007>.
- [9] Q. Wang, B. Hu, X. Hu, H. Kim, M. Squatrito, L. Scarpace, A. C. deCarvalho, S. Lyu, P. Li, Y. Li, F. Barthel, H. J. Cho, Y.-H. Lin, N. Satani, E. Martinez-Ledesma, S. Zheng, E. Chang, C.-E. G. Sauvé, A. Olar, Z. D. Lan, G. Finocchiaro, J. J. Phillips, M. S. Berger, K. R. Gabrusiewicz, G. Wang, E. Eskilsson, J. Hu, T. Mikkelsen, R. A. DePinho, F. Muller, A. B. Heimberger, E. P. Sulman, D.-H. Nam, and R. G. Verhaak, “Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment”, *Cancer Cell*, vol. 32, no. 1, 42–56.e6, Jul. 2017. [Online]. Available: <https://doi.org/10.1016/j.ccr.2017.06.003>.
- [10] D. W. Parsons, S. Jones, X. Zhang, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I.-M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, L. A. Diaz, J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. N. Marie, S. M. O. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler, “An integrated genomic analysis of human glioblastoma multiforme”, *Science*,

BIBLIOGRAPHY

- vol. 321, no. 5897, pp. 1807–1812, Sep. 2008. [Online]. Available: <https://doi.org/10.1126/science.1164382>.
- [11] H. Yan, D. W. Parsons, G. Jin, R. McLendon, B. A. Rasheed, W. Yuan, I. Kos, I. Batinic-Haberle, S. Jones, G. J. Riggins, H. Friedman, A. Friedman, D. Reardon, J. Herndon, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, and D. D. Bigner, “IDH1/IDH2 mutations in gliomas”, *New England Journal of Medicine*, vol. 360, no. 8, pp. 765–773, Feb. 2009. [Online]. Available: <https://doi.org/10.1056/nejmoa0808710>.
- [12] Q. SongTao, Y. Lei, G. Si, D. YanQing, H. HuiXia, Z. XueLin, W. LanXiao, and Y. Fei, “IDH mutations predict longer survival and response to temozolamide in secondary glioblastoma”, *Cancer Science*, vol. 103, no. 2, pp. 269–273, Nov. 2011. [Online]. Available: <https://doi.org/10.1111/j.1349-7006.2011.02134.x>.
- [13] C. W. Brennan *et al.*, “The somatic genomic landscape of glioblastoma”, *Cell*, vol. 155, no. 2, pp. 462–477, Oct. 2013. [Online]. Available: <https://doi.org/10.1016/j.cell.2013.09.034>.
- [14] H. Ma, C. Zhao, Z. Zhao, L. Hu, F. Ye, H. Wang, Z. Fang, Y. Wu, and X. Chen, “Specific glioblastoma multiforme prognostic-subtype distinctions based on DNA methylation patterns”, *Cancer Gene Therapy*, vol. 27, no. 9, pp. 702–714, Oct. 2019. [Online]. Available: <https://doi.org/10.1038/s41417-019-0142-6>.
- [15] A. F. Kazerooni, S. Bakas, H. S. Rad, and C. Davatzikos, “Imaging signatures of glioblastoma molecular characteristics: A radiogenomics review”, *Journal of Magnetic Resonance Imaging*, vol. 52, no. 1, pp. 54–69, Aug. 2019. [Online]. Available: <https://doi.org/10.1002/jmri.26907>.
- [16] C. Dupont, N. Betrouni, N. Reynolds, and M. Vermandel, “On image segmentation methods applied to glioblastoma: State of art and new trends”, *IRBM*, vol. 37, no. 3, pp. 131–143, Jun. 2016. [Online]. Available: <https://doi.org/10.1016/j.irbm.2015.12.004>.

BIBLIOGRAPHY

- [17] M. M. Letteboer, O. F. Olsen, E. B. Dam, P. W. Willem, M. A. Viergever, and W. J. Niessen, “Segmentation of tumors in magnetic resonance brain images using an interactive multiscale watershed algorithm1”, *Academic Radiology*, vol. 11, no. 10, pp. 1125–1138, Oct. 2004. [Online]. Available: <https://doi.org/10.1016/j.acra.2004.05.020>.
- [18] J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, “Efficient multilevel brain tumor segmentation with integrated bayesian model classification”, *IEEE Transactions on Medical Imaging*, vol. 27, no. 5, pp. 629–640, May 2008. [Online]. Available: <https://doi.org/10.1109/tmi.2007.912817>.
- [19] J. Jiang, Y. Wu, M. Huang, W. Yang, W. Chen, and Q. Feng, “3d brain tumor segmentation in multimodal MR images based on learning population- and patient-specific feature sets”, *Computerized Medical Imaging and Graphics*, vol. 37, no. 7-8, pp. 512–521, Oct. 2013. [Online]. Available: <https://doi.org/10.1016/j.compmedimag.2013.05.007>.
- [20] J. S. Cordova, E. Schreibmann, C. G. Hadjipanayis, Y. Guo, H.-K. G. Shu, H. Shim, and C. A. Holder, “Quantitative tumor segmentation for evaluation of extent of glioblastoma resection to facilitate multisite clinical trials”, *Translational Oncology*, vol. 7, no. 1, 40–W5, Feb. 2014. [Online]. Available: <https://doi.org/10.1593/tlo.13835>.
- [21] D. Selvaraj, “A review on tissue segmentation and feature extraction of mri brain images”,, 2013.
- [22] S. Udomhunsakul and P. Wongsita, “Feature extraction in medical MRI images”, in *IEEE Conference on Cybernetics and Intelligent Systems, 2004.*, IEEE. [Online]. Available: <https://doi.org/10.1109/iccis.2004.1460437>.
- [23] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation”, 2015. [Online]. Available: <https://arxiv.org/abs/1505.04597>.
- [24] L. Armi and S. F. Ershad, “Texture image analysis and texture classification methods - A review”, *CoRR*, vol. abs/1904.06554, 2019. arXiv: [1904.06554](https://arxiv.org/abs/1904.06554). [Online]. Available: <http://arxiv.org/abs/1904.06554>.

BIBLIOGRAPHY

- [25] C. G. Eichkitz, J. Davies, J. A. Amtmann, M. G. Schreilechner, and P. F. M. D. Groot, “Grey level co-occurrence matrix and its application to seismic data”, *First Break*, vol. 33, pp. 71–77, 2015.
- [26] J. Liu and Y. Shi, “Image feature extraction method based on shape characteristics and its application in medical image analysis”, in *Applied Informatics and Communication*, D. Zeng, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 172–178, ISBN: 978-3-642-23214-5.
- [27] L. Yang, F. Albregtsen, T. Lønnestad, and P. Grøttum, “Methods to estimate areas and perimeters of blob-like objects: A comparison.”, Jan. 1994, pp. 272–276.
- [28] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, “An introduction to decision tree modeling”, *Journal of Chemometrics*, vol. 18, no. 6, pp. 275–285, Jun. 2004. [Online]. Available: <https://doi.org/10.1002/cem.873>.
- [29] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, “Learning deep representations by mutual information estimation and maximization”, 2018. [Online]. Available: <https://arxiv.org/abs/1808.06670>.
- [30] A.-H. Jiang, X.-C. Huang, Z.-H. Zhang, J. Li, Z.-Y. Zhang, and H.-X. Hua, “Mutual information algorithms”, *Mechanical Systems and Signal Processing*, vol. 24, no. 8, pp. 2947–2960, 2010, ISSN: 0888-3270. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0888327010001585>.
- [31] N. Shah, X. Feng, M. Lankerovich, R. B. Puchalski, and B. Keogh, “Data from Ivy GAP”, Type: dataset, 2016. [Online]. Available: <https://wiki.cancerimagingarchive.net/x/jY9XAQ> (visited on 04/14/2022).
- [32] S. Pati, R. Verma, H. Akbari, M. Bilello, V. B. Hill, C. Sako, R. Correa, N. Beig, L. Venet, S. Thakur, P. Serai, S. M. Ha, G. D. Blake, R. T. Shinohara, P. Tiwari, and S. Bakas, “Multi-institutional paired expert segmentations and radiomic features of the ivy gap dataset”, 2020. [Online]. Available: <https://wiki.cancerimagingarchive.net/x/64MvB>.

BIBLIOGRAPHY

- [33] “G-mean is not suitable for medical image instance segmentation to find threshold when background is too much”, Apr. 2022. [Online]. Available: <https://pinkr1ver.com/posts/2022-04-03-g-mean-is-not-suitable-for-medical-image-instance-segmentation-to-find-threshold-when-background-is-too-much/> (visited on 04/14/2022).

Appendix A

Code

A.1 U-Net

```

import torch
from torch import nn
from torch.nn import functional as F

class convBlock(nn.Module):
    def __init__(self, inChannel, outChannel):
        super(convBlock, self).__init__()
        self.layer = nn.Sequential(
            nn.Conv2d(in_channels=inChannel, out_channels=outChannel, kernel_size=3,
                      padding_mode='reflect', bias=False),
            nn.BatchNorm2d(outChannel),
            nn.Dropout2d(0.3),
            nn.LeakyReLU(),
            nn.Conv2d(in_channels=outChannel, out_channels=outChannel, kernel_size=3,
                      padding_mode='reflect', bias=False),
            nn.BatchNorm2d(outChannel),
            nn.Dropout2d(0.3),
            nn.LeakyReLU()
        )

    def forward(self, x):

```

APPENDIX A. CODE

```
    return self.layer(x)

class downSample(nn.Module):
    def __init__(self, channel):
        super(downSample, self).__init__()
        self.layer = nn.Sequential(
            nn.Conv2d(in_channels=channel, out_channels=channel, kernel_size=3, stride=2,
                      padding_mode='reflect', bias=False),
            nn.BatchNorm2d(channel),
            nn.LeakyReLU()
        )

    def forward(self, x):
        return self.layer(x)

class upSample(nn.Module):
    def __init__(self, channel):
        super(upSample, self).__init__()
        self.layer = nn.Conv2d(channel, channel // 2, 1, 1)

    def forward(self, x, featureMap):
        up = F.interpolate(x, scale_factor=2, mode='nearest')
        out = self.layer(up)
        return torch.cat((out, featureMap), dim=1)

class UNet(nn.Module):
    def __init__(self):
        super(UNet, self).__init__()
        self.c1 = convBlock(1, 64)
        self.d1 = downSample(64)
```

APPENDIX A. CODE

```
self.c2 = convBlock(64, 128)
self.d2 = downSample(128)
self.c3 = convBlock(128, 256)
self.d3 = downSample(256)
self.c4 = convBlock(256, 512)
self.d4 = downSample(512)
self.c5 = convBlock(512, 1024)
self.u1 = upSample(1024)
self.c6 = convBlock(1024, 512)
self.u2 = upSample(512)
self.c7 = convBlock(512, 256)
self.u3 = upSample(256)
self.c8 = convBlock(256, 128)
self.u4 = upSample(128)
self.c9 = convBlock(128, 64)
self.out = nn.Conv2d(64, 1, 3, 1, 1)
self.th = nn.Sigmoid()

def forward(self, x):
    L1 = self.c1(x)
    L2 = self.c2(self.d1(L1))
    L3 = self.c3(self.d2(L2))
    L4 = self.c4(self.d3(L3))
    L5 = self.c5(self.d4(L4))
    R4 = self.c6(self.u1(L5, L4))
    R3 = self.c7(self.u2(R4, L3))
    R2 = self.c8(self.u3(R3, L2))
    R1 = self.c9(self.u4(R2, L1))

    return self.th(self.out(R1))
```

A.2 Read NIfTI file and convert it to PNG file

```

def read_nii_image(niifile):
    for root, dirs, files in os.walk(niifile):
        for file in files:
            if '.nii.gz' in file and '.png' not in file:
                savePathAX = os.path.join(root, file.replace('.nii.gz', '_AX'))
                savePathSAG = os.path.join(
                    root, file.replace('.nii.gz', '_SAG'))
                savePathCOR = os.path.join(
                    root, file.replace('.nii.gz', '_COR'))
                fileName = file.replace('.nii.gz', '')
                if not os.path.isdir(savePathAX):
                    os.mkdir(savePathAX)
                if not os.path.isdir(savePathSAG):
                    os.mkdir(savePathSAG)
                if not os.path.isdir(savePathCOR):
                    os.mkdir(savePathCOR)

                # read nii files
                img_path = os.path.join(root, file)
                img = nib.load(img_path)
                img_fdata = img.get_fdata()

                # transfer nii2png
                (x, y, z) = img.shape
                for i in range(z):
                    slice = img_fdata[:, :, i]
                    print(file)
                    imageio.imwrite(os.path.join(
                        savePathAX, '{}.png'.format(i)), slice)

                for i in range(y):

```

```

        slice = img_fdata[:, i, :]
        print(file)
        imageio.imwrite(os.path.join(
            savePathCOR, '{}.png'.format(i)), slice)

    for i in range(x):
        slice = img_fdata[i, :, :]
        print(file)
        imageio.imwrite(os.path.join(
            savePathSAG, '{}.png'.format(i)), slice)

```

A.3 Feature Extraction

```

import os
import pandas as pd
from six import u
from data import *
from radiomics import featureextractor
import SimpleITK as sitk

def ROI_not_one_dim(image_array):
    flag_i = False
    flag_j = False
    for val in (85, 170):
        image_array[image_array == val] = 255
    for i in range(image_array.shape[0] - 1):
        for j in range(image_array.shape[1] - 1):
            if(image_array[i + 1, j] == 255 and image_array[i, j] == 255):
                flag_i = True
            if(flag_i and flag_j):
                return True
            if(image_array[i, j] == 255 and image_array[i, j + 1] == 255):

```

```

flag_j = True
if(flag_i and flag_j):
    return True
return False

basePath = r''
dataPath = os.path.join(basePath, 'data')
dataFile = 'GBM_MRI_Dataset.csv'

if __name__ == '__main__':
    params = os.path.join(dataPath, "Params.yaml")
    FeatureDataset = FeatureExtractionDataset(dataPath, dataFile)
    i = FeatureDataset[30]
    if platform.system() == 'Linux' or platform.system() == 'Darwin':
        imageName = os.path.join(dataPath, (i['ImagePath'].loc[0]).replace('\\', '/'))
        maskName = os.path.join(dataPath, (i['MaskPath'].loc[0]).replace('\\', '/'))
    elif platform.system() == 'Windows':
        imageName = os.path.join(dataPath, i['ImagePath'].loc[0])
        maskName = os.path.join(dataPath, i['MaskPath'].loc[0])
    #print(imageName)
    extractor = featureextractor.RadiomicsFeatureExtractor(params)

    mask = sitk.ReadImage(maskName)
    mask_arr = sitk.GetArrayFromImage(mask)
    flag = 0
    if np.any(mask_arr):
        for val in (85, 170):
            mask_arr[mask_arr == val] = 255

    mask_merged = sitk.GetImageFromArray(mask_arr)
    mask_merged.CopyInformation(mask)

    feature = extractor.execute(imageName, mask_merged)

```

APPENDIX A. CODE

```

feature_row = pd.DataFrame({})

for key, val in feature.items():
    feature_row[key] = pd.Series(val)

result = pd.concat([i, feature_row], axis=1)
result_all = pd.DataFrame({}, columns=result.columns)

for i in FeatureDataset:
    if platform.system() == 'Linux' or platform.system() == 'Darwin':
        imageName = os.path.join(dataPath, (i['ImagePath'].loc[0]).replace('\\',
        maskName = os.path.join(dataPath, (i['MaskPath'].loc[0]).replace('\\',
    elif platform.system() == 'Windows':
        imageName = os.path.join(dataPath, i['ImagePath'].loc[0])
        maskName = os.path.join(dataPath, i['MaskPath'].loc[0])
#print(imageName)
extractor = featureextractor.RadiomicsFeatureExtractor(params)

mask = sitk.ReadImage(maskName)
mask_arr = sitk.GetArrayFromImage(mask)
if np.any(mask_arr) and ROI_not_one_dim(mask_arr):
    for val in (85, 170):
        mask_arr[mask_arr == val] = 255

mask_merged = sitk.GetImageFromArray(mask_arr)
mask_merged.CopyInformation(mask)

feature = extractor.execute(imageName, mask_merged)
feature_row = pd.DataFrame({})
...
for key, val in feature.items():
    if 'diagnostics' not in key:
        feature_row[key.replace('original_', '')] = pd.Series(val)

```

APPENDIX A. CODE

```
    ...
    for key, val in feature.items():
        feature_row[key] = pd.Series(val)

    result = pd.concat([i, feature_row], axis=1)
    result_all = result_all.append(result, ignore_index=True)
    flag += 1
    print(flag)

result_all.to_csv(os.path.join(dataPath, 'feature_extraction.csv'), index=False)
```