Radiomics Enhanced Machine Learning-Based Classifier to Improve Survival Estimation in Glioblastoma Multiforme

by

Abdulkerim Duman

A thesis submitted to fulfil the requirements for the degree

of Doctor of Philosophy in Engineering.

School of Engineering

CARDIFF UNIVERSITY

UK

May 2025

Thesis Abstract

In the field of neuro-oncology, precision oncology is advancing to improve patient survival outcomes. Radiomics, mainly through the use of standardised engineered (hand-crafted) features, has recently been utilised in neuro-oncology research and holds potential as biomarkers in the diagnosis and treatment planning of glioblastoma multiforme (GBM). However, the use of multiparametric Magnetic Resonance Imaging (mpMRI) data and the inclusion of datasets from multiple institutions pose substantial challenges for reproducibility. Therefore, establishing a standardised preprocessing pipeline and developing interpretable radiomic models could enhance the transition of radiomic studies in clinical settings.

This thesis investigated the optimisation of preprocessing pipelines to enhance reproducibility. The research addressed artefacts in registration and resampling on a widely used preprocessing pipeline from the Brain Tumor Segmentation Challenge (BraTS) and proposed an optimised version of this pipeline. For our domain-specific dataset (STORM_GLIO) with clinically defined contours used in radiotherapy treatment planning, we designed a preprocessing pipeline that excludes registration to a comprehensive Magnetic Resonance Imaging (MRI)-based reference of normal adult human brain anatomy and integrates a state-of-the-art brain extraction tool to improve accuracy and consistency. The proposed pipeline was assessed by our clinicians. Results demonstrated that the widely adopted preprocessing pipeline can be reliably reproduced through these optimisations, thereby ensuring consistency with our domain-specific clinical requirements.

In addition, a resource-efficient strategy, Region-Focused Selection Plus (RFS+), for enhanced automated tumour segmentation was implemented using state-of-the-art models to improve generalisability. By introducing weighted ensemble learning alongside different normalisation techniques (such as Z-score and Nyul), RFS+ enhanced segmentation performance and model generalisability when the model training process utilised three segmentation approaches (Multi-label, Binary class,

and Multiclass), incorporating tumour-specific characteristics such as overlapping and non-overlapping regions. Also, the strategy demonstrated competitive results, resource-efficiency, flexibility, as it can be applied to different models, including U-Net and nnU-Net.

The radiomic analysis studies focused on evaluating the effectiveness of using a limited number of radiomic features (RFs) from MRI sequences, in accordance with current radiomic study guidelines. For the radiomic analysis, RFs were combined with a single clinical variable due to incomplete clinical information across datasets. These studies were developed for overall survival (OS) prediction in GBM under two different settings while maintaining model interpretability by limiting the feature set to a maximum of 10 features. First, on BraTS 2020 and RHUH-GBM datasets, which used the same contouring format from the BraTS Challenge and were pre-processed through the widely used pipeline, we developed a novel hybrid feature selection method (LASSO-PSO). LASSO-PSO boosted radiomic model performance and achieved generalisable, state-of-the-art results, supported by external validation. Second, a radiomic model was developed using as few as two robust and reproducible RFs since many RFs extracted from the BraTS 2020 and STORM_GLIO datasets showed higher instability. This instability stemmed from differences in preprocessing pipelines and contouring formats across datasets. The radiomic model achieved moderate C-index performance when utilising a single contour and MRI sequence, suggesting potential applicability across different clinical challenges and limitations.

Acknowledgements

I would like to express gratitude to my primary supervisor, Prof. Dr. Emiliano Spezi, for his invaluable contributions and the guidance he has provided throughout this journey. His support and insightful feedback have played a crucial role in shaping both the direction and the quality of this work.

Secondly, I would like to thank my secondary supervisor, Dr. Xianfang Sun. I am grateful for the valuable guidance and the knowledge he has shared with me. His direction has been essential in helping me navigate my research, and your support has greatly impacted this study.

I would like to express my appreciation to all the academics who contributed to this research. Also, it was valuable to share an office space with the LIDA team. I appreciate the time spent together and the experiences we shared during this period.

I am thankful to my wife, Ecmel, my son, Erin, and my entire family, including my parents and siblings. Your steady support has accompanied me through every step of this journey, and I am profoundly grateful.

I also appreciate the financial support from the Turkish Ministry of National Education, which made the successful completion of this research possible.

Publications and Output

Key Publications:

Duman, A., Sun, X., Thomas, S., Powell, J.R. and Spezi, E., (2024). Reproducible and interpretable machine learning-based radiomic analysis for overall survival prediction in glioblastoma multiforme. Cancers, 16(19), p.3351. https://doi.org/10.3390/cancers16193351

Duman, A., Karakuş, O., Sun, X., Thomas, S., Powell, J., & Spezi, E. (2023). RFS+: A clinically adaptable and computationally efficient strategy for enhanced brain tumour segmentation. Cancers, 15(23), 5620. https://doi.org/10.3390/cancers15235620

Pre-print:

Duman, A., Sun, X., Powell, J.R. and Spezi, E., (2025). A Novel Swarm Intelligence-Driven Feature Selection for Interpretable Machine Learning in GBM Overall Survival Analysis. medRxiv, p.2025.04.16.25325927. https://doi.org/10.1101/2025.04.16.25325927

Conference Poster and Presentations:

Duman, A., Whybra, P., Powell, J., Thomas, S., Sun, X., & Spezi, E. (2023). PO-1620 Transferability of deep learning models to the segmentation of gross tumour volume in brain cancer. Radiotherapy and Oncology, Volume 182, S1315 - S1316. http://doi.org/10.1016/S0167-8140(23)66535-1

Duman, A., Powell, J., Thomas, S., Sun, X., & Spezi, E. (2024). Generalisability of Deep Learning Models on Brain Tumour Segmentation. In: Spezi E. & Bray M (eds.), Proceedings of the Cardiff University Engineering Research Conference 2023. Cardiff: Cardiff University Press. https://doi.org/10.18573/conf1.b

Duman, A., Powell, J., Thomas, S., & Spezi, E. (2024). Evaluation of Radiomic Analysis over the Comparison of Machine Learning Approach and Radiomic Risk Score on Glioblastoma. In: Spezi E. & Bray M (eds.), Proceedings of the Cardiff University Engineering Research Conference 2023. Cardiff: Cardiff University Press. https://doi.org/10.18573/conf1.f

Duman, A., Powell, J., Thomas, S., Sun, X., Spezi, E. (2024) Radiomics-based Risk Stratification for GBM: Training, Validation, and Clinical Applicability. Radiotherapy and Oncology, Volume 194, S5145 - S5148. http://doi.org/10.1016/S0167-8140(24)03079-2

Doherty, C., **Duman, A.**, Chuter, R., Hutton, M., & Spezi, E. (2024). Investigating the Feasibility of MRI Auto-segmentation for Image Guided Brachytherapy. In: Spezi E. & Bray M (eds.), Proceedings of the Cardiff University Engineering Research Conference 2023. Cardiff: Cardiff University Press. https://doi.org/10.18573/conf1.d/

Kim, K, **Duman, A.**, Spezi, E. (2024). RGU-Net: Computationally Efficient U-Net for Automated Brain Extraction of mpMRI with Presence of Glioblastoma. In: Spezi E. & Bray M (eds.), Proceedings of the Cardiff University School of Engineering Research Conference 2024. Cardiff: Cardiff University Press. https://doi.org/10.18573/conf3.h

Contributions

This thesis is my own words. The contributions with published materials were linked as follows:

(a) As presented in Chapter 2, the study served as the preprocessing pipeline for the radiomic analysis and the automated segmentation study on the local dataset.

- Our study showed that the Brain Tumor Segmentation Challenge (BraTS) preprocessing pipeline was not directly applicable to our local dataset (STORM_GLIO), which utilises clinically defined contours (Gross Tumour Volume, GTV) for radiotherapy treatment planning. Therefore, we optimised the pipeline to minimise potential contour distortions and MRI artefacts. These modifications included integrating a state-of-the-art skull stripping technique and removing registration to the MRI-based normal brain reference from the preprocessing pipeline.

- The proposed pipeline's output was validated by our clinicians, enabling interoperability between the tumour core (TC) and GTV contouring formats in segmentation and radiomic analysis.

(b) As presented in Chapter 3, the research explored novel feature selection strategies for survival prediction using a widely used preprocessing pipeline together with its corresponding contouring format:

- A novel hybrid feature selection method was developed to select up to 10 radiomic features (RFs) by following radiomic guidelines to enhance generalisability and interpretability of radiomic models across multi-institutional datasets.

- The model yielded performance comparable to that of leading Deep Learning (DL) models in stratification ability.

(c) As presented in Chapter 4, the study evaluated risk stratification across multi-institutional datasets, testing different preprocessing pipelines and assessing contouring interoperability by comparing the BraTS Challenge contours and clinically defined contours for radiomic overall survival (OS) analysis:

- The study conducted a robustness analysis with image perturbation techniques to identify stable RFs, addressing instability observed across open-access and local GBM datasets processed with different preprocessing methods and contouring formats.

- This study leveraged the largest reported cohort for OS analysis, including 289 GBM patients from multiple institutions. To support interpretability and address real-world limitations, the study explored a minimal feature set, with only age as the clinical variable and two stable RFs, using minimal imaging input: a single contour and one MRI sequence.

(d) As presented in Chapter 5, the study searched for computationally efficient strategies for enhanced brain tumour segmentation under the domain-specific requirements of our local dataset:

- The study proposed a strategy to improve the generalisability of DL models trained on TC segmentation (the BraTS contouring format) for clinically defined contours, specifically GTV.

- By integrating different intensity normalisation techniques (Z-score and Nyul) during preprocessing and combining the outputs of the trained models (across three segmentation approaches that incorporate tumour characteristics) through weighted ensemble learning, RFS+ enhanced generalisability and performance under limited VRAM and training time.

Contents

List of Figures

List of Tables

List of Abbreviations

3D:	Three-Dimensional.
3D-CRT:	Three-Dimensional Conformal Radiation Therapy.
3D-DSN:	3D Deeply Supervised Networks.
ADC:	Apparent Diffusion Coefficient.
AI:	Artificial Intelligence.
AUC:	Area Under the Curve.
BBB:	Blood-Brain Barrier.
BraTS:	the Brain Tumor Segmentation Challenge.
CaPTk:	the Cancer Imaging Phenomics Toolkit.
CAR:	Chimeric Antigen Receptor.
CI:	Confidence Index.
C-Index	Concordance Index.
CNN:	Convolutional Neural Network.
CNS:	Central Nervous System.
Cox-LASSO:	Regularised Cox Regression.
CSC:	Cancer Stem Cells.
CT:	Computed Tomography.
CTV:	Clinical Target Volume.
CV:	Cross-Validation.

DICOM:	Digital Imaging and Communications in Medicine.
DL:	Deep Learning.
DNA:	DeoxyriboNucleic Acid.
DSC:	Dice Similarity Coefficient.
ED:	Peritumoral Oedema.
EOR:	Extent of Resection.
ET:	Enhancing Tumour.
FBN:	Fixed Bin Number.
FBS:	Fixed Bin Size.
FCM:	Fuzzy C-means.
FLAIR:	Fluid-Attenuated Inversion Recovery.
GA:	Genetic Algorithms.
GBM:	GlioBlastoma Multiforme.
GBS:	Gradient Boosting Survival.
GLDZM:	the Grey Level Distance Zone Matrix.
GTR:	Gross Total Resection.
GTV:	Gross Tumour Volume.
HD95:	the 95th percentile Hausdorff distance.
HR:	Hazard Ratio.
HU:	Hounsfield Units.

iAUC:	integrated Area Under the Curve.
IBSI:	the Image Biomarker Standardisation Initiative.
ICC:	the Intra-class Correlation Coefficient.
KDE:	Kernel Density Estimate.
KM:	Kaplan-Meier.
LASSO:	Least Absolute Shrinkage and Selection Operator.
LPS:	the Left-Posterior-Superior.
LSQ:	Least Squares.
METRICS:	METhodological RadiomICs Score.
MGMT:	O6-Methylguanine-DNA Methyltransferase.
ML:	Machine Learning.
mpMRI:	multiparametric Magnetic Resonance Imaging.
MRI:	Magnetic Resonance Imaging.
mRMR:	Minimal Redundancy Maximum Relevance.
MutInfo:	Mutual Information.
NAWM:	Normal Appearing White Matter.
NCR:	Necrotic.
NET:	Non-Enhancing Tumour core.
NHS:	the National Health Service.
NIfTI:	Neuroimaging Informatics Technology Initiative.

NTR:	Near Total Resection.
OOB:	the Out-Of-Bag Bootstrap.
OS:	Overall Survival.
PACS:	Picture Archiving and Communication System.
PET:	Positron Emission Tomography.
PSO:	Particle Swarm Optimisation.
RANO:	Response Assessment in Neuro-Oncology.
RFs	Radiomic Features.
RFS:	Region-Focused Selection.
RFS+:	Region-Focused Selection Plus.
RHUH-GBM:	The Río Hortega University Hospital Glioblastoma Dataset.
ROI:	Regions Of Interest.
RP:	Radiofrequency Pulse.
RSF:	Random Survival Forests.
RTSTRUCT:	Radiotherapy Structure format.
SI:	Swarm Intelligence.
SPAARC:	SPAARC Pipeline for Automated Analysis and Radiomics Computing.
SRI24:	A comprehensive MRI-based reference of normal adult human brain anatomy.
T1:	T1-weighted.

T1ce:	T1-weighted Contrast-enhanced.
T2:	T2-weighted.
TC:	Tumour Core.
TCIA:	The Cancer Imaging Archive.
TE:	Time to Echo.
TMZ:	Temozolomide.
TR:	Repetition Time.
UPenn-GBM:	The University of Pennsylvania glioblastoma dataset.
UPHS:	The University of Pennsylvania Health System.
VOI:	Volumes Of Interest.
WHO:	World Health Organization.
WT:	Whole Tumour.
XAI:	Explainable AI.
ZDNU:	The Zone Distance Non-Uniformity.

1. Introduction

1.1 Outline

1.1.1 Aim of the work

The primary objective of this thesis is to develop a radiomics-enhanced machine learning (ML) classifier designed to assist clinicians in facilitating more personalised treatment decisions and potentially leading to enhancing survival outcomes for patients with glioblastoma multiforme (GBM). The aims to optimise radiomic analysis for the survival analysis of GBM are as follows:

i) Establishing a radiomic workflow in clinical settings that utilises standardised radiomic features (RFs) through the implementation of various preprocessing techniques, such as intensity normalisation, voxel resampling, and discretisation, while minimising resource requirements through the utilisation of a single clinical contour (Gross Tumour Volume, GTV) and one Magnetic Resonance Imaging (MRI) sequence. This approach addresses practical limitations regarding available imaging data and contouring resources in routine clinical practice.

ii) Enhancing automated segmentation that enables accurate, standardised, reproducible delineation of regions of interest (ROI) within medical images, improving efficiency and reliability in clinical settings. This automated approach establishes the essential groundwork for reliable downstream processing, facilitating reproducible radiomic feature extraction and comprehensive analytical evaluation.

iii) Developing a highly interpretable ML model that incorporates a minimal number of robust RFs and a limited number of MRI sequences for GBM survival analysis in clinical settings.

By achieving these objectives, the study seeks to create streamlined and transparent tools that facilitate reliable survival predictions and enhance the quality and personalisation of radiotherapy treatment planning, ultimately supporting clinical decision-making in the management of GBM patients. This research is specifically

designed to target patients diagnosed with GBM, a highly aggressive and malignant form of primary brain cancer classified as grade IV, according to the World Health Organization (WHO). Although the primary focus of this study is on GBM, the methodologies developed and employed in this thesis research have the potential to be extrapolated to other gliomas and may provide valuable insights to inform future studies in the realm of neuro-oncology, thereby contributing to the advancement of our understanding and treatment of malignant brain tumours.

1.1.2 Thesis Structure

Figure 1.1 presents a schematic map illustrating the flow and relationships between chapters.

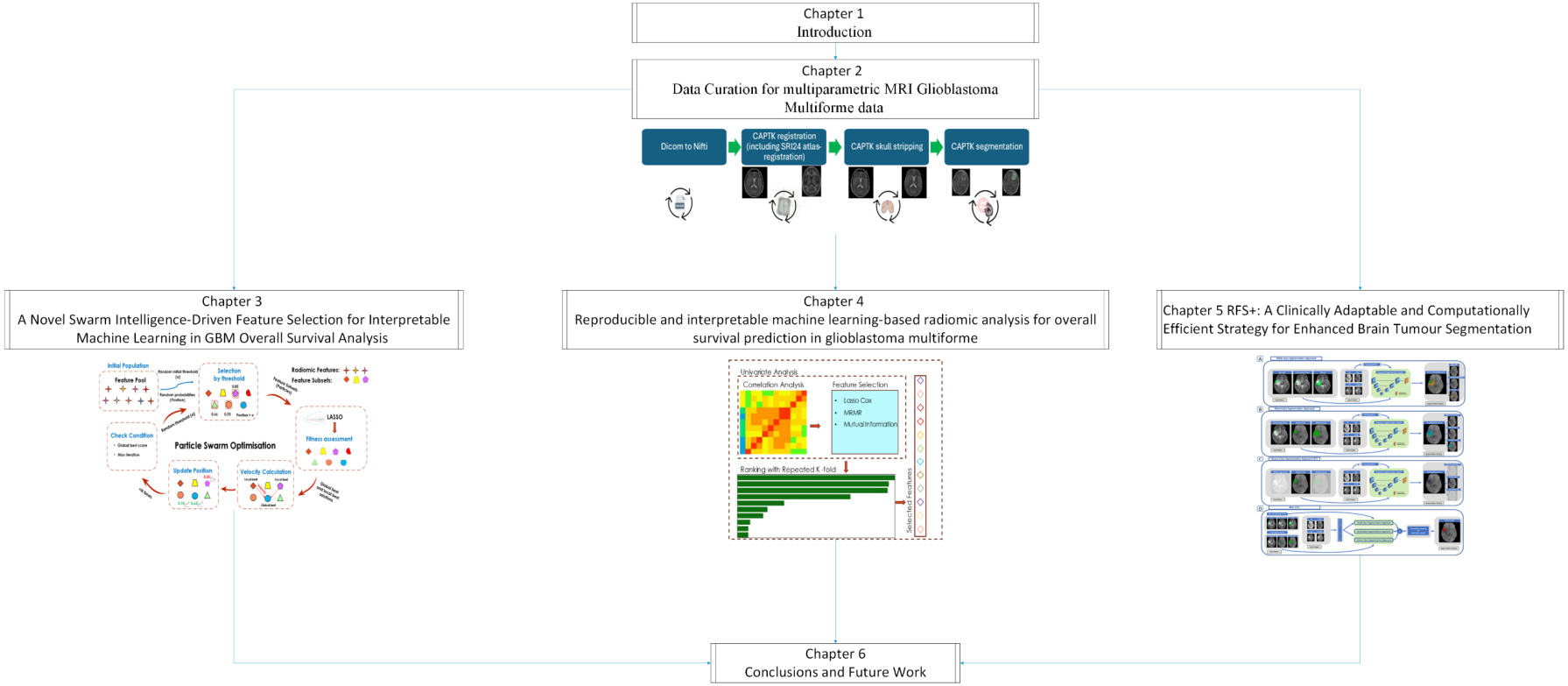


Figure 1.1 The thesis overview.

Chapter 1 provides the necessary background for the work presented. It introduces key concepts of GBM, medical imaging in GBM, treatments, and radiomics workflow. A brief general literature overview is provided, with more in-depth and critical discussions of specific literature appearing in subsequent chapters.

Chapter 2 provides a comprehensive overview of the preprocessing pipeline used in the Brain Tumor Segmentation Challenge (BraTS) for MRI sequences specific to GBM, including comparative analyses. It further details subsequent modifications made to the pipeline to facilitate radiomics analysis and support automated segmentation research in clinical settings, which are presented in Chapters 3, 4, and 5.

Chapter 3 develops a radiomic-based overall survival (OS) analysis, employing a novel hybrid Swarm Intelligence (SI)-based feature selection method to maximise the predictive performance under radiomic research guidelines. Designed for clinical applications and aligned with radiomic study guidelines, this approach integrates interpretable, traditional ML models to improve prognostic assessment. The chapter evaluates various feature selection methods, including the established Least Absolute Shrinkage and Selection Operator (LASSO)-based ranking method and two novel hybrid feature selection approaches. The aim is to maximise the risk-stratification model performance of OS analysis for GBM, incorporating up to ten RFs derived from three tumour regions (enhancing tumour (ET), tumour core (TC), and whole tumour (WT)) and two MRI sequences (T1-weighted (T1) and Fluid-Attenuated Inversion Recovery (FLAIR)).

Chapter 4 focuses on developing a reproducible and highly interpretable ML model for GBM survival analysis under clinical limitations. The model is designed to operate within clinical settings by utilising a minimal set of two robust RFs and MRI sequences, ensuring practical applicability while maintaining predictive performance.

Chapter 5 focuses on the development of Deep Learning (DL)-based auto-segmentation methods for brain tumours, with an emphasis on enhancing

automated segmentation techniques for clinical-based contouring (GTV). The chapter details the resource-efficient approach that enhances brain tumour segmentation in clinical settings by combining multiple models and normalisation techniques.

Chapter 6 provides a critical evaluation of the optimisation strategies explored for radiomic analysis, with a particular focus on their efficacy and potential for clinical implementation. The findings from this research not only consolidate the research presented in the thesis but also lay the groundwork for future studies in automated medical image analysis and personalised medicine in neuro-oncology.

1.2 Preview

This introductory chapter presents a comprehensive overview of GBM, highlighting the challenges associated with its treatment and the pivotal role of medical imaging in its management. It also discusses advanced medical image analysis techniques and outlines future directions, with a focus on emerging technologies such as novel imaging biomarkers and personalised treatment approaches.

1.3 Glioblastoma Multiforme

Globally, cancer affects millions of patients each year, with brain tumours representing a significant and particularly challenging subset. Among brain tumours, GBM stands out as the most frequent and aggressive primary malignancy. GBM accounts for approximately 57% of all gliomas and 48% of all primary malignant central nervous system (CNS) tumours. Its incidence increases with age, and it disproportionately affects men, accounting for approximately two-thirds of all cases. In the United States, the prevalence of GBM is reported to be 9.23 per 100,000 population [1]. GBM is characterised by its poor prognosis, with a median survival of only about 15 months [2], making it one of the most rapidly lethal forms of cancer.

1.4 Defining Brain Structure and Function

Understanding brain tumours requires a foundational knowledge of normal brain structure and function. The brain, as a part of the CNS, is a complex organ with distinct regions and cell types, each playing crucial roles in human physiology and cognition. The nervous system is broadly divided into two main components: the CNS and the peripheral nervous system. The CNS, which is the focus of this thesis, encompasses the brain and spinal cord. At its most basic level, the CNS is composed of neurons, the fundamental units of the nervous system, and supporting cells called neuroglia [3].

The brain itself can be categorised into three major divisions:

- Forebrain: This includes the cerebrum, the largest part of the brain responsible for higher-order functions, and the diencephalon.

- Midbrain: Connecting the pons and cerebellum to the forebrain.
- Hindbrain: Comprising the medulla oblongata, pons, and cerebellum.

The brainstem, a critical structure, includes the medulla oblongata, pons, and midbrain [4].

The CNS is comprised of two primary components: grey matter and white matter. Grey matter is characterised by the presence of neuronal cell bodies and dendrites, whereas white matter is composed exclusively of axons, also known as nerve fibres. The distinctive grey colouration of grey matter is due to the high concentration of neuronal cell bodies and their associated organelles. In contrast, white matter owes its name to the abundance of myelinated nerve fibres, which are enveloped by a lipid-rich myelin sheath comprising 70-80% lipid material [5]. This myelin sheath is responsible for the white appearance of white matter.

The nervous system covers two fundamental cell types: neurons and glial cells, which are shown in Figure 1.2. Neurons propagate electrical and chemical signals, while glial cells play important roles in modulating neuronal activity and supporting signal transmission [6]. Glial cells, accounting for approximately 90% of the brain's cellular composition, provide essential support and functional regulation for neurons. The CNS contains three main types of glial cells: astrocytes, oligodendrocytes, and microglia. Astrocytes are the most abundant glial cells, characterised by their star-shaped morphology. They provide essential structural and metabolic support to neurons. Oligodendrocytes, which have small cytoplasm, extend multiple processes that form myelin sheaths. Microglia, the smallest glial cells in the CNS, function as immune cells that remove debris[7]. Grade IV

astrocytoma, widely known as GBM, originates from astrocytes and is characterised by rapid proliferation and a poor prognosis [8].

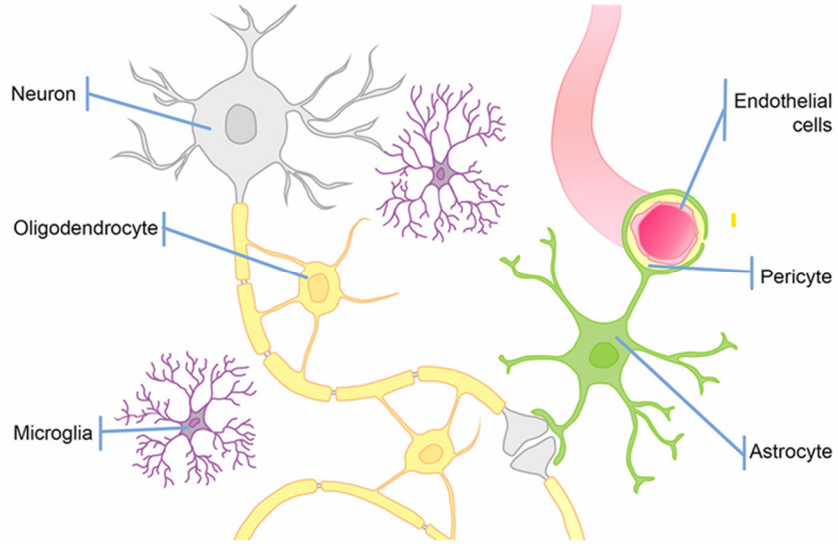


Figure 1.2 Healthy Brain Cells: Neuron, Oligodendrocyte, Astrocyte[8]. Astrocytic and oligoastrocytic glial tumour types of these healthy cells were classified as GBM[9].

1.5 Brain Tumours: Definition, Types, and Classification

A tumour is an abnormal growth of cells in the body, primarily caused by errors in the genetic code that controls cell division. These genetic alterations disrupt the normal cell cycle, inhibiting programmed cell death (apoptosis) and driving excessive cell growth (proliferation). As a result, these malfunctioning cells accumulate, forming a mass of tissue that we call a tumour. In the context of brain tumours, this process occurs within the confines of the non-elastic, stiff skull, leading to neurological symptoms and signs even before treatment begins [10].

Tumours can be classified into two main categories based on their behaviour and potential impact: i) Benign tumours: These are non-cancerous growths that generally remain localised. They tend to have well-defined borders and don't infiltrate surrounding tissues or metastasise to distant sites. However, their growth can still cause local pressure effects. ii) Malignant tumours: These cancerous

growths are characterised by their ability to invade surrounding tissues and potentially metastasise to other parts of the body through the bloodstream or lymphatic system. Brain tumours are further categorised based on their origin: i) Primary Brain Tumours: These originate in the brain itself and can be either low-grade or high-grade. ii) Secondary (Metastatic) Brain Tumours: These originate from cancers in other parts of the body and spread to the brain, which is the most frequent brain tumour in adults [11]. The WHO has established a grading system for brain tumours, which is crucial for treatment planning and prognosis: i) Grade 1 and 2: Low-grade tumours, which grow slowly. ii) Grades 3 and 4: High-grade tumours, which grow more rapidly. Glioblastoma, the focus of this thesis, is classified as a Grade 4 tumour in the 2021 WHO report on brain tumours [12], representing the highest level of malignancy. The first step in managing brain tumours, as with other cancers, is to achieve an accurate diagnosis by identifying the current extent of the disease, referred to as staging [13]. This process heavily relies on medical imaging, which forms the foundation for the radiomic analyses explored in this thesis. The unique challenges posed by brain tumours, such as their location within the skull and their potential to cause significant neurological deficits regardless of malignancy, underscore the critical importance of advanced imaging and analysis techniques in their management.

1.6 Traditional Treatments in Glioblastoma Multiforme

Although significant progress has been made in the molecular and cellular aspects of glioma biology, GBM patients have poor prognosis, highlighting the significant challenges in translating basic scientific insights into efficient clinical treatments. Currently, the conventional therapeutic options, encompassing surgical resection, radiotherapy, and temozolomide (TMZ) chemotherapy, yield a median OS of approximately 15-18 months [14] in selected patient cohorts enrolled in clinical trials, accompanied by a 5-year survival rate of under 10% [15], [16], [17]. GBM is a highly malignant and aggressive brain tumour that poses significant therapeutic challenges owing to its diffuse infiltrative growth pattern [18] and inherent resistance to established treatment modalities [19]. The standard therapeutic

strategy typically begins with extensive surgical resection to remove as much of the tumour tissue as possible.

1.6.1 Surgical Resection

Surgical resection of brain tumours emerged as a viable treatment option in the early 1980s. However, the field of neurosurgery experienced a paradigm shift with the advent of frameless stereotaxy in the 1990s. This groundbreaking innovation significantly advanced the precision and efficacy of surgical tumour removal techniques. Image-guided surgery has significantly enhanced the precision of surgical instrument placement in neurosurgical procedures. Advanced imaging techniques such as MRI have revolutionised the ability to accurately delineate tumour margins [20], [21].

The extent of tumour resection, often described as gross total resection (GTR), is an important determinant of treatment outcomes in brain tumour surgery. In the case of highly aggressive tumours such as GBM, advances in surgical methodologies and technologies have substantially improved the quality and effectiveness of treatments. A significant correlation has been established between the degree of tumour resection and patient survival rates, underscoring the importance of achieving GTR in the surgical management of brain tumours, particularly for patients with GBM [22].

Intraoperative imaging modalities, such as fluorescence-guided surgery, offer a real-time solution for accurately delineating tumour margins during neurosurgical procedures. This is particularly crucial in addressing the complex challenge of "brain shift", a phenomenon characterised by the dynamic displacement of brain tissue during surgery, resulting in discrepancies between the pre-operatively planned tumour location and its actual position in the operating room [23]. This imaging modality facilitates the identification of residual cancer tissue following tumour resection, thereby enabling a more precise removal of cancerous tissue and ultimately refining the surgical approach to improve patient outcomes.

1.6.2 Chemotherapy

Following surgical resection, the standard treatment protocol for GBM involves a multimodal approach combining TMZ chemotherapy with radiotherapy [24]. This regimen typically consists of a 6-week phase of concurrent TMZ and radiation, followed by an adjuvant TMZ phase. TMZ elicits its anti-tumour effects through a dual mechanism, involving both direct cytotoxic damage to tumour cells and indirect induction of programmed cell death pathways, including apoptosis, autophagy, and cellular senescence [25]. Additionally, TMZ has been shown to enhance the efficacy of concurrent radiotherapy, resulting in a synergistic increase in treatment efficiency [26]. Notwithstanding its therapeutic benefits, TMZ therapy is accompanied by considerable side effects, such as hematologic (blood-related) toxicity and thrombocytopenia (a decrease in thrombocyte count) [27].

1.6.3 Radiation therapy

Radiotherapy is an important treatment modality for GBM, especially for addressing microscopic cancer cells that are inaccessible to surgical resection. Modern radiotherapy employs X-ray photons, gamma photons, and protons, typically administered over a 6-week period. Three-dimensional conformal radiation therapy (3D-CRT) facilitates the delivery of precise radiation beams, informed by Computed Tomography (CT) and MRI guidance, with a 1-2 cm margin surrounding the tumour [28]. 3D-CRT utilises X-rays to target the tumour, inducing both direct and indirect Deoxyribonucleic acid (DNA) damage through low-linear energy transfer interactions. Despite the inherent complexity of 3D precise targeting, this approach enables the effective treatment of residual GBM cells while minimising side effects, thereby offering a therapeutic advantage over conventional methods [29]. The treatment of GBM has shown improved outcomes with the use of combination therapies. Notably, the application of carbon proton irradiation in conjunction with TMZ has demonstrated enhanced OS rates compared to TMZ paired with proton-induced irradiation. However, a significant obstacle hindering the effectiveness of radiotherapy at the cellular level is the issue of oxygenation, as the oxygen levels within cells play a crucial role in determining the success of radiotherapy, and hypoxic conditions can limit its efficacy [29], [30].

Heterogeneity in cancer is a challenging issue that significantly impacts the treatment of highly heterogeneous GBM. Additionally, drug delivery to the target site is hindered by multiple barriers: the blood-brain barrier (BBB) [31], cancer stem cells (CSCs) [32], the intertumoural heterogeneity in GBM [33] and the unique brain microenvironment [15], making it a major obstacle to overcome in GBM treatment. Researchers are exploring new treatments for glioblastoma, including immunotherapy (Chimeric antigen receptor (CAR) T-cell) [34], targeted therapy [35]. Nanomedicine is an emerging treatment approach for GBM, which seeks to effectively deliver therapeutic agents to the brain and is currently being evaluated in ongoing clinical trials [36]. Additionally, a ketogenic diet [37], high in fat and low in carbohydrates, has shown promise as an adjuvant therapy, potentially impairing cancer cell growth and survival. These novel approaches aim to improve the quality of GBM treatment. Ongoing clinical trials are investigating the effectiveness of these treatments, alone and in combination, to determine their optimal use in GBM treatment [38]. GBM treatment and management can be enhanced through personalised medicine (or precision medicine), aligning with the National Health Service (NHS) aims in the United Kingdom to enhance patient care [39]. This method can incorporate medical imaging analysis, specifically radiomics, as a tool for risk stratification and prognostic prediction [40]. Additionally, invasive methods such as biopsy have limitations due to the genetic heterogeneity of GBM tumours, as they can only obtain samples from a small, localised portion of the tumour [41]. In contrast, radiomics can assist a more comprehensive assessment of tumour characteristics by analysing the entire lesion non-invasively [42].

1.7 Diagnostic Imaging for Brain Tumours

Advanced imaging technologies are essential tools in modern medicine, providing non-invasive methods to visualise internal anatomical structures and physiological processes, particularly in the context of brain tumour diagnosis and treatment. These techniques allow clinicians to assess intracranial conditions without resorting to exploratory surgery, significantly improving patient care and outcomes. These imaging modalities serve multiple purposes throughout the patient care continuum, from initial detection and diagnosis to treatment planning, monitoring

therapy response, and long-term surveillance. The ability to generate detailed, three-dimensional (3D) representations of brain anatomy and function has revolutionised neuro-oncology, enabling more precise and personalised treatment strategies.

For individuals presenting with neurological symptoms, neuroimaging is a critical diagnostic step. Two primary modalities dominate this field: CT and MRI.

1.7.1 Computed Tomography

CT is a key imaging modality used to generate cross-sectional images representing a patient's anatomy [43]. The fundamental process involves an X-ray tube (source) and detector, with the patient positioned between them (shown in Figure 1.3a). The X-ray beam is rapidly rotated around the patient's body, producing tomographic (cross-sectional) images, or 'slices', which are then used by computer algorithms to reconstruct a 3D volume. The same figure demonstrated the first CT scan for prognostic tool (b) and the recent post-contrast CT scan (c).

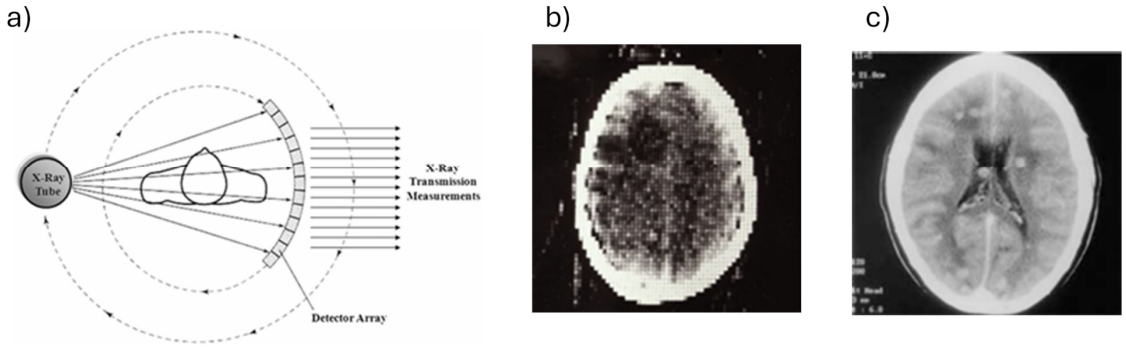


Figure 1.3 Introduction to Computed Tomography: a) a visual representation of a CT scanner [44] b) the first CT scan as a prognostic tool, 1971, London, UK [45] c) a modern post-contrast CT scan [46].

The principle behind CT imaging lies in the differential attenuation of X-rays by various tissues. Tissues with higher density, characterised by a higher atomic number, exhibit increased X-ray attenuation. This differential attenuation is quantified using Hounsfield Units (HU), creating a spectrum from air (-1000 HU) to bone (+1000 HU). Figure 1.4 illustrates the various tissue intensities of a human

brain in greyscale [47] and shows different tissue densities for CT brain scans in HU scale from -1000 to 1000+ for CT [48].

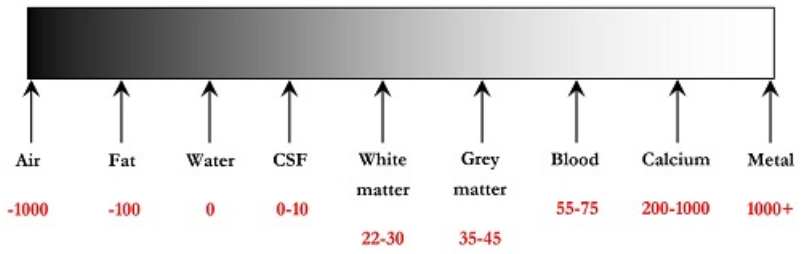


Figure 1.4 Hounsfield scale ranging from -1000 to 1000+ for different tissue intensities [48].

The ability of CT to differentiate tissues based on density makes it particularly useful for identifying certain types of brain abnormalities, especially those involving bone, bleeding, or calcification. However, it's important to note that CT involves ionising radiation, which carries potential health risks that must be considered when choosing imaging modalities. CT imaging offers several key benefits: a) Speed: CT scans can be performed rapidly, which is crucial for time-sensitive cases like trauma. b) Haemorrhage detection: CT excels in visualising blood in meningeal spaces and brain tissue. c) Paediatric trauma: CT is often the preferred choice for assessing injuries in children. d) Bone visualisation: CT provides exceptionally clear images of skull fractures and other bone abnormalities. e) Cost-effectiveness: CT is generally less expensive than MRI. However, CT also has limitations: a) soft tissue contrast: CT is less effective at detecting ischemia, infarcts, and brain oedema compared to MRI. b) Grey-white matter differentiation: CT cannot distinguish as clearly between grey and white matter as MRI can. These characteristics make CT an invaluable tool in emergency settings and for initial assessments, but it may be complemented by MRI for more detailed soft tissue evaluation in non-urgent scenarios.

1.7.2 Magnetic resonance imaging

MRI exploits the abundant presence of protons, particularly hydrogen nuclei, in human tissues to generate high-resolution images of anatomical structures[49], which is shown in Figure 1.5. The fundamental principle underlying MRI is based on the magnetic moments arising from the nuclear spin [50].

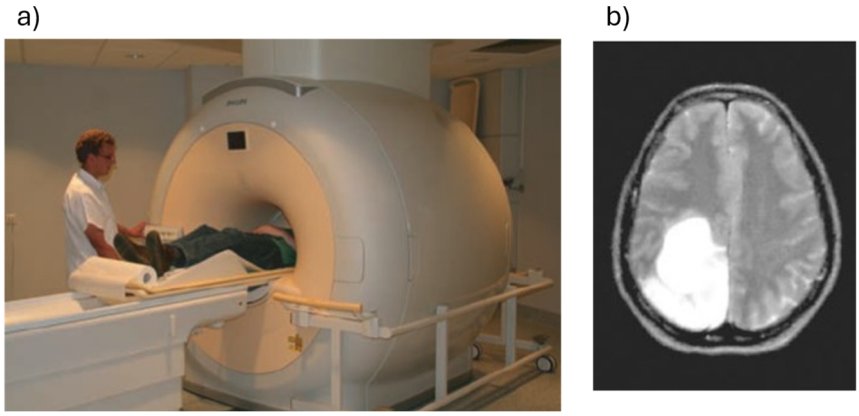


Figure 1.5 Introduction to Magnetic Resonance Imaging: a) A whole-body 3T MRI scanner is designed to provide high-resolution imaging of all anatomical regions. This system includes a superconducting magnet with a horizontal, solenoid main field. During imaging, the patient is positioned at the centre of the tunnel. b) T2-weighted (T2) axial brain scans with hyperintense tumoural lesion [51].

These protons, behaving analogously to miniature bar magnets, exhibit a spin characteristic that causes them to rotate or align when subjected to an external magnetic field [52]. In the absence of such a field, the magnetic moments of these protons are randomly oriented due to the influence of local magnetic fields generated by surrounding electrons. MRI technology harnesses the magnetic susceptibility of these protons to produce detailed images of the brain and other bodily structures [47]. The imaging process involves the application of a radiofrequency pulse (RP), which is essentially a brief transmission of radio waves within the magnetic field encompassing the patient. The intensity of this RP can be modulated depending on the specific imaging protocol employed. The concept of resonance in MRI refers to the phenomenon where protons absorb the radio wave energy when the frequency of the RP matches their precession frequency. Upon termination of the RP, excited protons revert to their equilibrium state through relaxation processes. This relaxation results in the emission of electromagnetic energy, manifesting as a detectable signal. In clinical MRI, this signal is typically measured in the form of an "echo". This echo is captured by specialised receiver coils and then processed and digitised to construct a visual representation of the targeted anatomical region, such as the brain [47]. The unique ability of MRI to manipulate and detect these subtle magnetic interactions at the atomic level allows for the

generation of highly detailed soft tissue images, making it an invaluable tool in neuroimaging and the diagnosis of brain tumours. The "echo" from relaxed protons is detected, and the variations in these time constants contribute to the construction of T1 and T2 images. T1 represents the time constant for protons realigning with the magnetic field axis, while T2 denotes the time constant for proton dephasing, also known as T2 decay. Increased T1 and T2 relaxation times result in darker T1 scans and brighter T2 scans compared to surrounding normal tissues (shown in Figure 1.6).

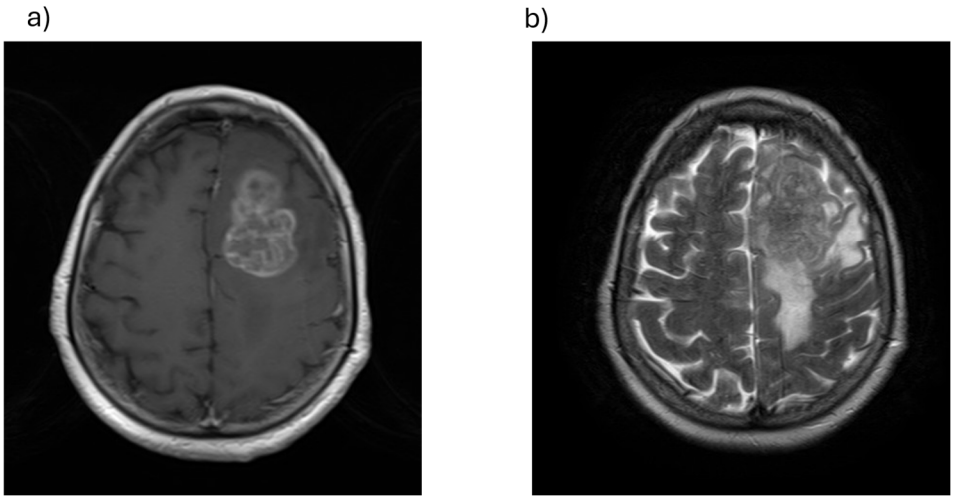


Figure 1.6 Magnetic Resonance Imaging Sequences: a) T1-weighted Contrast-enhanced (T1ce) axial MRI scan showing a large enhancing, isointense lesion with oedema b) T2 axial MRI scan showing a hypointense mass with surrounding oedema (from our local dataset: STORM_GLIO) [53].

By manipulating radiofrequency pulse timing and sequences, clinicians can selectively produce T1 or T2 images. Advanced techniques like diffusion-weighted imaging and magnetic resonance spectroscopy offer additional tumour characterisation capabilities [54]. A standard MRI system incorporates a superconducting magnet generating a static magnetic field, typically 1.5 or 3 Tesla. This magnet houses radiofrequency transmitter and receiver coils for nuclear spin excitation and signal detection, alongside gradient coils for spatial encoding. The detected radiofrequency signals undergo analogue-to-digital conversion and Fourier transformation to reconstruct images. While MRI offers superior soft tissue contrast compared to CT, it necessitates longer acquisition times. Patient motion,

including physiological movements, can induce phase errors and consequent image artifacts. Advanced acquisition strategies and fast imaging sequences, such as breath-hold sequences and cardiac gating, have been developed to mitigate motion-related artifacts and reduce scan duration. MRI provides unique tissue characterisation capabilities, which are particularly advantageous in neuroimaging. Notably, MRI's utilisation of non-ionising electromagnetic radiation presents a significant safety advantage over CT's X-ray-based approach [54]. Three primary terms are employed to describe signal intensities in T1 and T2 images: hyperintense, hypointense, and isointense. These descriptors are used to characterise lesions relative to healthy brain tissue. In T1 imaging, hyperintensity refers to a signal shift towards the appearance of fat tissue, manifesting as increased whiteness compared to surrounding brain tissue. Conversely, in T2 imaging, hyperintensity denotes a signal shift towards the appearance of cerebrospinal fluid, which typically appears white in normal subjects. In both T1 and T2 sequences, hypointensity describes a signal shift towards the appearance of air or bone, resulting in a darker appearance relative to surrounding brain tissue. Isointensity is characterised by similar grey shades or textures between the lesion and adjacent brain tissue, indicating comparable signal intensities [47]. These relative signal intensities play a crucial role in lesion detection and characterisation, aiding in differential diagnosis and treatment planning.

Comparative Analysis of MRI and CT: Advantages and Limitations

MRI offers significant advantages over CT in neuroimaging. Primarily, MRI's versatility allows for the visualisation of a wide spectrum of both physiological and pathological brain structures through various pulse sequence manipulations. Furthermore, MRI provides superior soft tissue contrast, enabling detailed characterisation of both normal and abnormal brain tissue. However, MRI is not without limitations. Several key disadvantages warrant consideration:

Limited sensitivity to acute haemorrhage: MRI may fail to adequately depict acute or subacute subarachnoid haemorrhage or intraparenchymal bleeding, potentially leading to missed diagnoses in critical cases.

Prolonged acquisition times: The extended duration required for MRI scanning renders it suboptimal for acute cases or trauma scenarios where rapid imaging is crucial.

Higher cost: MRI examinations generally require greater expenses compared to CT scans, which may impact resource allocation and patient access.

Acoustic noise: The considerable noise generated during MRI sequences can be problematic, particularly for paediatric patients or those with heightened sensitivity to auditory stimuli.

These factors underscore the importance of judicious selection between MRI and CT modalities based on clinical context, patient characteristics, and resource availability [47]. Additionally, MRI provides limited bone detail and suboptimal visualisation of calcifications, which can be crucial in specific diagnostic scenarios.

Magnetic Resonance Imaging for Characterising Brain Tissue

MRI offers multiplanar capabilities, allowing for visualisation of the brain in axial, coronal, and sagittal planes (Figure 1.7). While axial imaging remains the standard for routine brain examinations, all planes provide valuable diagnostic information and can be utilised based on specific clinical requirements.

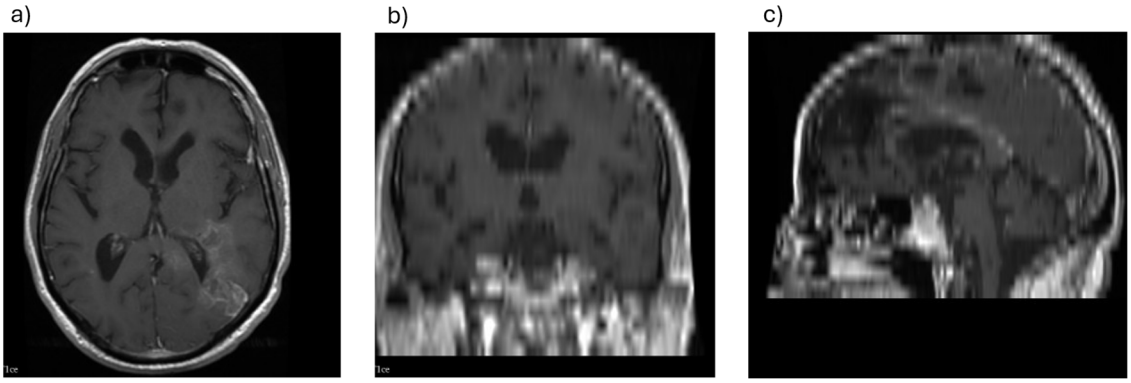


Figure 1.7 Planes for Brain Imaging: a) Axial plane b) Coronal plane c) Sagittal plane for T1ce MRI sequence (obtained from our local dataset: STORM_GLIO [53])

The superiority of MRI over CT in neuroimaging is primarily attributed to its enhanced spatial and contrast resolution. MRI facilitates exquisite delineation of brain anatomy and provides superior characterisation of pathologies, particularly those in proximity to the skull base, which are often poorly visualised on CT. The FLAIR sequence (shown in Figure 1.8), a specialised MRI technique, further augments the visibility of brain pathologies by suppressing the cerebrospinal fluid signal, thus increasing lesion conspicuity.

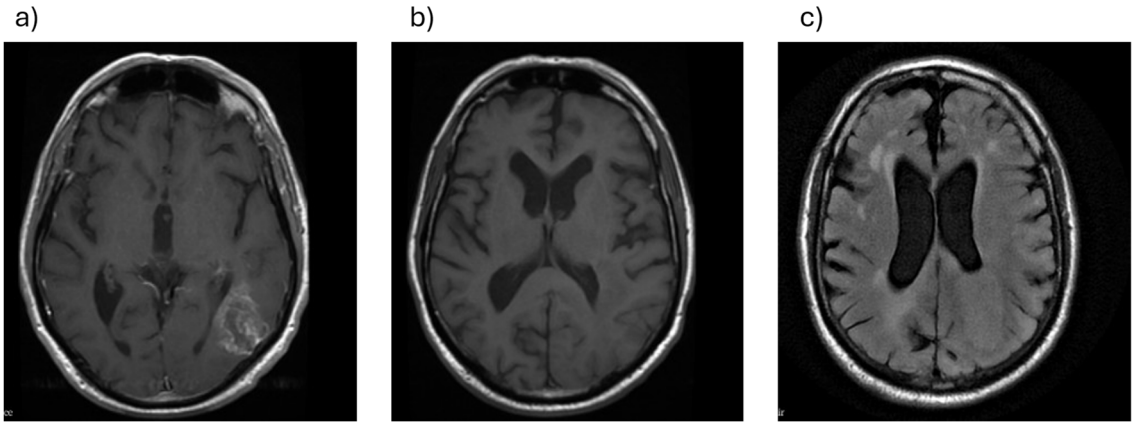


Figure 1.8 Widely used Magnetic Resonance Imaging Sequences: a) T1ce sequence, b) T1 sequence c) FLAIR Sequence for a GBM patient in the STORM_GLIO dataset [53].

The application of contrast enhancement in MRI significantly improves diagnostic yield, particularly in the evaluation of neoplastic processes. Paramagnetic contrast agents, typically gadolinium-based, accumulate in areas of BBB disruption, resulting in hyperintense signals on T1 images, thereby ET visibility and characterisation.

While the spectrum of intracranial neoplasms is vast, this research focuses specifically on GBM, an aggressive entity within the broader category of gliomas, tumours arising from glial cell lineages. GBMs typically exhibit heterogeneous signal characteristics on MRI. On T1 sequences, these lesions generally demonstrate hypointensity relative to healthy brain tissues, whereas T2 sequences reveal hyperintensity. This signal pattern reflects the complex histopathological features of GBM, including areas of necrosis, haemorrhage, and vasogenic oedema. The integration of advanced MRI techniques, including perfusion-weighted imaging, diffusion tensor imaging, and magnetic resonance spectroscopy, offers additional

avenues for tumour characterisation, treatment planning, and response assessment in GBM management. These modalities provide insights into tumour vascularity, cellular density, and metabolic profile, respectively, enhancing our understanding of tumour biology, potentially guiding personalised therapeutic approaches and assessing treatment response [55].

1.7.2.1 Repetition Time (TR) and Time to Echo (TE)

In MRI, two critical parameters that govern image contrast and quality are TR and TE, which are shown in Figure 1.9. These parameters are fundamental to pulse sequence design and optimisation. TR is defined as the temporal interval between successive radiofrequency excitation pulses applied to the same slice or volume of tissue. This parameter primarily influences T1 contrast in the resultant images. TE, conversely, refers to the duration between the initial radiofrequency excitation pulse and the peak of the echo signal during signal acquisition. TE is a critical determinant of T2 contrast in MRI images. The manipulation of TR and TE allows for the generation of various contrast mechanisms in MRI, including T1, T2, and proton density-weighted images. This versatility in contrast manipulation underlies the diagnostic utility of MRI across numerous clinical applications, particularly in neuroimaging for the characterisation of brain tumours.

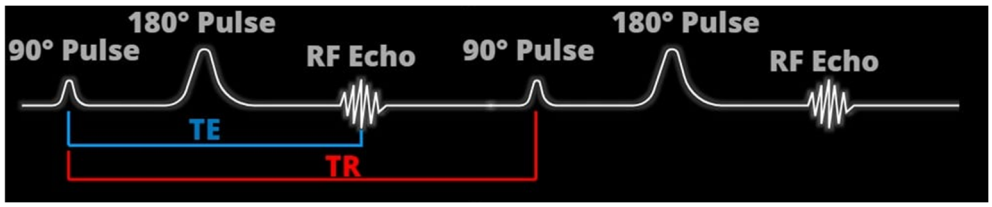


Figure 1.9 TR and TE of spin echo sequence[56].

MRI Sequences in Advanced Neuroimaging: MRI employs a variety of pulse sequences to generate images with different tissue contrasts, each offering unique diagnostic information. The most frequently utilised sequences in clinical practice are T1 and T2. T1 sequences are characterised by short TR and TE parameters. These sequences predominantly reflect the T1 relaxation properties of tissues, determining image contrast and signal intensity. T1 images are particularly useful for delineating anatomical structures and detecting fat-containing lesions.

Conversely, T2 sequences employ longer TR and TE times. The resultant images primarily reflect the T2 relaxation characteristics of tissues, providing excellent contrast for pathological processes associated with increased tissue water content, such as oedema or inflammation. FLAIR sequences represent a modification of T2 imaging, utilising an inversion recovery pulse to nullify the signal from cerebrospinal fluid. This technique enhances the visibility of periventricular and cortical lesions by suppressing the high signal intensity typically observed in fluids on T2 images. An additional important sequence in neuro-oncological imaging is T1ce scans. This technique involves the intravenous administration of a gadolinium-based contrast agent prior to image acquisition. Gadolinium, a non-toxic paramagnetic substance, enhances the visibility of lesions with disrupted BBB, such as many primary and metastatic brain tumours. The strategic application of these diverse MRI sequences allows for the comprehensive characterisation of brain pathologies, facilitating accurate diagnosis and treatment planning in neuro-oncology. The integration of advanced quantitative and functional MRI techniques further augments the diagnostic capabilities of conventional sequences, providing insights into tumour biology and treatment response. Advancements in neuroimaging techniques have expanded their applicability in the diagnosis and prognosis of GBM, facilitating their integration into personalised medicine (precision oncology) [40].

1.8 Precision Oncology in the Treatment of Glioblastoma Multiforme

Supporting the NHS's objective of improving treatment quality [39], personalised medicine is important in oncology, also known precision oncology, particularly for handling the challenges of GBM management [57]. Precision oncology describes a paradigm shift of therapeutic strategies in cancer, utilising a sophisticated approach that treats each tumour as a unique fingerprint rather than applying a one-size-fits-all strategy. It is a personalised approach to cancer treatment that aims to enhance the quality of care by tailoring therapy to individual tumours' unique genetic and molecular characteristics. This can be achieved through omics analysis, which encompasses genomics, pathomics, and radiomics, providing a comprehensive understanding of the tumour's genetic, pathological, and radiological

characteristics. By leveraging these precision biomarkers [58], precision oncology enables the identification of optimal therapy options, resulting in more effective treatment strategies that are specifically tailored to each patient's distinct tumour characteristics. By addressing the challenges outlined in the previous section and harnessing biomarkers, the outcomes for patients with GBM can be improved. Radiomics, the focus of this thesis, is poised to play a substantial role in the future of precision oncology [40].

1.9 Radiomics Overview

Radiomics is an evolving field in medical image analysis that contains the extraction of complex, high-dimensional quantitative features from medical images, offering potential enhancement for the understanding of tumour biology [59]. The diagnosis, grading, and characterisation of brain tumours typically rely on invasive biopsies [60]. However, due to the inherent genetic heterogeneity within GBM tumours, biopsy samples often yield limited information, as they are restricted to a small, localised region of the tumour tissue [41]. According to Beig et al. [42], image analysis utilising radiomics holds promise as a potential alternative to biopsies, especially in cases where biopsy procedures are not feasible or pose significant risks to the patient. Radiomic analysis targets to uncover hidden patterns and subtle features that are imperceptible to the human eye, transcending the limitations of visual evaluation by physicians and potentially enhancing the outcomes of personalised and precise patient care, as noted by Gillies et al. and Shaheen et al. [61], [62].

As artificial intelligence (AI) continues to advance, the field of radiomics has diverged into two distinct areas. Despite the well-defined engineered features grounded in mathematical concepts, the deep features extracted from AI-based medical imaging analysis are hindered by the "black box" issue, stemming from complex decision-making processes involving non-linear relationships. This is occurring against a backdrop of increasing calls for transparency in AI-driven medical image analysis. The opaque nature of AI poses significant challenges in healthcare, where interpreting the rationale behind decisions is crucial [63].

Although Explainable AI (XAI) techniques provide a means to glimpse into these models, they occasionally fail to capture the full scope of computational complexity, which could lead to discrepancies with established clinical practices [64]. Due to the limitations posed by the black box issue, this thesis focuses exclusively on engineered features.

On the other hand, the field of engineered radiomics faces significant challenges regarding reproducibility and validation. The diverse software implementations available can produce varying RFs from identical medical images, leading to inconsistent outcomes. Moreover, the lack of feature reproducibility across different datasets presents a substantial obstacle for the external validation of radiomics-based models [61], [65]. In response to these challenges, the Image Biomarker Standardisation Initiative (IBSI) conducted a comprehensive study examining radiomics reproducibility, with the goal of establishing standardised RFs for the clinical use [66].

Based on the radiomics guidelines [67], [68], there are several steps for radiomic analysis, namely: image acquisition, data curation (image processing), image segmentation, feature extraction, feature selection, and model building. The radiomic workflow is shown in Figure 1.10.

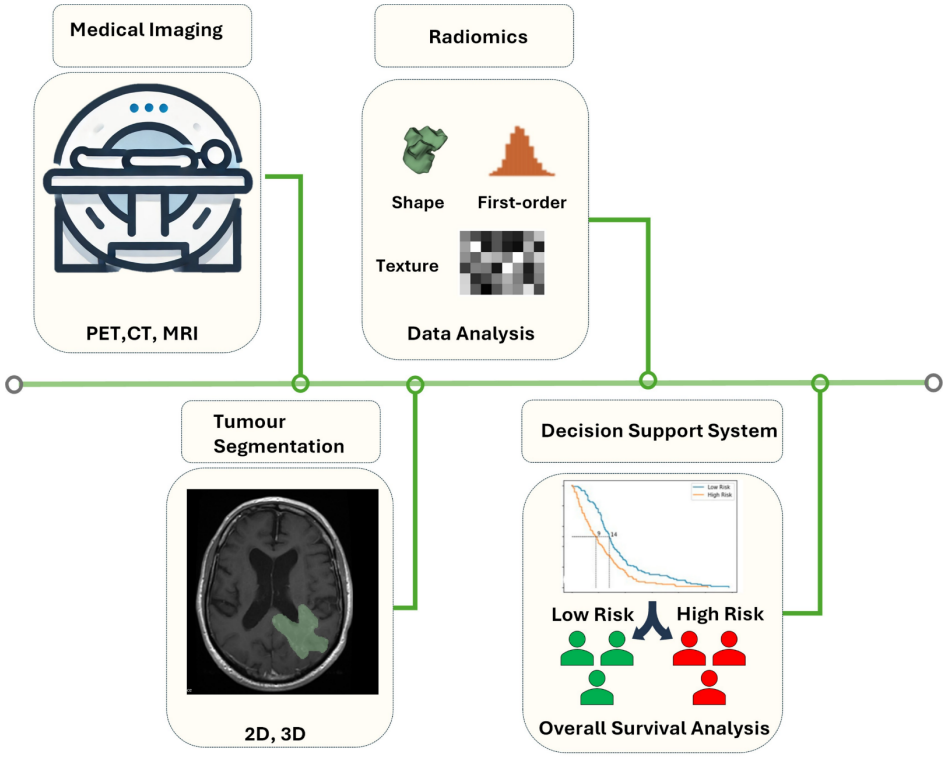


Figure 1.10 The radiomic workflow.

1.9.1 Acquisition and Data Curation

The acquisition step of the radiomics analysis requires a high-quality image, which is obtained using specialised scanners such as CT, MRI or Positron Emission Tomography (PET). These modern imaging devices have a wide range of acquisition and image reconstruction protocols, but standardisation across medical imaging centres is often challenging. While routine radiologic features used in clinical practice are not affected by this variability, variations in acquisition and reconstruction parameters can introduce changes in image analysis that are unrelated to biological effects, thereby impacting the extraction of meaningful information from numeric biomarkers in radiomics [61].

Complex medical imaging studies, like those utilising multi-sequence MRI, are becoming more prevalent in GBM research, the focus of this thesis. However, analysing the resulting data poses challenges due to intensity variability in MRI scans, which makes comparisons between study visits or subjects difficult. The intensity variability in MRI necessitates a critical preprocessing step known as

intensity normalisation, which reduces inconsistencies originating from variations in scanner parameters, patient positioning, and acquisition protocols. Additionally, this preprocessing step minimises significant intensity fluctuations across imaging datasets, which could compromise the accuracy of radiomic analyses [69], [70], [71]. This standardisation process improves the comparability of MRI scans across different imaging sessions and subjects, enabling more reliable analysis and interpretation of the data [72].

Consequently, the development and implementation of optimised preprocessing protocols are essential for achieving analytical precision, methodological reproducibility, and clinical applicability in GBM radiomics analysis. Preprocessing methodologies remain heterogeneous across radiomics research, reflecting the absence of a definitive standardised pipeline. While IBSI [66] works toward protocol standardisation for reproducible engineered radiomics features, significant variation persists in practical applications. Global initiatives exemplified by BraTS have emerged as pivotal drivers of neuro-oncologic research advancement. By offering extensively standardised datasets with a preprocessing pipeline via the Cancer Imaging Phenomics Toolkit (CaPTk) software [73], [74] for tumour segmentation, survival prediction, and radiogenomic investigation, these platforms facilitate systematic methodology assessment and accelerate innovation in clinical neuroimaging analytics.

CaPTk implemented the BraTS preprocessing pipeline, which is particularly valuable, as the BraTS challenge applies the same pipeline structure to its multi-institutional datasets. This alignment is important for reproducibility and helps enhance clinical translation efforts. CaPTk provides detailed information on the BraTS preprocessing pipeline, including intermediate outputs that facilitate comparison with custom pipelines. Its graphical user interface-based workflow requires no additional adjustments, minimising effort for reproducibility. We aimed to optimise the pipeline to meet the specific needs of our local dataset. Furthermore, skull stripping and automated tumour segmentation steps are optional in the preprocessing pipeline, and we explored enhancing generalisability by replacing

tools in these steps with up-to-date alternatives. The skull stripping tool, included as an optional step in CaPTk, is outdated. In Chapter 2, we adopted HD-BET [75], an up-to-date DL model was trained on a large multi-institutional dataset from 37 European centres, which outperforms alternative publicly available tools such as FSL BET [76] and AFNI 3dSkullStrip [77].

Radiomic analysis of multiparametric Magnetic Resonance Imaging (mpMRI) data necessitates various preprocessing procedures, with the CaPTk handling several but not all of these steps:

1. Format Conversion: Digital Imaging and Communications in Medicine (DICOM) images are converted to standard formats like Neuroimaging Informatics Technology Initiative (NIfTI) for easier processing [57]. DICOM is the standard format utilised in clinical settings, integrating image data, comprehensive metadata, and communication protocols for hospital-wide applications. In contrast, NIfTI is the preferred format in neuroimaging research, providing broader data type support and efficient handling of multidimensional data for post-processing and analysis [78].
2. Resampling: Images are resampled to a uniform voxel size (e.g., 1 mm^3) to correct differences in scanner settings and slice thickness [57], [66], [71]
3. Co-registration: Different image sequences from the same patient are aligned to a reference coordinate for accurate multi-modal comparisons [57], [71].
4. Brain Extraction (skull-stripping): Non-brain tissues (e.g., skull) are removed to focus on the brain region and reduce intensity variation [70], [71], [79].
5. Intensity Normalisation: Various normalisation techniques, encompassing Nyul, WhiteStripe, and Z-score normalisation, serve to standardise intensity distributions across datasets [70]. Although these approaches effectively harmonise intensity distributions, the field lacks definitive evidence to support the selection of a single superior methodology [80].
    - Nyul: Matches intensity distributions using a reference histogram.
    - WhiteStripe: Uses Z-score normalisation based on Normal Appearing White Matter (NAWM).
    - Z-Score: Averages intensity values across the whole brain mask.

6. Bias Field Correction: Intensity inhomogeneity within tissues is reduced [81]
7. ROI/ Volumes of interest (VOI): The precision of ROI segmentation represents a cornerstone of radiomics analysis, directly affecting feature extraction and model dependability [69]. Automated delineation methods have gained prominence due to their inherent reproducibility and the reduction of inter-rater variability, a crucial consideration for biomarker development and clinical translation [82].
8. Harmonisation: Harmonisation techniques are essential in multicentre radiomics studies to address the "centre effect," which refers to the variability in RFs arising from differences in scanners, acquisition protocols, and reconstruction settings across institutions [83].
    - ComBat: Originally developed for genomic batch effect correction, ComBat harmonisation has been successfully repurposed for radiomics applications, where it has shown promise in reducing centre-specific variations and enhancing the reproducibility of multi-institutional results [83], [84].
9. Grey-level discretisation (binning): It is essential to calculate textural features by grouping similar intensity levels (bins), simplifying image representation, and reducing noise impact [67]. Two fundamental discretisation strategies are employed:
    - Fixed Bin Number (FBN): This method adaptively modifies bin widths to maintain consistent bin quantities throughout the intensity range, proving especially advantageous for MRI data by compensating for contrast variations and augmenting feature reproducibility [66].
    - Fixed Bin Size (FBS): This approach implements uniform bin width across the intensity spectrum, demonstrating efficacy in CT and PET imaging [85] but limited applicability in MRI due to its non-standardised intensity metrics.

The choice of discretisation method (FBS or FBN) and the number of bins significantly impact the generalisability and accuracy of radiomics models across

diverse datasets. However, the potential benefits of increased bin quantities for feature consistency must be weighed against the potential risks of decreased classification accuracy in certain diagnostic scenarios, emphasising the need for a nuanced understanding of the relationships between bin quantity, feature consistency, and classification performance in radiomics model development [70]. Reporting preprocessing steps, including tools, parameters, and public code releases, is critical for reproducibility. IBSI provides guidelines, recommending relative discretisation for MRI to handle variable intensity ranges.

1.9.2 Segmentation

Image segmentation, the delineation of ROI in 2D or VOI in 3D, is the crucial first step in any radiomics pipeline, defining the area for feature calculation. It is also the most critical, challenging, and debated step of radiomic analysis [61]. Segmentation methods range from manual delineation to semi-automatic techniques using algorithms like region-growing or thresholding, and finally to fully automated approaches, often employing DL [67].

Manual and semi-automated segmentation, frequently corrected by physicians, are widely used but possess inherent limitations. Manual segmentation is notably time-consuming, particularly with large datasets. Both approaches introduce observer bias, weakening the robustness and generalisability of RFs due to intra- and inter-observer variability in ROI/VOI delineation [82]. Therefore, studies employing these methods should thoroughly assess this variability by excluding non-reproducible features, utilising thorough evaluations of intra- and inter-observer bias [67].

Initially, ML approaches utilising hand-crafted features were employed for brain tumour segmentation in BraTS [86]. However, DL has demonstrated superior performance due to its ability to extract complex features [87]. Fully automated image segmentation in medical imaging uses DL models like U-Net [88] and others, such as atlas-based methods [89], to achieve accurate and reproducible results. Many open-source algorithms have been developed for segmenting various organs, mainly focusing on entire organs rather than specific tumour regions. Automated segmentation is beneficial as it can enhance the consistency and generalisability of

RFs, thereby overcoming variability issues caused by differing observations. However, the generalisability of these algorithms remains challenging across multi-institutional datasets. External validation with such datasets is ongoing, emphasising the need for further research to create more robust and generalisable segmentation algorithms [61]. This study examines the performance of automated segmentation models driven by DL algorithms, with particular emphasis on their generalisability when tested on external validation datasets under clinical settings.

1.9.2.1 Impact of Domain-Specific Preprocessing and ROI Characteristics

The variability in MRI intensity arises from factors such as differences in scanner models, manufacturers, and acquisition techniques, posing challenges for the generalisation of ML and DL-based segmentation methods. Consequently, standardisation of MRI intensity is necessary, which might improve the generalisability of ML/DL-based segmentation models. The application of intensity normalisation techniques has shown improvements in the metrics of convolutional neural network (CNN)-based brain tumour segmentation [90]. The primary objective of brain tumour segmentation is to delineate active tumour tissue, including ET, necrotic (NCR) tissue, and oedema (swelling adjacent to the tumour). In radiotherapy, GTV is delineated for treatment planning, defined as "the gross palpable or visible/demonstrable extent and location of the malignant growth" [91]. Duman et al. highlighted the significant similarity between the GTV and TC [92]. Due to the difficulty in differentiating between tumour and healthy tissues, given their overlapping imaging characteristics, multiple MRI modalities such as T1, T1ce, T2, and FLAIR are commonly employed. The Response Assessment in Neuro-Oncology (RANO) working group provides guidelines regarding specific MRI modalities for GBM [93]. In clinical settings, the diversity of MRI modalities in brain tumour segmentation introduces challenges, including varying sequences, image quality, resolution, and slice thickness across different modalities, as well as the complexity involved in integrating these diverse data sources for precise tumour delineation.

1.9.2.2 A general overview of automated medical image segmentation

Current medical image segmentation methods can be grouped by the type and availability of annotations utilised when training a model. As the most traditional method, fully supervised approaches require a large amount of precisely labelled data at the pixel level, yielding outstanding performance benchmarks [94], [95] while simultaneously demanding a considerable manual annotation burden [96]. Recent studies have increasingly explored non-fully supervised paradigms [97], such as semi-supervised [98] and weakly supervised [99] learning techniques to reduce dependence on large-scale labelled data [96]. By utilising a small quantity of labelled data and a large amount of unlabelled data, semi-supervised techniques aim to achieve a competitive performance and less labelling effort [96]. Weakly supervised methods aim to reduce annotation complexity with lower-effort annotations such as image-level labels, point annotations, bounding boxes and scribbles [97]. It needs models to learn detailed boundaries from sparse and limited forms of supervision. On the other hand, unsupervised segmentation eliminates dependence on annotated data, detecting patterns or representations from images via approaches [100]. While these methods effectively reduce the dependency on labelling workload, they often achieve lower accuracy compared to more supervised techniques [97].

Traditional brain tumour segmentation methods, such as thresholding [101], regional growth [102], active contour [103] and feature-based machine learning, can segment tumours with a moderate performance. However, the big challenge is complex tumour shapes and unclear boundaries due to hand-crafted features, which are time-consuming, need expert input, and have low generalisability. Also, tumour heterogeneity and variations across imaging devices limit the model performance [104]. As a result, researchers are increasingly exploring more advanced methods to address these challenges. DL-based approaches have recently advanced the accuracy of brain tumour segmentation. Research has focused on improving network architectures and model robustness. Although CNN-based models are good at capturing local correlations [105], their ability to model long-range dependencies

and capture global context is still limited [104]. CNN-based U-Net and its variants have dominated developments until around 2020, leveraging skip connections to maintain high-resolution features [104]. U-Net has an encoder-decoder setup connected by skip-connections: the encoder extracts features at various scales, and the decoder utilises these connections to generate the segmented image. U-Net is widely known for its effectiveness in complex medical image segmentation across different tasks [88], [106]. Additionally, the challenge of 3D medical image analysis is solved by replacing all 2D convolutional layers in U-Net with 3D convolutional layers, consequently introducing the 3D U-Net [107]. Therefore, 3D medical images with full spatial context can be used directly for U-Net model training. Another model, V-Net [108], was introduced by utilising shortcut connections from ResNet [109] in 3D medical images. For several segmentation tasks, U-Net and 3D U-Net achieve good performance. However, the models may have issues for different datasets and different domain-specific segmentation tasks. The models need parameter changes for optimum performance [105]. Following its introduction, thousands of studies have cited the U-Net, with many architectural changes and extensions for improving segmentation accuracy and model adaptability. Accordingly, Isensee et al. [95] noted that properly optimised U-Net architectures establish a good performance benchmark, and outperforming them is still a challenge with alternative models. As a result, the authors proposed nnU-net ("no new net"), which trains a basic U-Net model by automatically adapting by using the dataset features. nnU-net has shown high performance in 49 different segmentation tasks, ranking first in 21 of them [95], and continues to be a highly competitive framework in medical image segmentation [105].

The attention mechanism enhances neural networks by allowing them to focus on important regions or features in input data, improving their ability to capture critical information. It has been integrated into various models to increase performance on tasks involving small or complex structures [105]. The U-Net model with an attention mechanism was introduced by helping to focus more on important areas in the input image, which significantly improves the model's ability to localise abdominal organs [110]. A Cross-Task Guided Attention module was proposed

[111], using information from previous tasks to guide attention. Attention mechanisms can be classified into spatial and channel attention. Spatial attention is widely applied to help models concentrate on important areas by expanding the receptive field [105]. A U-Net variant integrated spatial attention modules, such as SA-UNet [112], proposed attention modules at the interface between the encoder and decoder, allowing the network to focus on more informative spatial regions for retinal segmentation. Unlike spatial attention, channel attention focuses on inter-channel connections within a feature map, enhancing task-relevant channels by changing their weights for better feature representations [105]. Channel attention only considers the channel-wise information; it may miss important spatial relationships between features, which can reduce its performance in some medical image tasks. To handle this issue, CPCANet [113] was proposed by combining Channel Prior Convolutional Attention module with a spatial attention module to balance attention weights. The model achieved good performance in cardiac diagnosis and skin lesion segmentation. A dense connection is a densely connected structure in deep learning where each layer's output is passed to all later layers as input, creating rich forward and backward connections [114]. This dense structure enhances information propagation and introduces a solution for gradient vanishing. A densely connected 3D model [115] was introduced by adding dense connections to traditional CNNs. Despite its increased GPU memory consumption, this model helped the model preserve information between layers and achieved higher performance than that of 3D U-Net. Li et al. [116] introduced a dense U-Net model with a lower GPU requirement, utilising a hybrid architecture that combines a 2D Dense U-Net with a hierarchical fusion of 3D contextual information, tailored for liver tumour segmentation.

Multi-scale methods in image processing work by analysing the input at different scales, allowing models to understand both local details and broader contextual information [105]. Unlike the standard U-Net, which utilises a single scale, these methods give a broader and more effective feature representation. UNet++ improves upon the original U-Net by introducing sub-networks with multi-scale approach by replacing long skip connections across diverse medical image

segmentation tasks [117]. However, Huang et al. [118] defined that UNet++'s dense skip connections limit the sufficient capturing of the full-scale. Therefore, UNet 3+ [118] addresses this problem with full-scale skip connections utilising low-level features with semantic information in multiscale approach. The previous structures are known as inter-layer multi-scale methods due to collecting features from different encoder levels and combining them in the decoder. Another type, known as intra-layer multi-scale methods, captures features at different scales, such as Atrous Spatial Pyramid Pooling module [119] and a densely connected version of this module [120]. MSNet [121] was introduced for polyp segmentation, which uses inter-layer multi-scale method and aims to reduce redundant information caused by combining these features. Additionally, the intra-layer version of this unit was added to the structure, leading to M^2SNet model [122].

Since 2020, Transformer-based architectures like Vision Transformer have gained popularity for capturing long-range dependencies, showing superior segmentation performance compared to CNNs [104]. The Transformer [123], initially introduced for sequence data such as language sequences, has been successfully extended to image processing through the Vision Transformer. It operates by dividing images into patches and utilising self-attention mechanisms to capture global contextual relationships. In medical image segmentation, Transformer-based models can be classified into two groups: pure Transformer methods and hybrid methods that utilise the strengths of Transformers and other methods, such as CNNs [105]. Pure Transformer models are still not common in medical image segmentation, as most methods continue to rely on convolution layers [105]. To explore convolution-free model use in 3D segmentation, Karimi et al. [124] proposed dividing 3D images into patches, flattening them into 1D embeddings, and applying Transformer blocks with self-attention to capture global information. The model matched or outperformed the state-of-the-art CNNs in the experimental results. To adapt Transformers for vision tasks, Liu et al. [125] proposed the Swin Transformer, which utilises Shifted Window Multi-Head Self-Attention to reduce computational cost. On the other hand, Cao et al. [126] developed Swin-Unet, a Transformer-based model using shifted windows for 2D medical image segmentation. Although the model is similar to U-

Net with skip connections, windowed attention limits to capture of global features while improving the extraction of local features. Recent research has explored combining CNNs and Transformers to improve accuracy on medical image segmentation. These hybrid models are usually grouped as using serial, parallel, or skip connections, taking advantage of both convolution and attention mechanisms [105]. One common hybrid approach is the serial connection of CNNs and Transformers, where image patches converted from CNN feature maps are fed into a Transformer module. TransUNet [127] is a version of this design by using a hybrid CNN-Transformer encoder and a Transformer in the decoder to preserve fine details like organ shapes and boundaries. nnFormer [128] is a hybrid model that interleaves CNN and Transformer blocks. Additionally, it stands out by using volume-based multi-head self-attention to handle effectively 3D medical images. In a parallel connection, CNNs and Transformers process information side by side [105]. TransFuse [129] integrates both CNN and transformer in parallel to combine features from different levels of the encoder. It proposed a new fusion method with a shallow network for better inference speed and efficient model size. Additionally, Yuan et al. [130] developed CTC-Net with two parallel branches: one CNN encoder and one Transformer encoder with a Swin Transformer decoder. They also introduced a feature complementary module to fuse features from two branches, which extracts local features and long-range dependency. In hybrid models using skip connections, Transformers and CNN are linked in the U-shaped structure, connected by skip connections [105]. UNETR [131] is a hybrid model utilising skip connections that utilise a Transformer in the encoder and connect it to a CNN decoder through multiple levels. Although UNETR improves segmentation accuracy, it increases model size. UNETR++ [132], which includes an Efficient Pairwise Attention module, separating spatial and channel attention and sharing weights among attention branches to reduce model size while keeping high-quality segmentation outputs.

CNNs are not effective at capturing global features. Also, Transformers need large datasets, which is an important challenge in the medical domain due to the scarcity of medical image data. Because of these limitations, researchers have started

exploring new architectures to improve segmentation [104]. Mamba, introduced for natural language processing in 2023 [133], is a selective state space model. On the other hand, state space models have also performed well in visual tasks [134]. Recently, more medical image segmentation methods have started using Mamba, showing it could be a promising direction [105]. It addresses the ineffectiveness of CNNs for global contextual information while simultaneously preserving the computational efficiency through linear complexity, in contrast to the quadratic complexity of self-attention mechanisms in Transformers [104]. To use Mamba in computer vision, Li et al. developed VMamba [134], a backbone network. Its main part, the VSS block, applies a 2D-Selective-Scan module that scans images to combine information from several directions. Recent research based on VMamba has mainly focused on improving accuracy with the pure selective state space model [135], reducing computational cost [136], and adjusting the model for different cases such as such as avoiding training from scratch [137]. VM-UNet [135], inspired by V-Mamba, was developed as the first model entirely based on selective state space models for medical image segmentation tasks. The model uses VSS blocks in both encoder and decoder within a novel U-Net-like architecture and adds addition operation instead of concatenate operation in the skip connections. Although it outperformed the state-of-the-art models, the authors noted further improvements, such as decreasing the model size, can enhance the applicability of real-world medical scenarios [135]. Real-world clinical settings have challenges due to limited computational resources. Thus, developing approaches that enhance U-Net's performance in capturing global features, without increasing computational cost, is important. Mamba-based architectures provide a promising direction in this regard [105]. LightM-UNet [136] is a lightweight model that integrates U-Net with Mamba, significantly reducing the model size. Despite its compactness, it surpasses state-of-the-art methods on various medical segmentation tasks. Additionally, Swin-UMamba [137], pretrained on ImageNet, outperformed state-of-the-art models. In addition to VMamba-based models, recent research has explored the direct integration or enhancement of Mamba blocks in novel architectures. U-Mamba [138], utilising a hybrid CNN-Mamba block, is another model that designed for medical image segmentation. U-Mamba, based on nnU-net, can automatically adjust

to different datasets. It performs well in areas like abdominal imaging, endoscopy, and cell segmentation. On the other hand, a recent study [139] showed that the model doesn't always outperform nnU-net, meaning its success might depend on the dataset characteristics [105]. Previous Mamba-based models were not specifically designed for 3D medical segmentation. To develop a Mamba-based model for 3D medical imaging, SegMamba [140] was proposed with a new tri-orientated Mamba module, which improves 3D feature extraction by using three different directions. The model is good at capturing global features while reducing training memory and inference time [140].

For brain tumour segmentation, Isensee et al. [141] used their model as the baseline 3D U-Net with minor modifications in the BRATS 2017 Challenge. Also, A lightweight 2D U-Net with an attention mechanism was proposed, outperforming state-of-the-art models in the same dataset [142]. Although U-Net variants were widely used until 2020 [104], alternative models to U-Net were also proposed, such as a decision tree-based method utilising SegNet for the same dataset [143]**.** Several of the U-Net variants have been proposed that integrate attention mechanism, DenseNet and ResNet architectures, including Res-U-Net [144], Hybrid ResU-Net [145], JGate-AttResU-Net [146], Hybrid DenseNet121-U-Net [147]. The U-Net model, along with variants, has achieved strong performance in brain tumour segmentation by using skip connections [104]. Despite its advantage, these models often have issues in capturing global features, which affects segmenting tumours for complex patterns. Transformer-based approaches with self-attention to represent global contextual information surpassed U-Net on the BRATS 2020 dataset [148]. Wang et al. [149] introduced TransBTS for brain tumour segmentation, utilising a Transformer in a 3D CNN. On the other hand, CoTr [150] was proposed, which performs a sparse-attention Transformer with 3D CNN, aiming more efficient global context representation for 3D segmentation tasks. A recent study [151]**,** covering various CNN-based and Transformer-based models, showed that transformer-based models, including the study's proposed TransUNet architecture, achieved competitive or superior results while notably increasing model size, training time and memory usage compared to nnU-net [95] across various medical datasets.

Additionally, TransUNet achieved a slight improvement over the extended nnU-net [152] on the BraTS 2021 dataset [151].

On the other hand, MambaBTS [153], inspired by the Mamba architecture, proposed a U-Net-based network incorporating a cascade residual multi-scale convolutional strategy for brain tumour segmentation. The model has fewer parameters and superior accuracy over state-of-the-art models on the BraTS 2019 dataset. MUNet [154] combined U-Net with Mamba by introducing an SD-SSM module for both global and local features and an SD-Conv module for minimising feature redundancy. Thus, the model outperformed state-of-the-art models on the BraTS 2020 dataset while including a lower number of parameters relative to most models, excluding U-Net. CDA-mamba [155] was introduced to balance accuracy and efficiency for brain tumour segmentation, which outperforms state-of-the-art models on the BraTS 2023 dataset while providing the lowest inference time. Also, SegMamba ranked second in accuracy in the same study, while yielding low inference time. Additionally, the study was computationally intensive for training, with GPU memory consumption reaching 32 GB [155]. One-dimensional selective scanning mechanism of the original Mamba is limited to the extraction of spatial information in high-dimensional visual data [156]. Although current methods attempt to mitigate this limitation, they are insufficient, highlighting the necessity for further research [156]. Based on this, current Mamba-based methods in 3D medical imaging remain limited in their capacity to represent spatial dependencies in multiple directions and to extract high-resolution spatial features, which is important for accurate segmentation [155].

1.9.2.3 Current challenges in medical image segmentation

Although deep learning has driven significant improvements in medical image segmentation, important challenges remain that compromise methodological robustness, computational efficiency, and clinical applicability. These obstacles originate from both the medical imaging data and the design limitations of current segmentation algorithms [96]**.** Medical imaging data present heterogeneity across modalities (e.g., MRI, CT, ultrasound), acquisition parameters, and institutional

settings. For instance, MRI scans from varying scanners may exhibit discrepancies in resolution, contrast, and artefact characteristics, leading to domain shifts that can significantly degrade model performance [96]. A major challenge is overfitting [96], posing a critical limitation in medical image segmentation, characterised by strong performance on training samples coupled with diminished accuracy on unseen data. This challenge often exists when model complexity surpasses the heterogeneity and scale of the training data due to the limited availability of annotated medical imaging datasets. Data scarcity in deep-learning medical imaging arises not only from image availability but also lack of annotation [157]. Producing pixel-precise labels is time-consuming and expensive [158], particularly for rare diseases [159]. Additionally, privacy and governance frameworks limit cross-institutional sharing [160], [161] while many researchers have no clinical access. Variability across scanners, acquisition settings, and demographics introduces systematic domain shift, while weak or noisy boundaries introduce annotation noise [160]. These limitations hinder the effectiveness of deep learning models and motivate data-efficient alternative methods [159], including data augmentation [162] and non-fully supervised methods [97]. Additionally, the overfitting issue is deepened by the poor generalisation ability of many models, which leads to decreased performance on unseen clinical datasets. This significantly limits their clinical application [104]. Current research on medical image segmentation rarely provides sufficient validation to determine which models have the potential for clinical translation [163]. Although a model may achieve high overall segmentation accuracy, high variability in its performance can make it unsuitable for clinical applications, where patient safety is a critical concern [163]. For example, low performance on even a few cases could lead to serious consequences for patients [163]. Also, robust domain adaptation and transfer learning methods are important to enhance the transition of models to clinical applications [104]. For instance, Sharma et al. [164], Yang et al. [165], and Dai et al. [166] have introduced innovative methods aimed at enhancing cross-domain performance, thereby addressing this issue [104].

DL models with large parameters need significant computational resources and long training times, a challenge that is amplified when models are deployed on large-

scale medical imaging datasets or perform inference on high-resolution 3D medical images. Transformer-based models achieve state-of-the-art performance in tasks like brain tumour segmentation but demand significant GPU memory and extended training times due to their self-attention mechanisms, multi-scale feature fusion methods [96]. On the other hand, CNN-based 3D U-Net [107] achieves high segmentation accuracy in 3D medical images. However, its high computational burden in both training and inference reduces efficiency, especially in clinical settings with limited resources [96]. Lightweight architectures [150], [167] have been designed to enhance computational efficiency. However, these architectural simplifications often impair the capacity to capture detailed anatomical structures and to generalise effectively across heterogeneous datasets [96].

1.9.2.4 CNN-based U-Net variants

Considering the challenge of data scarcity, CNN-based U-Net variants such as nnU-net achieved robust performance under different clinical scenarios. In a comprehensive study using 12 multi-institutional real-world clinical datasets [168], nnU-net maintained high performance for brain tumour segmentation even when MRI sequences were incomplete or varied in quality. Notably, nnU-net demonstrated strong generalisability, maintaining consistent segmentation performance across all datasets, supporting its ability in different clinical settings. Additionally, recent comparative studies using real-world datasets have demonstrated that nnU-net outperformed both Mamba-based and Transformer-based models in various medical image segmentation tasks, including clinically defined contouring formats [169], [170]. On BraTS 2021 dataset, a recent study noted that nnU-net achieved segmentation results on par with state-of-the-art Transformer-based and Mamba-based models for brain tumour segmentation [139]. In addition to addressing challenges of data limitation or generalisability on real-world scenarios, CNN-based U-Net variants are known for cost-effectiveness. For example, nnU-net has proven to be an effective choice, demonstrating high segmentation accuracy alongside shorter training times and lower computational costs when compared to current Transformer-based and Mamba-based models

[171]. While requiring significantly less GPU memory and training time, nnU-net matches the performance of more resource-intensive, state-of-the-art segmentation models [139]. Although the extended nnU-net [152] with retraining achieved slightly lower segmentation accuracy than TransUNet on the BraTS 2021 dataset, nnU-net required nearly half the GPU memory, training time, and inference time compared to TransUNet, confirming the efficiency and competitiveness of U-Net-based models like nnU-net in different scales, such as 2D and 3D [151].

In our study, the U-Net architecture, along with its self-configuring variants, nnU-net [95] and the extended nnU-net [152] were selected as the segmentation models due to their demonstrated efficacy, adaptability, and suitability in real-world scenarios. The ability of nnU-net to autonomously adapt preprocessing, model training, and postprocessing methods [95] makes it particularly well-suited for domain-specific segmentation tasks in resource-limited clinical settings. In addition, our research explored domain-specific adaptations to preprocessing and data handling strategies, aiming to enhance both resource efficiency and segmentation performance relative to nnU-net's default self-configuring pipeline. Our research assessed whether customised, domain-specific modifications, particularly in areas such as focusing on different GBM contouring formats, normalisation methods and rigid registration-related resampling, can improve ensemble segmentation performance. Through a systematic assessment of alternative preprocessing approaches coupled with modified U-Net variants at different scales (2D, 2.5D and 3D), our study aimed to deliver competitive performance with enhanced computational efficiency, while improving model generalisability on our institution-specific dataset.

1.9.2.5 The related work for the proposed strategy

In early applications, computer vision techniques demonstrated success under specific conditions for analogous tasks; however, medical image segmentation is still a challenge due to the complexities of feature representation [172]. Although this challenge persists, DL methods have shown promise in image segmentation tasks. CNNs are renowned for their capacity to learn complex patterns and features.

In order to segment tumours effectively using CNNs, extra feature extraction methods are often implemented [173]. In recent years, transformers have found widespread adoption in computer vision, including image segmentation tasks [123]. These models, utilised either independently or in conjunction with CNNs, effectively capture both local and global information in medical image segmentation. Most studies integrate transformer architectures with the U-Net or similar variations [174]. Given the numerous advancements and diverse methodologies in computer vision, it is important to classify the predominant models essential for navigating the complexities of medical image segmentation tasks.

Contemporary models are generally categorised into two principal types: (1) multi-class segmentation, and (2) cascaded versions of binary class segmentation, which provide all-in-one, end-to-end solutions for each sub-tumour, namely ET, TC, and WT of brain tissue. Multi-class segmentation is effective in delineating multiple tumour classes. Unlike the multi-class segmentation, binary class segmentation may offer distinct advantages, such as simpler optimisation processes [175]. In binary class segmentation, the multi-class problem is subdivided into three separate models, each targeting a specific sub-region for each class. Subsequently, all sub-regions are segmented using either cascaded or simple binary class models [175], [176]. Additionally, 2D models employing binary classification may outperform 3D models using the multi-class segmentation approach [177]. Tumour classes are represented in two ways: labels (non-overlapping masks) and sub-regions (overlapping masks). Labels are categorised as ET, NCR tissue, and oedema, while sub-regions include (i) ET, (ii) TC; encompassing ET and necrosis, and (iii) WT; including ET, necrosis, and oedema [178]. On the other hand, multi-class segmentation primarily focuses on label segmentation rather than region-based techniques. However, previous studies indicate that optimisation based on sub-regions rather than labels yields superior results [179], [180], [181], [182].

It is vital to highlight that the generalisability of DL models poses a significant challenge, particularly in the field of medical imaging research. For instance, a state-of-the-art model was developed using the BraTS dataset but evaluated on a local

dataset, which resulted in a marked discrepancy. The segmentation performance on the local dataset did not match the high accuracy observed with the BraTS dataset [183]. Furthermore, another study investigated the substantial impact of MRI scanner variability on medical image analysis. This study analysed datasets from two different scanners, each containing data from 50 patients, and demonstrated the potential improvements achievable through diverse methodological approaches [184]. The root cause of these scanner-induced differences lies in variations in MRI acquisition parameters, such as slice thickness, matrix size, echo time, and TR. These findings underscore the requirement for more comprehensive research efforts focused on improving segmentation accuracy on local datasets. Such efforts are important in enhancing the universal applicability and reliability of DL models across varied clinical settings. Addressing these challenges will contribute significantly to the advancement of medical imaging technologies. Therefore, we proposed a novel strategy to handle these challenges in Chapter 5. The main contributions of this research include the following:

1. This work constitutes the first thorough investigation within the literature into the application of various normalisation techniques employed on MR, specifically in the context of segmentation tasks for DL models.

2. This study introduces Region-Focused Selection Plus (RFS+), a novel and adaptable framework applicable to a wide range of DL models. By combining various segmentation techniques, normalisation methods with ensemble learning, RFS+ achieves superior accuracy in brain tumour segmentation while enhancing its applicability across heterogeneous datasets.

3. Through methodically investigating the impact of various normalisation methods on U-Net architectures, RFS+ framework attains higher Dice Similarity Coefficient (DSC) metrics across all regions. This paradigm facilitates the discovery of region-specific optimal normalisation strategies when utilising a unified model. For example, the migration of models conditioned under one methodology, particularly multi-class segmentation, to characterise domains such as ET, TC, and WT for GTVs frequently yields suboptimal outcomes. Nevertheless, RFS+ excels at identifying the

most effective model for specific contours, especially when translating knowledge between different contour categories (e.g., from TC to GTV) [92].

4. Via ensemble learning, RFS+ combines the top three models exhibiting superior DSC metrics in training data evaluation, as identified through its algorithmic framework. Implementation within a 2D U-Net infrastructure yields segmentation performance that surpasses the state-of-the-art models, representing a substantial boost in accuracy. Notably, RFS+ demonstrates a quantifiable enhancement in DSC performance, achieving a 1% improvement over its predecessor methodology.

5. This research thoroughly examines a state-of-the-art model, recognised as the BraTS 2021 challenge winner and implemented using its original Docker image, by testing it on a local dataset. Such an evaluation addresses an important gap in the field, where models trained on the BraTS training dataset are predominantly validated and tested on BraTS-specific datasets, with minimal attention given to their performance on local datasets. By presenting the segmentation performance of this model on local datasets, the study highlights its generalisability, offering meaningful insights and establishing a cornerstone for future research and practical developments in this field.

1.9.3 Feature extraction

After image acquisition and data curation, feature extraction is performed to calculate quantitative characteristics from the ROI/VOI. This study employs RFs that are standardised according to IBSI [66] and implemented through the SPAARC Pipeline for Automated Analysis and Radiomics Computing (SPAARC) software package [185], [186].

The following radiomic feature families are used, identified by IBSI:

- Shape-based Features:
  - Morphological
- First-order Features:
  - Intensity-based Statistics

- Intensity Histograms
- Intensity-Volume Histogram

➢ Texture Features:

- Gray Level Co-occurrence Matrix
- Gray Level Run Length Matrix
- Gray Level Size Zone Matrix
- Grey Level Distance Zone Matrix
- Neighbourhood Gray Tone Difference Matrix
- Neighbourhood Grey Level Dependence Matrix

1.9.4 Machine Learning and Deep Learning Models

After extracting RFs, the final step involves developing predictive and prognostic models for clinical applications in cancer research, such as prognosticating patient outcomes, predicting treatment responses, and evaluating tissue malignancy characteristics. Radiomic models often suffer from multicollinearity, with many redundant RFs. In a recent radiomics study, the majority of extracted features showed strong inter-feature correlation across different tumour types [187]**.** Due to the large number of RFs extracted, feature selection becomes an important factor of the radiomics workflow by eliminating redundant and non-informative data [68]. This removes highly correlated and non-discriminative features, preventing the models from overfitting risk, thereby enhancing their robustness and generalisability with only the most relevant and distinctive features contributing to the analysis [61]. Additionally, an effective method is to exclude features that are not robust under perturbations. As an example of GBM survival analysis, this requires examining radiomic feature consistency under slight modifications in ROI delineation, preprocessing protocols, or scanner-related factors [188]. Another challenge is that radiomic models trained on single-centre datasets tend to have a significant performance drop on external multicentre datasets [188], limiting their clinical translation due to low generalisability. Also, these models must be designed to handle incomplete data, addressing issues of data sparsity and scarcity since comprehensive information is not always available for every patient [61], [189].

DL-based models that capture intra-tumour heterogeneity and provide successful differentiation of patient cohorts [190]. While DL-based models offer efficiency by bypassing extensive preprocessing and have shown strong performance in medical imaging, their 'black box' nature limits interpretability and hinders clinical use [191]. Although many interpretability approaches have been developed for DL-based models, the lack of biological grounding in these explanations still exists [192]. Also, DL-based models are effective on large datasets, which are often unavailable for rare diseases [191]. When considering these challenges, DL–based models perform on par with traditional radiomic models, but the improvement is not statistically significant and comes at the cost of low interpretability [193]. Additionally, a recent review reported no consistent evidence supporting a clear advantage of deep radiomics approaches [194]. These challenges might be solved by curating larger, heterogeneous, and publicly accessible datasets [195]. On the other hand, traditional ML-based radiomic models offer higher interpretability but moderate performance compared to DL-based models [192]. Additionally, traditional ML-based radiomic models are well-suited to limited data, while DL-based radiomic models have good performance with large datasets [191].

Clinicians must consider a trade-off between accuracy and interpretability, with the latter, in some cases, prioritised in clinical decision-making [192]. Complex radiomic models with a high number of texture features limit interpretability for researchers and clinicians. Smaller, well-defined feature sets improve interpretability and clinical trust, since each feature can be directly examined through its mathematical definition [187]. The establishment of interpretable radiomic models necessitates systematic feature engineering methods. Current guidelines, particularly those outlined by van Timmeren et al. [67], advocate for dimensional reduction, specifically recommending feature sets comprising 3-10 parameters to mitigate overfitting risks and enhance model transparency. Also, the biological rationale of selected features can foster clinical trust, which is considered good radiomic practice [196].

1.9.4.1 Feature Selection for Interpretable ML-based Radiomic Models

Among traditional feature selection methods in radiomics, LASSO is the most frequently utilised method [197]. Despite their widespread use, such traditional methods, including LASSO, mostly collect unstable feature sets under different preprocessing strategies or Cross-Validation (CV), limiting reproducibility and robustness [198], [199]. These methods consider RFs as an independent variable, without focusing on their intercorrelations [198]. When there is a high correlation among a group of features, LASSO preserves only one variable arbitrarily while excluding the others [200]. Despite this limitation, the selection results in easier-to-interpret models due to a small feature set [201]. Additionally, LASSO cannot capture potential nonlinear dependencies among features due to its inherently linear nature [202], [203]. However, LASSO has computational advantages [204]. Evolutionary algorithms, such as Genetic Algorithms (GA), are good at global search and have recently been utilised for feature selection [205], [206]. On the other hand, SI-based Particle Swarm Optimisation (PSO) exhibits a trade-off between exploration and exploitation [207]: particles initially explore the search space widely and then refine the search around promising regions in the exploitation phase [208]. PSO frequently suffers from two limitations: random initialisation, which delays convergence, and ignoring feature redundancy, resulting in poor model performance [209]. Designing an initialisation strategy based on the feature selection task can enhance PSO performance [206]. For example, a hybrid method employs PSO for global search while using traditional methods to exploit these regions [206]. To increase performance and achieve results competitive with the state of the art, we explored alternative feature selection methods, including PSO and GA, using up to 10 RFs within a similar preprocessing pipeline and the same contouring format, as well as multiple contours from the BraTS challenge. PSO generally performs well across tumour types in survival analysis [208], while GA has shown promising results in the prediction of genetic mutations [210]. Considering the limitations of LASSO for feature selection, it collects unstable feature sets and is limited by its linear nature. Nature-inspired algorithms such as PSO and GA can explore nonlinear relationships among features. While GA is good at global search,

it cannot keep information from the previous iterations [208]. On the other hand, PSO is a good solution for a balanced feature selection method when searching for feature subsets. However, PSO has limitations also, such as the initialisation of feature subset generation and feature redundancy. To overcome these limitations, LASSO was utilised, which is good at feature redundancy and model robustness. Therefore, our study focused on a novel hybrid LASSO-PSO feature selection method to combine these strengths and achieve a competitive result in OS analysis of GBM using traditional ML-based models in Chapter 3.

The investigation leveraged two distinct datasets: BraTS 2020 [211], [212], [213], a publicly accessible, multi-institutional resource, and a single-institution dataset from The Río Hortega University Hospital Glioblastoma Dataset (RHUH-GBM) [214]. Both datasets demonstrate compliance with radiomics quality metrics through their implementation of the standardised BraTS preprocessing protocol and consistent delineation of anatomically defined regions (ET, TC, and WT). The model choice was established upon dual criteria: first, adherence to the principle of model selection as proposed by Meneghetti et al. [215] in radiomic research, which emphasises the relationship between reduced complexity and enhanced interpretability and generalisability; second, observed evidence from a systematic literature review [216] identifying the top three most frequently implemented ML models in radiomics research. Consequently, only these two models, Random Survival Forests (RSF) and regularised Cox Regression (Cox-LASSO), as these models exclusively satisfied both aforementioned criteria. Traditional ML models offer advantages over DL models, particularly in terms of interpretability. While DL approaches have demonstrated significant capabilities, their complex architectural designs and extensive parameterisation often hinder transparency in decision processes [67].

Although Meneghetti et al. [215] relied on traditional feature selection methodologies to extract a limited set of two RFs in conjunction with a single clinical feature, resulting in moderate performance within a distinct cancer type, van Timmeren et al.'s framework [67] capitalised on a more expansive upper threshold, incorporating up to 10 features as described in their guideline. Based on a recent

review on radiomics and GBM OS analysis [216], most time-to-event radiomic studies [217], [218], [219] integrate multiple clinical features and more than 10 RFs. However, the inclusion of numerous clinical features may be constrained by data sparsity, particularly in rare medical domains [189]. Furthermore, the use of more than 10 RFs and RFs that are not aligned with IBSI guidelines [66] can reduce reproducibility and generalisability in multi-institutional studies. Therefore, there is a need for an alternative feature selection method that satisfies these constraints while ensuring interpretability and maintaining performance without degradation. Such a method, leveraging engineered RFs and interpretable ML models, should achieve performance on par with or surpassing that of DL-based models or traditional ML models utilising more than 10 RFs and multiple clinical features. This underscores an important gap in the current literature. This study employed LASSO regression as the baseline feature selection method, a well-established and widely utilised technique in radiomics and other fields. Its utility and interpretability in the context of radiomics have been confirmed by previous research, including recent radiomic studies [215], [216] and various benchmarking analyses [220]. Nature-inspired algorithms such as evolutionary-based GA and SI-based PSO are gaining traction in radiomics research [208], [210]. SI-based algorithms offer effective approaches to feature selection [221] and model building [208], offering the potential to enhance prediction performance in the challenging context of radiomic analysis for patients with GBM. The methodological framework extends the application domain of SI optimisation beyond its implementation in DL contexts, such as SwarmDeepSurv [208]. While SwarmDeepSurv utilised an SI-based feature selection and DL-based modelling across multiple cancer types in the study, the results did not achieve statistical significance for risk stratification in GBM time-to-event analysis. Various radiomic research tasks have leveraged nature-inspired hybrid feature selection methods. For instance, O6-Methylguanine-DNA Methyltransferase (MGMT) status prediction in GBM has been explored using a hybrid nature-inspired method (GA) [210], while treatment response prediction in oropharyngeal cancer [209], and breast tumour classification [222] have utilised hybrid nature-inspired (PSO) approaches. Based on this, we explored a novel hybrid feature selection method, representing the first application of PSO in traditional ML-

based radiomic models for OS analysis in GBM. The proposed feature selection method includes three phases: (i) Correlation analysis was applied to reduce multicollinearity among RFs, ensuring that highly correlated features were minimised. (ii) LASSO with CV not only reduced the number of selected features but also explored a wider range of potential feature subsets, from 2-feature to 9-feature, resulting in a total of eight enriched feature pools. (iii) PSO with a balanced search prioritised local search while still maintaining exploration of the broader feature space across each feature pool.

In Chapter 4, our study aimed to enhance clinical translation efforts by utilising our local dataset (STORM_GLIO), clinically defined ROIs, and clinically driven modifications to the preprocessing pipeline applied to this real-world data. Compared to Chapter 3, the survival analysis in Chapter 4 was more challenging as it included STORM_GLIO alongside the BraTS 2020 dataset. Unlike the BraTS 2020 dataset, STORM_GLIO was curated according to the clinical requirements outlined in Chapter 2. To address performance drops on unseen datasets due to differences in preprocessing pipelines and contouring formats, we aimed to develop a reproducible and interpretable radiomic model using a small set of robust features, achieving reliable yet moderate performance. A recent review suggests employing traditional feature selection methods, such as LASSO or Minimal Redundancy Maximum Relevance (mRMR), in radiomic research to achieve high performance with limited complexity [191]**.** This will result in radiomic models that are easier to interpret. Although traditional feature selection methods are widely used for their advantages in redundancy reduction and enhancing model robustness, they often collect unstable feature sets under different preprocessing pipelines, which limit reproducibility [198]. To address the reproducibility issue introduced by traditional feature selection and the heterogeneity of the datasets (different contouring formats and preprocessing pipelines), we performed a robustness analysis. This analysis resulted in the development of a radiomic model that achieved moderate performance when applied to a different cancer type [215].

Additionally, the BraTS 2020 dataset was utilised in both Chapter 3 and Chapter 4, since it originates from 19 institutions with diverse clinical protocols and scanner parameters, making it a highly heterogeneous resource. Therefore, this dataset is highly suitable for addressing the generalisability issue of single-centre radiomic analyses [188]. On the other hand, the BraTS data include only two clinical variables in addition to OS. However, age was the only clinical variable consistently available for all patients across all datasets, along with all four MRI sequences, in Chapters 3 and 4. This limitation could be addressed in the future by incorporating additional clinical variables across multi-institutional datasets. While this represents a limitation of the study, it also had the advantage of reducing model complexity, thereby facilitating interpretability.

To ensure interpretability and applicability within limited clinical settings, such as using a minimal number of clinically based contours and MRI sequences, we selected ML models over DL models due to the latter's "black box" nature. ML techniques facilitate more transparent decision-making processes, which are essential for developing survival analysis models in GBM. Ultimately, our goal is to establish reliable and interpretable tools that support and enhance clinical decision-making.

2. Data Curation for multiparametric MRI Glioblastoma Multiforme data

2.1 Introduction

The implementation of radiomics preprocessing protocols necessitates careful consideration of institution-specific imaging parameters and local data characteristics. Although standardisation remains crucial for broader applicability, the efficacy of preprocessing methodologies can be significantly influenced by local imaging conditions and protocols [70]. A systematic framework for preprocessing our local dataset STORM_GLIO was developed and is described herein, incorporating purpose-built adaptations that address the complexities of translating radiomics analysis into clinical practice while maintaining concordance with standardised medical procedures. An analytical assessment of preprocessing methodologies highlights the importance of tailored modifications, whereby this chapter systematically investigates their integration into the established BraTS pipeline, providing qualitative comparative evidence to substantiate these adaptations' efficacy in addressing the distinct parameters and limitations inherent to our institutional dataset. By introducing targeted workflow enhancements, this study aims to mitigate the complexities and inconsistencies inherent to multi-sequence MRI data while also optimising the performance and robustness of radiomic model-based analysis, thereby improving the overall reliability and generalisability of the developed radiomic models. A thorough investigation of fundamental preprocessing steps is performed, encompassing the conversion of DICOM to NIfTI, rigid registration protocols, brain tissue extraction, application of noise reduction algorithms, implementation of intensity standardisation techniques, delineation of tumour tissues, to improve the reproducibility and clinical utility of the radiomic-based model results. Essentially, the evaluation criteria are grounded in a dual consideration of technical performance metrics and clinical applicability, allowing for a nuanced assessment of the developed approach and ensuring that it is optimised for real-world clinical deployment, where technical capabilities must be balanced against practical clinical constraints. The methodological framework established herein serves as a cornerstone for

subsequent thesis chapters, wherein sophisticated radiomics analytical approaches and automated segmentation investigations are extensively explored and validated.

2.2 Material and Methods

2.2.1 Datasets

In this study, four different GBM collections were utilised.

BraTS 2020: BraTS 2020 dataset [211], [212], [213] is a unique and valuable resource for the medical imaging research community, offering a standardised collection of pre-operative MRI scans, as NIfTI files, with annotations that precisely delineate three tumour sub-regions (ET, TC and WT), and facilitating the development and evaluation of automated segmentation algorithms and survival analysis. Table 2.1 shows BraTS 2012~2021.

Table 2.1 BraTS datasets with three tasks: segmentation, disease progression, survival prediction and MGMT classification [178].

Year	Total Data	Training Data	Validation Data	Testing Data	Tasks	Timepoint
2012	50	35	NA	15	Segmentation	Pre-operative
2013	60	35	NA	25	Segmentation	Pre-operative
2014	238	200	NA	38	Segmentation	Longitudinal
2015	253	200	NA	59	Segmentation, Disease Progression	Longitudinal
2016	391	200	NA	191	Segmentation, Disease Progression	Longitudinal
2017	477	285	46	146	Segmentation, Survival prediction	Pre-operative
2018	542	285	66	191	Segmentation, Survival prediction	Pre-operative
2019	626	335	125	166	Segmentation, Survival prediction	Pre-operative
2020	660	369	125	166	Segmentation, Survival prediction	Pre-operative
2021	2040	1251	219	570	Segmentation, MGMT classification	Pre-operative

In order to facilitate research in both automated tumour segmentation and survival prediction, BraTS 2020 dataset, has been carefully curated to include 660 cases, of which 236 GBM cases are accompanied by survival analysis data, and leveraging the expert manual labels established in BraTS'12-'13 and maintained in BraTS'17-'20 [212]. By incorporating four standardised MRI sequences per case, the dataset enables a detailed examination of brain tumour characteristics, featuring T1 images for anatomical reference, T1ce images for assessing gadolinium uptake, T2 images for visualising pathological features, and FLAIR images for delineating oedema, thereby supporting the development of more accurate and reliable diagnostic and therapeutic strategies. The dataset was pre-processed using a standardised pipeline [212], which included three key steps: co-registration to align all images to the same anatomical template, resampling to achieve a uniform 1 mm^3 isotropic resolution within a 240 × 240 × 155 matrix, and skull-stripping to remove non-brain tissues, thereby ensuring data consistency and quality. In order to ensure accurate and reliable tumour segmentation, the dataset features expert-validated labels for three tumour sub-regions: ET: label 4, peritumoral oedema (ED: label 2), and necrotic/non-enhancing tumour core (NCR/NET: label 1), which are used to define clinically relevant tumour regions, including WT (label 1,2 and 4), TC (label 1 and 4), and ET (label 4), with a robust quality assurance process involving multiple raters, standardised labelling protocols, and neuroradiologist verification in Figure 2.1.

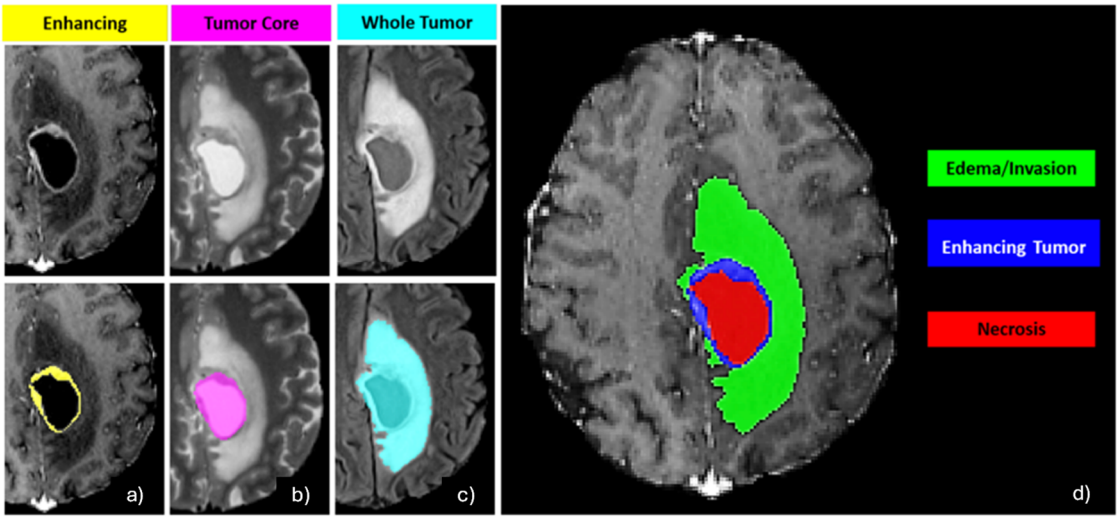


Figure 2.1 GBM sub-regions a)ET, b)TC, c)WT and sub-tumours d) ED, ET, NCR [178].

With the inclusion of three key clinical parameters (OS, Age, and Resection Status), the dataset offers a unique opportunity for researchers to investigate the relationships between these parameters and patient outcomes, enabling survival prediction tasks.

BraTS 2021: BraTS 2021 challenge dataset [211], [212], [213], comprising 1251 training cases, 219 validation cases, and 570 testing cases, is a unique and valuable resource for the medical imaging research community, offering a standardised collection of pre-operative MRI scans, as NIfTI files, with annotations that precisely delineate three tumour sub-regions (ET, TC and WT), and facilitating the development and evaluation of automated segmentation algorithms and the classification of genetic mutation, MGMT, status. This comprehensive dataset incorporates four different MRI sequences: T1, T1ce, T2, and FLAIR. The dataset was established to facilitate the advancement and validation for tumour segmentation tasks and radiogenomics Table 2.1. In order to facilitate accurate tumour segmentation and radiogenomics analysis, the image pre-processing pipeline consisted of three key stages: spatial co-registration to standardise a comprehensive MRI-based reference of normal adult human brain anatomy (SRI24) [223], skull-stripping to isolate brain tissue, and volumetric standardisation to 1 mm^3 isotropic resolution within a 240 × 240 × 155 matrix, which enabled the creation of high-

quality images for expert neuroradiological assessment and segmentation. With a standardised annotation protocol, experienced neuroradiologists conducted manual tumour segmentation to identify and label three primary tumour compartments: NCR as label 1, ED as label 2, and ET as label 4, and subsequently calculated two derivative tumour metrics: TC, which combined NCR and ET regions, and WT, representing the entire tumour region in Figure 2.1.

STORM_GLIO: The STORM_GLIO dataset, a clinical repository of GBM cases collected in Wales between 2014 and 2018, was utilised to validate our methodology, with a focus on a subset of 53 cases that provided complete imaging profiles, featuring a comprehensive MRI protocol including T1, T1ce, T2, and FLAIR sequences. In contrast to BraTS, the STORM_GLIO dataset is characterised by several key methodological distinctions, including the implementation of DICOM formatting and the preservation of acquisition-specific image resolutions and matrix dimensions. Additionally, a manual annotation approach was employed, utilising Clinical Target Volume (CTV) and GTV, where GTV represented the delineations of the observable tumour extent [91]. The preprocessing workflow integrated two essential tools: a) CaPTk to handle image registration, b) HD-BET to handle skull-stripping [75], ensuring BraTS-compliant standardisation. The preprocessing workflow integrated two essential tools: a) CaPTk to handle image registration, b) HD-BET to handle skull-stripping [75], ensuring BraTS-compliant standardisation. This preprocessing pipeline, compatible with the BraTS protocols and designed for seamless integration into clinical use, was complemented by clinical information, including patient age and survival outcomes. Notably, the technical implementations of BraTS and STORM_GLIO differ in terms of file formatting, with BraTS utilising NIfTI and STORM_GLIO employing DICOM. In addition, the two datasets vary in image standardisation approaches, as BraTS maintains uniform parameters (1 mm^3 isotropic resolution within a 240 × 240 × 155 matrix), whereas STORM_GLIO preserves acquisition-specific, variable parameters. Furthermore, in contrast to the BraTS dataset, which utilises a detailed sub-regional classification system (ET, TC and WT), the STORM_GLIO dataset employs a consolidated GTV and CTV approach to tumour annotation, highlighting the diversity of methodological approaches to

tumour segmentation and analysis in brain tumour research (can be seen in Figure 2.2)

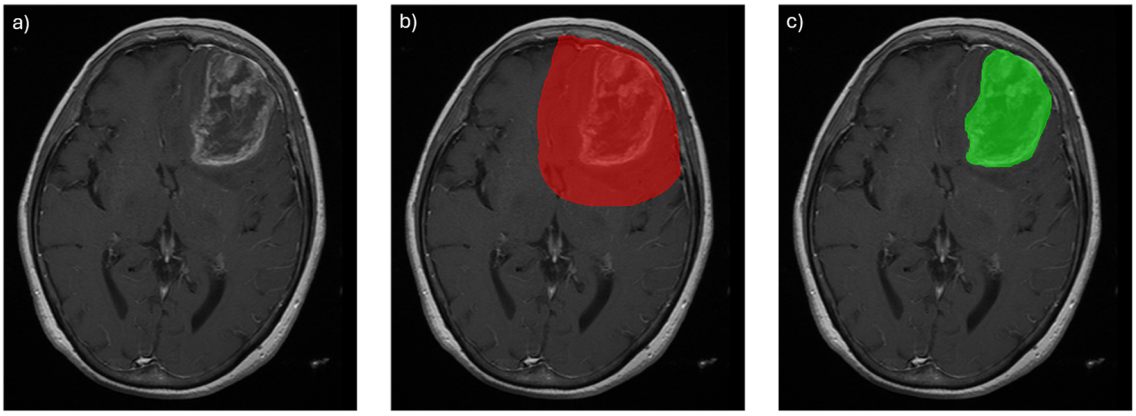


Figure 2.2 A patient from STORM_GLIO with clinical contours , a) T1ce MRI sequence, b) CTV, c) GTV.

In light of these distinctions, it is essential to carefully evaluate dataset selection, model development strategies, and the interpretation of cross-dataset analyses, as the differences between the BraTS and STORM_GLIO datasets can significantly influence the validity and reliability of research findings in brain tumour segmentation and radiomic-based survival analysis.

UPenn-GBM: The University of Pennsylvania glioblastoma dataset (UPenn-GBM) repository represents the most comprehensive open-access dataset for de novo glioblastoma cases currently available to researchers [224]. The collection pairs advanced mpMRI with rich patient metadata, along with clinical, demographic, and molecular information. Comprising 611 cases, all imaging data originated from standardised pre-operative assessments at the University of Pennsylvania Health System (UPHS), with additional follow-up imaging available for selected patients prior to their second surgical intervention. The repository contains 671 total scans (611 pre-operative, 60 follow-up) from 630 patients. Patient demographics show an age distribution of 18-89 years, with males comprising 60% of the cohort. The imaging protocol encompasses multiple MRI sequences: T1, T1ce, T2, FLAIR, and for the majority of cases, Diffusion tensor imaging, and dynamic susceptibility contrast acquisitions. The dataset also consists of expert-curated annotations defining key

tumour components (ET, NCR and ED), initially computer-generated and subsequently manually verified. Data preparation for the UPenn-GBM cohort followed a systematic pre-processing approach. Each imaging study underwent file format transformation from DICOM to NIfTI, followed by reorientation of all mpMRI volumes to conform to the left-posterior-superior (LPS) coordinate framework. The registration pipeline utilised the T1ce scan as the reference, which was initially aligned to the SRI24 atlas and resampled to uniform 1 mm^3 voxels. Subsequent rigid registration brought all other MRI sequences into spatial alignment with this standardised T1ce image. The N4 bias field correction technique was implemented across all MRI modalities to normalise signal intensities [81]. The 'mri_deface' algorithm was applied to remove facial features from all aligned scans, followed by transformation of the defacing masks to match the original image coordinates. The BrainMaGe tool performed skull-stripping operations to extract brain tissue, maintaining effectiveness across all imaging sequences. The scans before and after skull-stripping were available for comparison. Multiple stages of the preprocessing pipeline incorporated expert manual verification and adjustment to maintain optimal data quality. Researchers can access the entire dataset through the Cancer Imaging Archive (TCIA) platform. The fifty patients from the dataset were utilised for skull-stripping as it provided both pre- and post-skull-stripping scans, allowing it to serve as a reference.

2.2.2 Implementation details

Preprocessing was performed utilising a PC running the Windows operating system, equipped with 32 GB of system RAM and an Intel i7-11700 processor. For each step of the BraTS preprocessing pipeline, the CaPTk 1.9 software was utilised [73], [74]. HD-BET was employed as an alternative method for skull-stripping. To ensure an independent and clinically adaptable pipeline, custom code was developed utilising Python libraries. Preprocessing and analysis were performed using the following tools and libraries in Python v3.9.19. The utilised libraries were detailed in Appendix A. The clinical relevance of the pipeline results was confirmed through clinician approval in a clinical setting. The complete, Python-based alternative workflow is available at [https://github.com/krmdmn/preprocesing_pipeline].

For quantitative comparison, DSC was utilised. The DSC was calculated to evaluate the overlap between segmented regions, specifically for brain tissue and tumour segmentation. This metric is commonly used to assess the performance of medical image segmentation tasks. It assesses the similarity between the ground truth segmentation and the predicted segmentation. The DSC metric ranges from 0 to 1, where a score of 1 indicates perfect overlap, and a score of 0 indicates no overlap.

$$DSC = \frac{2|Y_{true,pos} \cap Y_{pred,pos}|}{|Y_{true,pos}| + |Y_{pred,pos}|} \quad (2.1)$$

Here:

- Y_(true,pos) represents the set of positive voxels (or pixels) in the ground truth segmentation.
- Y_(pred,pos) represents the set of positive voxels (or pixels) in the predicted segmentation.
- |Y_(true,pos) ∩ Y_(pred,pos)| denotes the intersection of the two sets, i.e., the number of correctly predicted positive voxels (true positives).
- |Y_(true,pos)| is the total number of positive voxels in the ground truth.
- |Y_(pred,pos)| is the total number of positive voxels in the prediction.

2.2.3 Study Design

In our study, all GBM datasets, except STORM_GLIO, underwent standardisation through the BraTS preprocessing pipeline, which can be accessed and replicated using CaPTk software. The standardised preprocessing workflow (shown in Figure 2.3) consisted of the following steps: (a) DICOM to NIfTI conversion, (b) N4 bias field correction (temporary: due to prevent possible information loss), (c) reorientation of all scans to LPS coordinate system, (d) registration of all scans to the T1ce sequence, and (e) a two-phase rigid registration was implemented: T1ce was first registered to SRI24, after which all other sequences were aligned to the same atlas space. (f) skull-stripping: Removal of non-brain tissue.

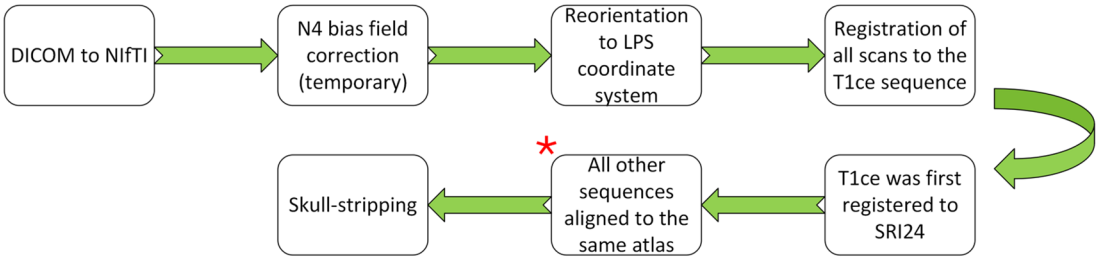


Figure 2.3 The BraTS preprocessing workflow is integrated in CaPTk [73]. N4 bias field correction is applied for optimal registration. However, this is not applied in the final co-registered output images, which indicated with a red asterisk.

The workflow of the STORM_GLIO dataset includes the following steps: (a) DICOM to NIfTI conversion, (b) registration of all scans to the T1ce sequence, and (c) skull-stripping: Removal of non-brain tissue (d) NIfTI to DICOM conversion (for automated contours). The dataset initially acquired in DICOM format, originated from radiotherapy planning cases. The radiotherapy planning process involved delineating both GTV and CTV on T1ce sequences that were co-registered with CT scans. Registration to the SRI24 was not performed in this study. The manual contours on T1ce scans represented critical clinical decisions, so we chose to preserve their original form by avoiding this registration step that might introduce unwanted deformations. When we attempted to reverse-transform the GTV contour or automated segmentations (generated from BraTS-pre-processed STORM_GLIO scans) back to their original DICOM space, we encountered problematic deformation artefacts (can be seen in Figure 2.4,b).

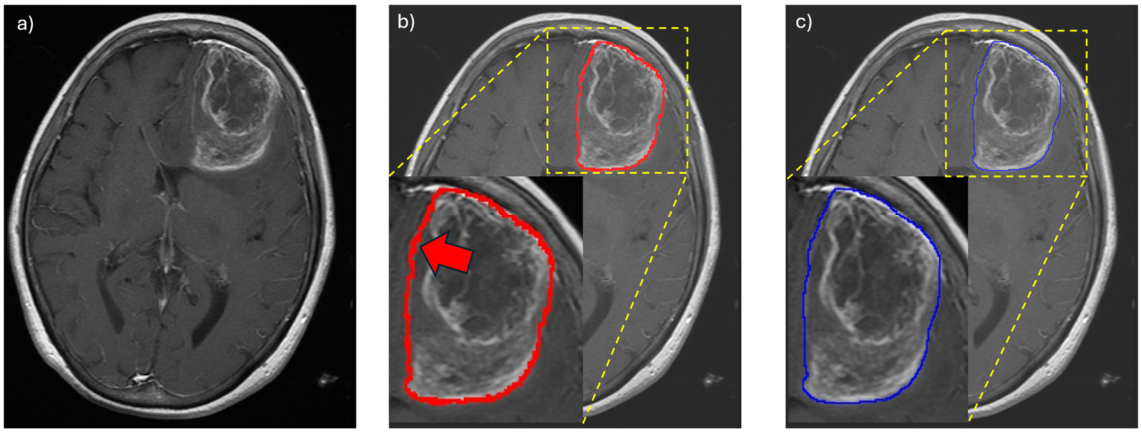


Figure 2.4 Deformation was observed on the contour border (red arrow) during the re-registration of the GTV contour for a STORM_GLIO patient after applying the standardised preprocessing pipeline: a) T1ce MRI sequence b) deformation of the re-registered GTV contour c) GTV original contour

Interpolation artefacts in the resampling step of the BraTS pipeline were observed across different patient cases within the STORM_GLIO dataset (shown in Figure 2.5). A modified version of the BraTS preprocessing pipeline tailored to our specific institutional needs, was developed to overcome these limitations and support quantitative medical imaging analyses (radiomics analyses and automated brain tissue extraction and automated tumour segmentation). Preprocessing procedures were performed within the native DICOM coordinate system, thereby preserving the spatial integrity of the data throughout the entire pipeline. CaPTk, which follows the BraTS preprocessing pipeline, was employed to collect the intermediate and final outputs from preprocessing pipelines. The primary modification implemented in our preprocessing pipeline, distinguishing it from the standard pipeline, was the removal of the spatial alignment step involving registration to the SRI24 atlas.

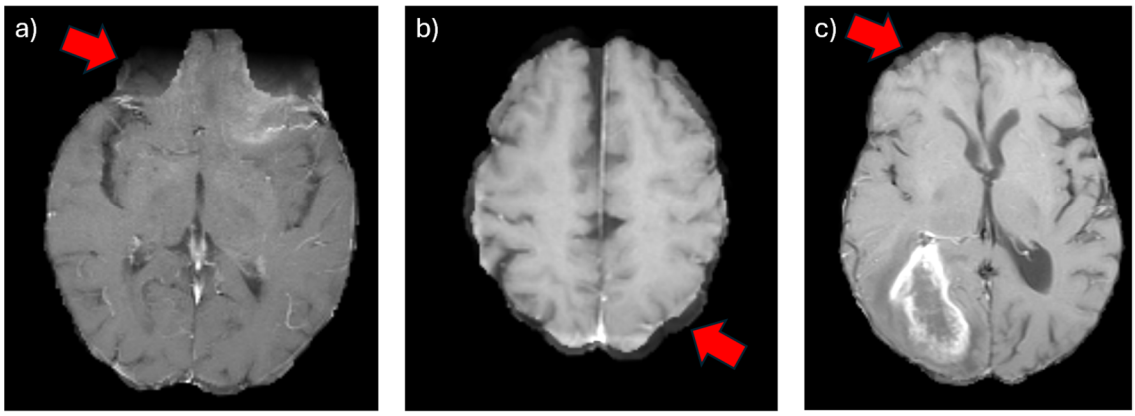


Figure 2.5 The interpolation artefacts were demonstrated with red arrows in a), b) and c) for different STORM_GLIO patients.

A comparative analysis using the DSC metric was performed on the UPenn-GBM dataset (providing both raw and skull-stripped reference scans) to evaluate the effectiveness of CaPTk's DeepMedic-based brain extraction tool [225] against the state-of-the-art HD-BET method for skull-stripping. The DeepMedic model is designed for multi-modal input, requiring four MRI sequences for processing. Unlike DeepMedic, HD-BET processes a single MRI sequence, typically T1, and demonstrates exceptional segmentation accuracy with this sequence [75]. A quantitative comparison of the tumour segmentation results generated by the DeepMedic model [72] (integrated within the CaPTk implementation of the BraTS pipeline) was performed using the DSC metric. These segmentations were derived from the STORM_GLIO dataset, which was pre-processed using both the standard and modified pipeline configurations.

2.2.3.1 BraTS Pipeline vs. Proposed Clinical Workflow

The CaPTk implementation of the BraTS preprocessing pipeline was evaluated and modified to meet the requirements of our clinical workflow. Following a review of the standard BraTS preprocessing workflow, we presented our clinically adapted pipeline and its associated optimisations.

In the BraTS workflow, initial MRI preprocessing involves LPS coordinate reorientation, preparing volumes for the subsequent rigid registration step using the SRI24 neuroanatomical template. To achieve preliminary intensity

normalisation and noise reduction, a temporary N4 bias field correction is applied to the imaging data. All sequences (T1, T2, and FLAIR) are initially registered to the T1ce volume, which is subsequently aligned with the SRI24 atlas template. Using the computed transformation matrix from this registration process, other MRI sequences are spatially normalised to match the atlas coordinate space. The complete set of four MRI sequences is then input into the DeepMedic AI model from CaPTk to facilitate automated skull-stripping of the brain volumes. Using automated segmentation method from CaPtk's DeepMedic AI model, the pipeline generates BraTS labels by identifying three distinct regions: ET, TC, and WT. The full pipeline is demonstrated in Figure 2.6.

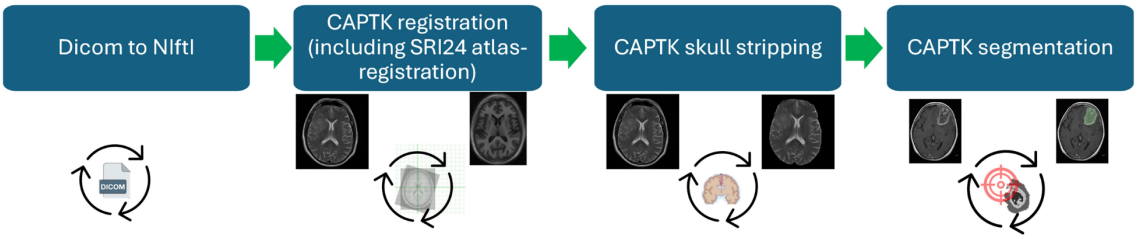


Figure 2.6 The BraTS standardised preprocessing pipeline.

The clinical implementation framework incorporates preprocessing algorithms by leveraging intermediate computational outputs generated by CaPTk. This approach enhances the quality of outputs while ensuring seamless integration into clinical practice workflows. In contrast to traditional atlas-dependent such as SRI24, Colin27, MNI152, ICBM452, and LPBA40, registration frameworks, our modified approach implements reorientation-based image processing to facilitate direct spatial alignment of multimodal sequences (T1, T2, FLAIR) with the T1ce reference volume. By implementing this approach, we maintain the integrity of the original spatial matrices of the T1ce scan and its aligned sequences, effectively avoiding potential distortions of the clinical contours associated with atlas registration Figure 2.4. Our modified framework replaced the DeepMedic-based implementation in CaPTk by adopting the HD-BET model. This approach offers superior performance metrics and a unique capability to perform skull-stripping using a single-sequence input, with optimal results achieved when utilising T1 sequences.

After the completion of the skull-stripping step, the DeepMedic algorithm, implemented as part of the CaPTk software, was conducted to generate automated tumour segmentations. For the final step, the generated contour was converted to Radiotherapy Structure format (RTSTRUCT) for compatibility with clinical systems (Figure 2.7).

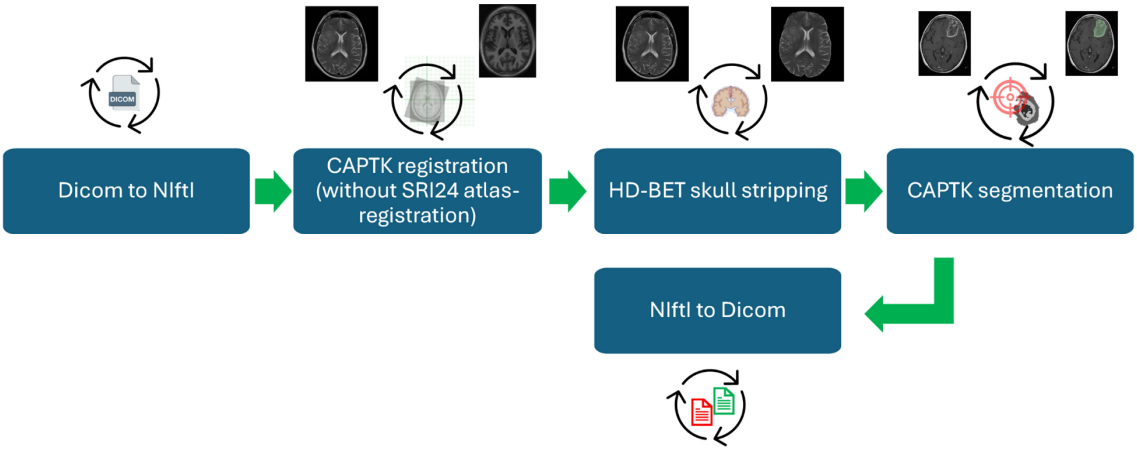


Figure 2.7 The proposed workflow aligned with the clinical settings.

The experimental study was designed to facilitate a systematic comparative analysis between the conventional BraTS preprocessing pipeline and our clinically adaptable preprocessing pipeline. The proposed methodology underwent clinical validation via a comprehensive evaluation, encompassing quantitative metrics and qualitative assessments across preprocessing pipelines.

By eliminating the need for atlas registration and utilising optimised computational tools, particularly HD-BET, the proposed framework aimed to enhance preprocessing efficiency by reducing undesirable deformations of the local scans (due to having lower quality MRI scans compared to the open-access datasets) and the clinical contours. This results in an adaptable pipeline with automated brain tumour segmentation, designed for radiomics applications in clinical settings.

2.3 Results

Automated brain tissue extraction was accomplished by employing the CaPTk platform and the HD-BET computational architecture. The outputs of this procedure were illustrated in Figure 2.8 for STORM_GLIO dataset.

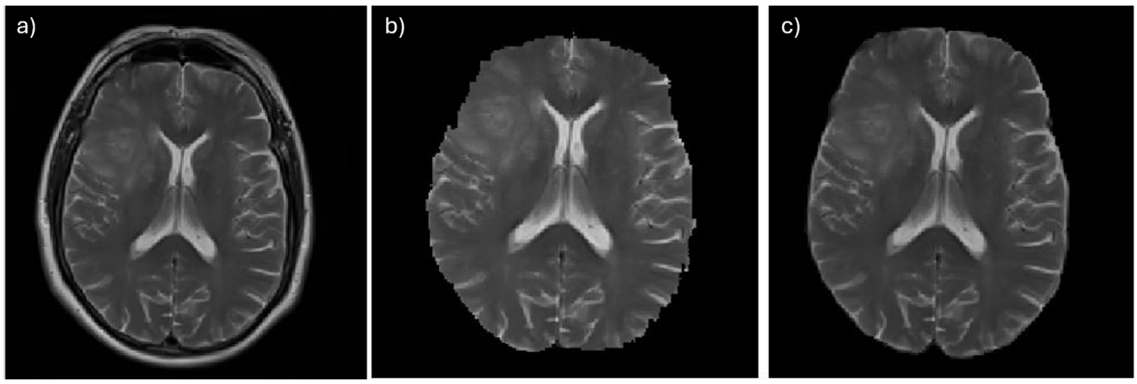


Figure 2.8 Visual Representation of Skull Stripping Step: a) The MRI Scan for STORM_GLIO b) CaPTk result c) HD-BET result: HD-BET provided superior brain tissue segmentation for the STORM_GLIO dataset, producing well-defined boundaries compared to the output of CaPTk. In contrast, CaPTk resulted in significant information loss in brain tissue segmentation, yielding notably poor outcomes.

Figure 2.9 presents a comparative analysis, quantified using the average DSC, to evaluate the performance of the skull-stripping tools. A fifty-patient cohort of the UPenn-GBM dataset was used for this analysis. Quantitative analysis revealed that HD-BET exhibited superior segmentation accuracy, achieving an average DSC of 97.9% when utilising only T1 sequences, which yielded the highest DSC among all individual MRI sequences. In contrast, CaPTk achieved an average DSC of 94.6% by incorporating all four MRI sequences.

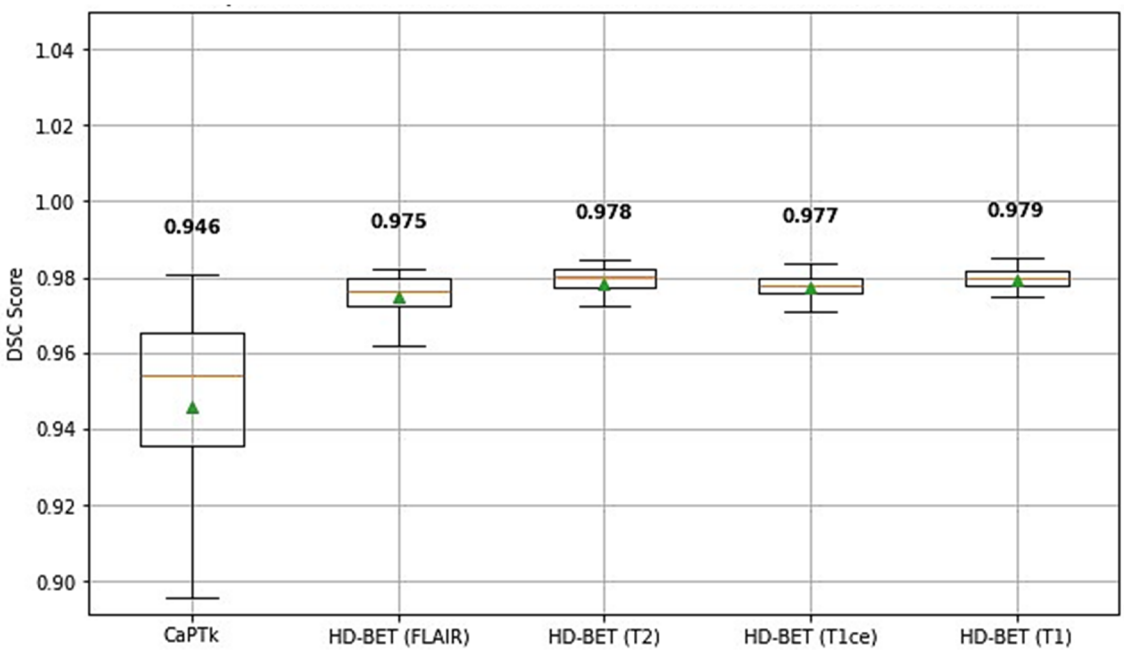


Figure 2.9 Skull Stripping Result Comparison: The box plot compares CaPTk using all four MRI sequences against HD-BET applied separately to T1, T1ce, T2, and FLAIR, with mean values displayed at the top.

The execution time (in seconds) was demonstrated in Table 2.2. When running on a CPU, HD-BET required a significantly longer processing time than CaPTk. An automated tumour delineation tool for comparative analysis, integrated with CaPTk's DeepMedic algorithm, was performed after the successful execution of both the BraTS and the proposed clinical pipelines.

Table 2.2 The average DSC (%) for skull stripping, Execution time (seconds) between the CaPTk and HD-BET tools. Execution time was assessed on i5 4-core CPU with 16 GB RAM, GTX 1050 4 GB VRAM.

	MRI-Scans						
Models	T1	T1ce	T2	FLAIR	ALL Scans	Execution Time	
HD-BET	97.9	97.7	97.8	97.5	-	88.62 (GPU)	1920.36 (CPU)
CaPTk	-	-	-	-	94.6	315.56 (CPU)	

TC segmentation region was replaced with a clinician-approved GTV region after expert clinical consultation and thorough validation [92]. A comparative visualisation of the standard and modified pipeline implementations within an example patient from the STORM_GLIO dataset is presented in Figure 2.10 with each pipeline's DSC. The DSC metric obtained from the quantitative assessment of the

standard BraTS protocol was 84.01%. The integration of advanced methodological modifications in the clinically optimised pipeline framework demonstrated superior quantitative outcomes, yielding an DSC of 89.38%.

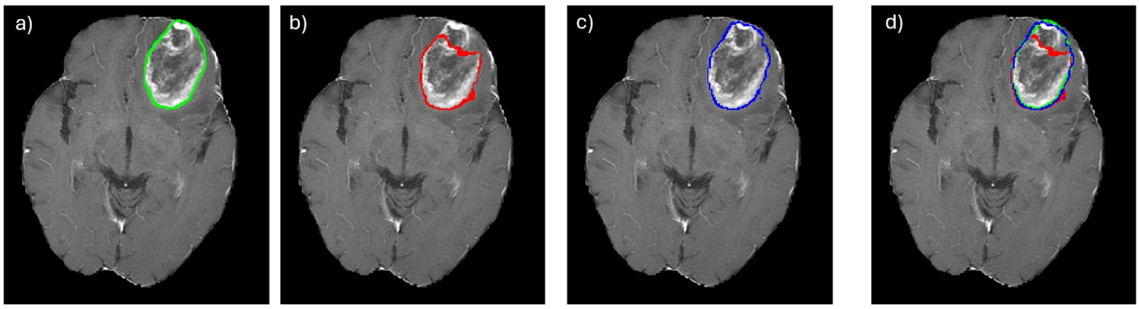


Figure 2.10 DICOM-Compatible Alignments: The alignment results are demonstrated using the T1ce MRI scan, showing a) the GTV region (green, approved by clinicians), b) the automated TC region generated by the BraTS pipeline (red; 84.01% DSC), and c) the automated TC region produced by the proposed pipeline (blue; 89.38% DSC). d) The overlapping of each label is visualised on the same scan.

Average DSC metrics of automated tumour segmentation were derived from the STORM_GLIO dataset for both the BraTS Pipeline and the Proposed Pipeline, yielding values of 68.09% and 73.74%, respectively, as depicted with Box plots in Figure 2.11.

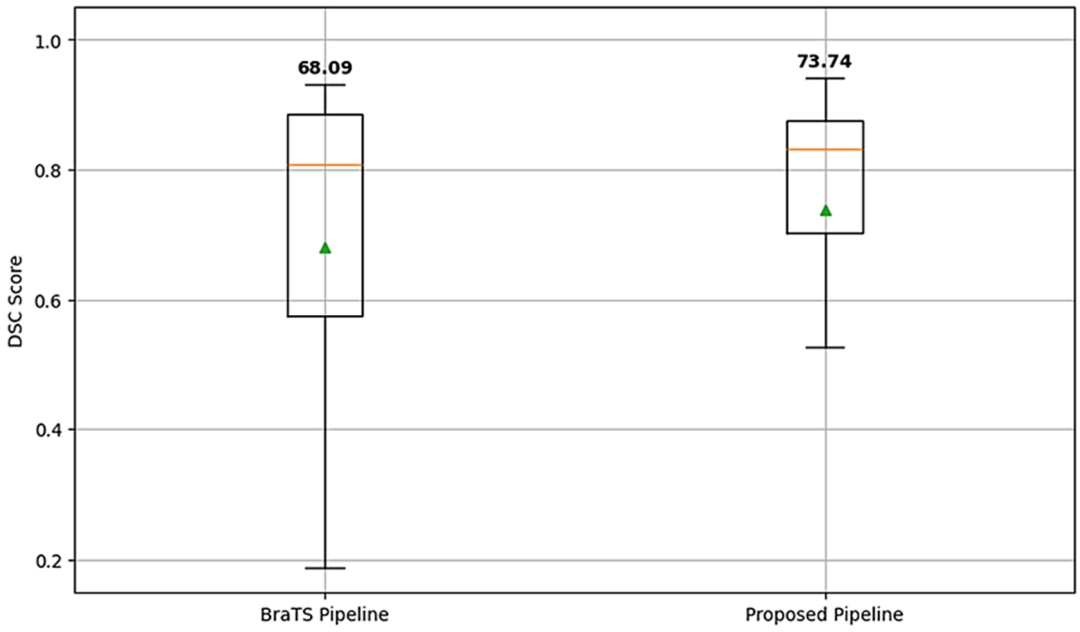


Figure 2.11 Tumour segmentation on MRI scans using CaPTk showed 68.09% DSC accuracy for the BraTS Pipeline and 73.74% for the Proposed Pipeline.

2.4 Discussion

The preprocessing steps plays an important role in medical image analysis protocols. Therefore, it may require modification and optimisation to effectively address the challenges introduced by the variability and complexity of clinical scenarios. The experimental results showed that spatial registration to original DICOM coordinates improves automated tumour segmentation, with a notable impact on boundary region segmentation accuracy. The change in dimensions from standardised matrix parameters (240×240×155 for SRI24) to native scan resolutions (e.g., 256×256×20 slices) creates significant challenges for quantitative medical image analysis. These challenges are more pronounced in lower-quality clinical MRI acquisitions, which demonstrate significant heterogeneity compared to the standardised, high-fidelity imaging protocols prevalent in public datasets. Notably, the recent BraTS challenges (BraTS 2023, 2024 and 2025) emphasised limited-resource, lower-quality MRI scans, such as those from Sub-Saharan African datasets [226], [227], [228], aiming to encompass diverse clinical settings, an important consideration for future research directions. Notably, the recent BraTS challenges (BraTS 2023, 2024 and 2025) emphasised limited-resource, lower-quality MRI scans, such as those from Sub-Saharan African datasets [226], [227], [228], aiming to encompass diverse clinical settings, an important consideration for future research directions. The observed qualitative divergence reflects the challenges of real-world practice, where image acquisition frequently occurs under non-ideal conditions. This underscores the necessity of modified preprocessing pipelines to optimise both the performance and reliability of automated segmentation and radiomic analysis.

When MRI scans are subjected to interpolative up-sampling to achieve enhanced spatial resolutions (such as SRI24; 240×240×155), inconsistencies and morphological perturbations are commonly introduced. The perturbations arising from interpolative up-sampling negatively affect the accuracy of automated skull stripping and automated tumour segmentation. This, in turn, leads to a cascading degradation in downstream quantitative imaging analyses, with a particularly noticeable impact on radiomic analysis [70], [71], [79]. The quantitative

methodology of radiomics, which highly relies on accurate tumour delineation for hand-crafted feature extraction [67], exhibits reduced robustness and impaired external validity when subjected to segmentation (ROI) errors. These findings highlight the fundamental importance of preserving native scan integrity during preprocessing to avoid distortions and ensure methodological consistency. Further research is necessary to assess the proposed pipeline's effectiveness in radiomic studies. Additionally, the algorithmic refinements within our modified preprocessing framework resulted in significant increases in segmentation accuracy, as evidenced by improved DSC metrics.

Skull stripping task remains an important step in medical image analysis for brain tumours. However, there is ongoing research exploring the feasibility of conducting brain tumour segmentation without skull stripping. HD-BET has been regarded as one of the closest approaches to manual segmentation references (gold standard), yet skull stripping still has a significant impact as part of the preprocessing pipeline, as highlighted by Pacheco et al. [79]. Additionally, HD-BET requires only a single MRI scan to perform brain tissue extraction, providing critical flexibility in scenarios where data scarcity or sparsity poses a challenge in clinical applications [69], [189]. These observed improvements, achieved through the systematic mitigation of diverse technical challenges such as preserving image quality, reducing interpolation artefacts, and improving delineation accuracy, suggested that the proposed preprocessing pipeline may enhance computational reliability. This can potentially lead to more accurate and adaptable segmentation outcomes across varying imaging conditions. This advancement in automated segmentation tasks is essential for radiomic research, as it enhances both the reproducibility and generalisability of radiomic models [69]. By enabling precise tumour segmentation, it may facilitate more reliable feature extraction, with the potential for wider applications across varied patient datasets and clinical environments [71]. This might enable the generation of reproducible and transferable results, ultimately enhancing the clinical utility of radiomic analyses and related computational procedures and paving the way for their broader adoption in clinical practice.

2.5 Conclusions

This chapter presented refinements to the standardised pipeline for medical image analysis, demonstrating substantial gains in the precision of both automated brain tissue extraction and tumour segmentation. These methodological adjustments are designed to be flexible even within the constraints of typical clinical environments, where data availability and quality can be limited, by utilising a single MRI scan with HD-BET for skull stripping instead of CaPTk. Moreover, these enhancements are intended to facilitate straightforward integration into existing clinical workflows, supporting advancements in medical image analysis, such as radiomics. The modified clinical pipeline's ability to address technical challenges may enhance its practical usability and support efforts to bridge the gap between current research and real-world clinical implementation. This work lays the groundwork for subsequent chapters that will demonstrate its practical application and explore its potential in both data-rich and resource-constrained environments, contributing to ongoing developments in the field of medical imaging and its impact on patient care.

3. A Novel Swarm Intelligence-Driven Feature Selection for Interpretable Machine Learning in Glioblastoma Multiforme Overall Survival Analysis

3.1 Introduction

This chapter investigated the potential of radiomic analysis using high-quality imaging cohorts, free from the typical limitations of clinical data. This investigation will employ the BraTS standardised preprocessing and utilise up to ten interpretable RFs derived from three different tumour regions aligned with the BraTS contouring format. This approach not only aimed to match the performance of DL-based analyses but also to enhance the transparency of ML models with a novel feature selection method and uncover potential biomarkers.

This research introduced a novel hybrid feature selection framework, building upon the foundations laid by Meneghetti [215] and Al-Tashi [208], integrating PSO-supported LASSO into a conventional ML pipeline to enhance feature selection efficacy. The research aimed to improve upon the reproducibility of a radiomic-based survival analysis in GBM by implementing a standardised approach to data preprocessing and segmentation. The validation strategy incorporated a dual approach: direct external validation utilising open-access institutional data repositories and systematic comparison with previously published studies based on established radiomics research guidelines. For the first time, this study explored an SI-based feature selection approach with traditional ML models applied to GBM time-to-event radiomic analysis, evaluating its capability to achieve statistically significant patient risk stratification.

3.2 Material and Methods

The radiomic-based model development and validation framework utilised a total of 276 GBM cases derived from two distinct sources: 1) a multi-institutional dataset from BraTS 2020 Challenge, comprising 236 cases [211], [212], [213], and (2) a single-institutional dataset from RHUH-GBM, contributing 40 cases [214].

RHUH-GBM covers 40 patients suffering from GBM who underwent surgical treatment between January 2018 and December 2022. To be included in this study, patients were required to meet three criteria: (1) a confirmed diagnosis of GBM; (2) surgical resection classified as either GTR or Near total resection (NTR), ensuring an Extent of resection (EOR) ≥ 95% with no residual enhancement; and (3) availability of complete MRI data, including preoperative, early postoperative (within 72 hours), and recurrence imaging. However, this particular analysis focused solely on preoperative MRI scans. The imaging protocol included T1, T2, T1ce, FLAIR, and apparent diffusion coefficient (ADC) sequences, ensuring comprehensive radiological assessment. For both datasets, four preoperative MRI sequences (T1, T1ce, T2, and FLAIR) were utilised in this study, following the RANO guidelines [93] to ensure standardised imaging assessment.

The study framework, represented in Figure 3.1, implemented time-to-event analysis, stratifying into low- or high-risk groups to assess OS, with the endpoint defined as the interval between initial pathological diagnosis and either death (censored=1) or final follow-up (censored=0). To ensure robust model development and validation, BraTS 2020 dataset was randomly divided into a discovery cohort (n = 188, 80%) for internal validation during model training and selection, and an unseen test set (n = 48, 20%) for the final evaluation of the selected model. RHUH-GBM cohort was preserved as an independent external validation dataset.

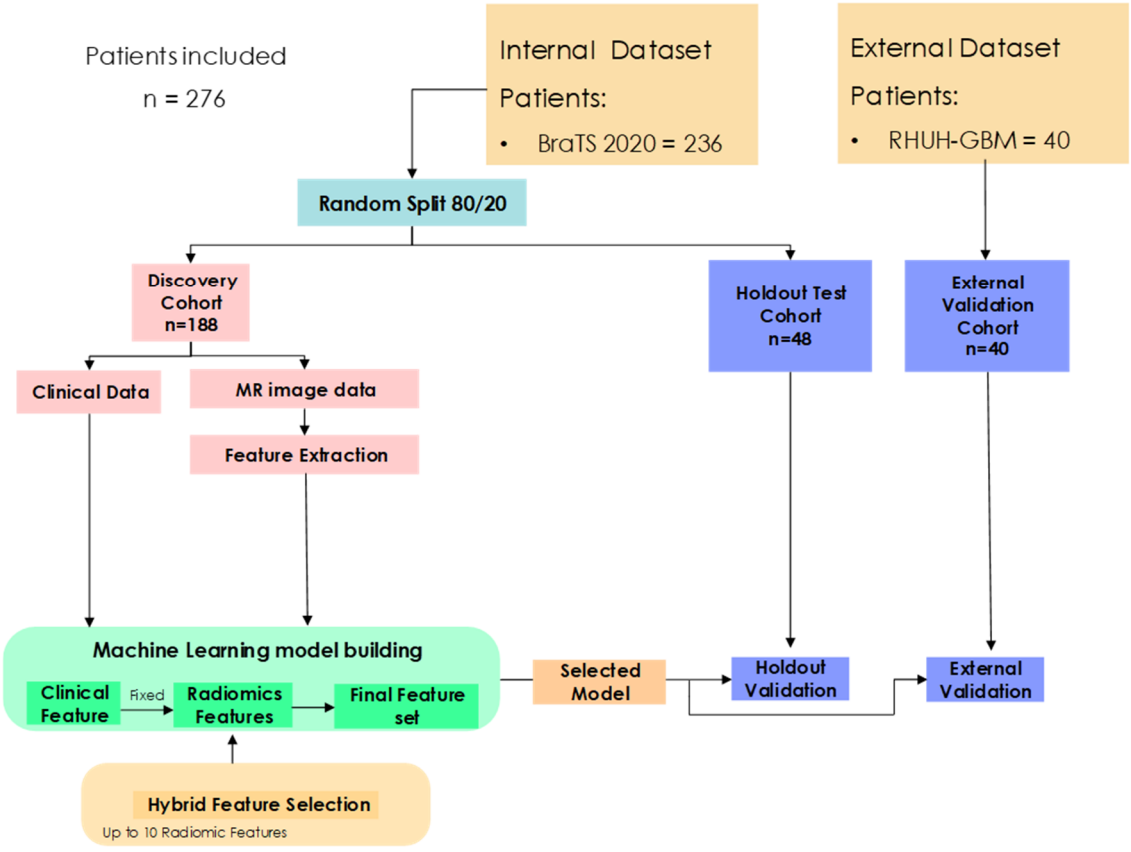


Figure 3.1 The study Design: This study utilised an internal dataset as the discovery cohort for feature selection and model development, a holdout test cohort for validating the model on unseen data, and an external dataset for additional validation.

Clinical parameters were initially derived from the discovery cohort and subsequently integrated with engineered RFs extracted from three distinct tumour regions (ET, TC and WT) on pre-treatment MRI scans. Radiomic feature selection, followed by the construction and optimisation of risk stratification models, was performed using the discovery cohort. Model parameters were tuned to maximise predictive performance on this dataset. The predictive performance of these models was then evaluated via a two-stage validation strategy: first, internal validation using a hold-out test set from BraTS 2020 dataset; and second, external validation using the independent RHUH-GBM dataset to rigorously assess model generalisability. Risk stratification model performance was evaluated using the Concordance Index (C-index), Kaplan-Meier curves (KM plots), and the log-rank test.

The preprocessing steps of the BraTS dataset was described in Chapter 2; in summary, uniform spatial sampling was established through isotropic voxel resampling ($1 \times 1 \times 1$ mm^3) and consistent $240 \times 240 \times 155$ matrix dimensions was yielded across all MRI scans.

RHUH-GBM dataset was processed following an image pre-processing pipeline [214], which maintains consistency with BraTS 2020 dataset. The pipeline included the following steps: (1) conversion of imaging data from DICOM to NIfTI format; (2) rigid registration of T1ce scans to the SRI24 anatomical atlas, followed by alignment of T1, T2, FLAIR to the transformed T1ce scan; (3) brain extraction on all co-registered volumes using the DL-based tool; and (4) Z-score intensity normalisation. Both datasets maintain uniform voxel resolution ($1\times1\times1$ mm^3) and matrix size ($240\times240\times155$) across all MRI scans. To ensure reproducibility and facilitate potential clinical translation for this study, image preprocessing followed the standardised guidelines of IBSI [66]. To maintain consistency across datasets, tumour segmentation followed BraTS 2020 Challenge protocol, identifying ET, TC, and WT regions. TC included ET and necrosis; WT encompassed ET, necrosis, and oedema. A two-step methodology was employed for tumour contouring in both BraTS 2020 and RHUH-GBM datasets. First, DL-based automatic segmentation was performed to segment tumour regions. This was subsequently reviewed and validated by neuroradiologists, ensuring consistency with clinical standards [211], [214]. A total of 1,980 ($4\times3\times165$) RFs were extracted per patient, derived from four MRI sequences, three tumour regions, and 165 features per region. These features were extracted using the MATLAB version of SPAARC (https://www.spaarc-radiomics.io/, accessed on 1 January 2025) [185], [186]. The extracted imaging features quantify tumour characteristics such as shape, texture, and intensity patterns. To ensure reproducibility and comparability, all features were extracted using a 3D approach and standardised following the guidelines of IBSI. The important image pre-processing parameters and extracted RFs are summarised in Table 3.1. For reproducibility, the details of preprocessing parameters were in Appendix Figure B- 1.

Table 3.1 The IBSI standardised preprocessing parameters for radiomic analysis

Parameters	
Voxel Spacing	[1, 1, 1]
Interpolation Method	Spline
Bin method	FBN
Bin value	64
Analysis Type	3D

Model development followed a systematic workflow designed with three different feature selection approaches to ensure robust model development and minimise overfitting risks. The model development workflow, as outlined in Figure 3.2, comprised four sequential phases: (i) feature pre-processing,

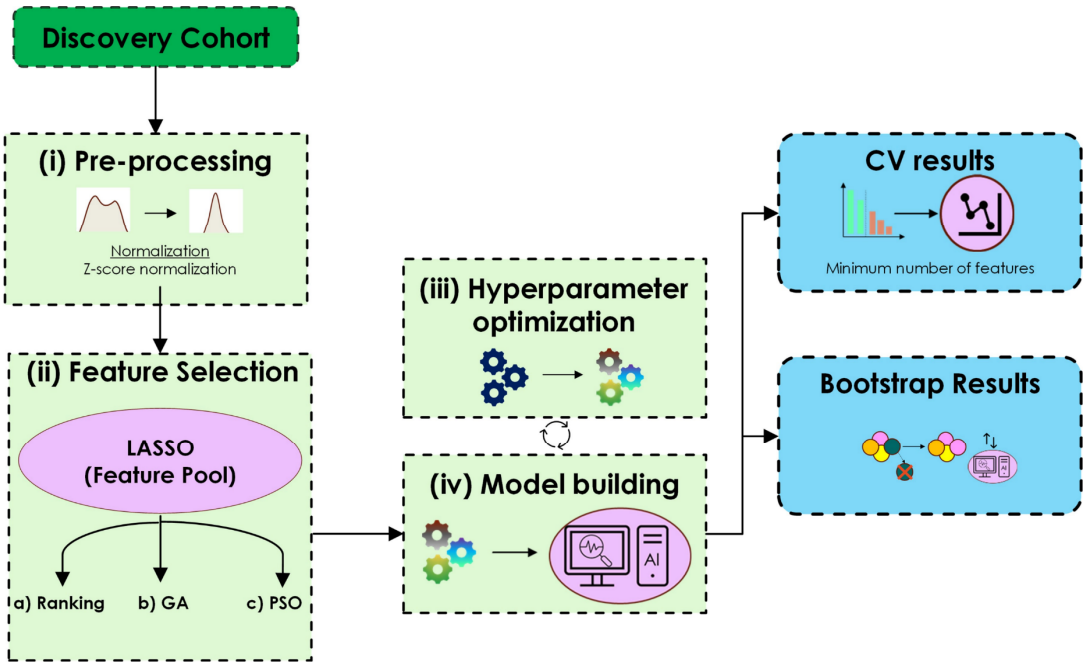


Figure 3.2 The study workflow for Feature Selection and Hyper-parameter optimisation.

(ii) feature selection (detailed in Appendix Figure B- 2 for LASSO-RANK, Appendix Figure B- 3, Figure B- 4 and Figure B- 5 for LASSO-GA, and Appendix Figure B- 3, Figure B- 6 and Figure B- 7 for LASSO-PSO), (iii) hyperparameter tuning, and (iv) model training with internal validation using data from the discovery cohort. All steps were performed using five-fold CV. The detailed description of the steps is as follows.

i) Z-score normalisation was implemented to address feature scale variability, ensuring reliable RFs. This normalisation was applied to the training dataset, with the computed parameters later used to normalise the holdout test set and external validation dataset.

ii)The feature selection process included three approaches: one established method and two novel variations utilising stochastic feature selection algorithms [221]. The established LASSO-RANK algorithm [229], documented in Leger et al. [230], provided the methodological foundation. Additionally, this study introduced a novel two-phase hybrid feature selection approach. The first phase was based on LASSO-RANK, which excluded only the feature ranking step while generating a feature pool. This pool was subsequently processed using algorithms, specifically GA [231] and

Table 3.2 Hyperparameters for PSO and GA were set based on the example source codes, with the exception of Maximum Features, which was adjusted to 9 (the total feature number for each feature subset) for GA, leading to improved model performance.

Algorithm	PSO	Algorithm	GA
Particle Number	30	Population Number	50
Estimator	LASSO	Estimator	LASSO
CV	3	CV	5
Scoring	Negative Mean Squared Error	Scoring	Negative Mean Squared Error
Max Iteration	10	Crossing-over Probability	0.5
		Mutation Probability	0.2
		Number of Generations	40
		Maximum Features	9
		Population Number	50

PSO [232] to optimise feature selection. The proposed models are referred to as LASSO-GA and LASSO-PSO. The LASSO-RANK method, serving as the baseline approach, employed a five-fold CV strategy to identify a subset of up to nine RFs, ranging from 2-feature to 9-feature subsets. However, the novel two-phase approach employed GA or PSO to refine the feature pool obtained from the regularised Cox regression for feature selection (Lasso). GA and PSO, unlike the deterministic selection of LASSO-RANK, use stochastic (randomness), iterative search processes to identify optimal feature subsets. The hyperparameter configurations for these algorithms are provided in Table 3.2. The details of each algorithm were provided in Appendix Figure B- 3. The estimator was LASSO, with a negative mean squared error metric for each nature-inspired feature selection step. Risk stratification was performed using two survival analysis models: Cox-LASSO and RSF [233]. These models have been tuned for survival analysis, optimising the ability to handle censored time-to-event data while prioritising predictive accuracy, model interpretability and reproducibility.

iii) Hyperparameters for each model were tuned using bootstrap resampling of the training (discovery) dataset. This approach was selected to mitigate the risk of overfitting and enhance model performance on unseen data.

iv)Following the radiomics guidelines outlined by van Timmeren et. al [67], the total number of features, including the clinical variable of patient age, was restricted to a maximum of ten. LASSO-RANK method identified from 2-feature to 9-feature RFs in each of the five CV folds, and these eight variant feature pools were then ranked by their selection frequency, utilising 200 bootstrap iterations.

Additionally, LASSO-GA and LASSO-PSO methods utilised the eight different feature pools for the final RFs, utilising k-fold CV and bootstrap methods. Model validation employed a bootstrap resampling approach (200 iterations) applied to the discovery cohort, utilising the selected feature subset to assess risk stratification performance. The C-index served as the principal metric for evaluating model robustness and predictive accuracy. The established workflow involved the development and optimisation of prognostic models using the discovery cohort.

Model performance was then assessed on two independent datasets: a holdout test set (the hold-out portion of BraTS 2020 data) and an external validation set (RHUH-GBM). This two-stage validation process was implemented to evaluate both model accuracy and generalisability to unseen data.

The log-rank test was used to compare survival distributions between the discovery and the holdout test sets. Additionally, these comparisons were conducted between the discovery cohort and the external validation cohort. Differences in continuous variables were evaluated using the Mann-Whitney U test. For evaluating the patient risk stratification performance of radiomic models, the KM curve was employed to examine the risk scores generated by the prognostic models, grouping patients into low- and high-risk cohorts based on the median risk score serving as the cut-off or the threshold. The statistical significance of the difference in survival distributions between the two risk groups was evaluated using the log-rank test. C-index was used to evaluate the stratification performance of the prognostic models, with 200 bootstraps applied to the discovery, the holdout test and the external validation cohorts to compute the confidence index (95% CI) [234].

Statistical analyses, including survival analyses, were conducted with the Python programming language (version 3.10). Statistical significance was defined as a p-value less than 0.05. The study of image preprocessing and statistical analysis is illustrated in Figure 3.2. The relevance of features was assessed using permutation feature importance, as implemented in Sklearn v1.5.2. Additionally, PSO and GA algorithms were implemented in ps-optimize v2.0.4 and sklearn-genetic v0.6.0, respectively.

3.3 Results

The clinical characteristics of the discovery, hold-out test, and external validation cohorts are outlined in Table 3.3. The discovery cohort had a median OS of 12.05 months, while the hold-out test cohort demonstrated a slightly higher median OS of 14.44 months; however, there was no statistically significant difference ($p = 0.58$). Likewise, in the external validation cohort, the median OS was recorded at 12.13 months, which did not differ significantly from the discovery cohort ($p = 0.59$).

Table 3.3 Characteristics of clinical variables for discovery and hold-out test and external validation datasets.

Variable	Discovery Dataset Median (range)	Holdout Test Dataset Median (range)	Statistical Cohort Comparison	Discovery Dataset Median (range)	External Validation Dataset Median (range)	Statistical Cohort Comparison
Age (Years)	62.4 [19.0-86.7]	60.6 [27.8-85.9]	U: 0.55, p-value: 0.3	62.4 [19.0-86.7]	64.0 [45.0-78.0]	U: 0.46, p-value: 0.55
OS (months)	12.05 [0.17-52.03]	14.44 [1.0-58.9]	U: 0.47, p-value: 0.58	12.05 [0.17-52.03]	12.13 [3.0-41.47]	U: 0.47, p-value: 0.59

LASSO-RANK and LASSO-GA methods performed best with the 6-feature pool, whereas LASSO-PSO demonstrated superior performance with expanded feature sets, requiring a 9-feature pool to achieve the highest performance. To mitigate potential multicollinearity among the RFs, the first step was performed before feature selection. This involved calculating the Spearman correlation coefficients between all pairs of RFs and excluding features exhibiting high correlations (Spearman's $\rho > 0.95$). This reduced the initial feature set to 767 RFs. The final feature pool for each selection method was established as the collection of features identified across all five folds, excluding repetitions. This approach yielded 16 RFs for both LASSO-RANK (the 6-feature pool) and LASSO-GA (the 6-feature pool), and 18 RFs for LASSO-PSO (the 9-feature pool). In the LASSO-RANK approach, the six features with the highest selection frequency were incorporated into the final model, setting the selection threshold to six due to the 6-feature configuration. Conversely, the internal validation procedures demonstrated guided the decision of the final feature sets for LASSO-GA and LASSO-PSO, which comprised 2 RFs and 10 RFs, respectively, as presented in Table 3.4.

Table 3.4 The selected RFs for each feature selection method, with their respective MRI sequences and Labels indicated in parentheses. Notably, Morphological features are associated with contour morphology (shape) rather than a specific MRI sequence. The MRI sequences utilised for radiomic feature extraction are specified in parentheses (FLAIR sequence, T1 sequence), along with ROIs used for feature extraction (ET label, TC label, WT label).

LASSO-RANK (6 feature)	LASSO-GA (2 feature)	LASSO-PSO (10 feature)
morph_volume (TC label) morph_av (TC label) morph_comp_1 (TC label) morph_diam (TC label) morph_pca_maj_axis (TC label) morph_pca_elongation (TC label)	stat_skew (T1 sequence, TC label) dzm_zdnu_3D (FLAIR sequence, TC label)	morph_pca_maj_axis (TC label) morph_pca_flatness (TC label) morph_comp_1 (TC label) morph_vol_dens_aee (TC label) morph_area_dens_aee (TC label) ngl_dc_entr_3D (FLAIR sequence, WT label) dzm_zdnu_3D (FLAIR sequence, TC label) szm_lgze_3D (FLAIR sequence, ET label) szm_lgze_3D (FLAIR sequence, TC label) stat_skew (T1 sequence, TC label)

Hyperparameter tuning for Cox-LASSO and RSF was performed through 200 bootstrap iterations on the entire discovery cohort by utilising the selected RFs and the clinical feature (age). The final hyperparameter configurations, utilising k-fold CV and bootstrap methods, for each model with each feature selection method, are provided in Appendix Table B- 1. Prognostic models were developed using the entire discovery cohort, utilising a clinical-radiomic signature that combined patient age with the final set of selected RFs. The LASSO-based feature selection approaches identified a subset of highly predictive RFs across multiple MRI sequences, effectively reducing the number of selected features while preserving predictive strength. The resulting Cox-LASSO model, recognised for its interpretability among ML methods [235], achieved the highest C-index of 0.64 on the internal validation dataset during the model-building step when utilising LASSO-PSO or LASSO-GA feature selection approach (Figure 3.3 b). In the out-of-bag bootstrap (OOB)

evaluation, the prognostic model utilising LASSO-PSO achieved a C-index of 0.67, outperforming the same model with LASSO-GA (C-index = 0.64) on the discovery dataset (Figure 3.3c), confirming its superior model robustness.

The results of RSF models using LASSO-PSO achieve a C-index with 0.64 in the internal validation dataset (Figure 3.3b). However, due to the overall poor performances, C-index values were included for the discovery, holdout, and external validation datasets for the best RSF model with LASSO-PSO, which achieved a bootstrap OOB C-index of 0.74, as presented in Appendix Table B- 2.

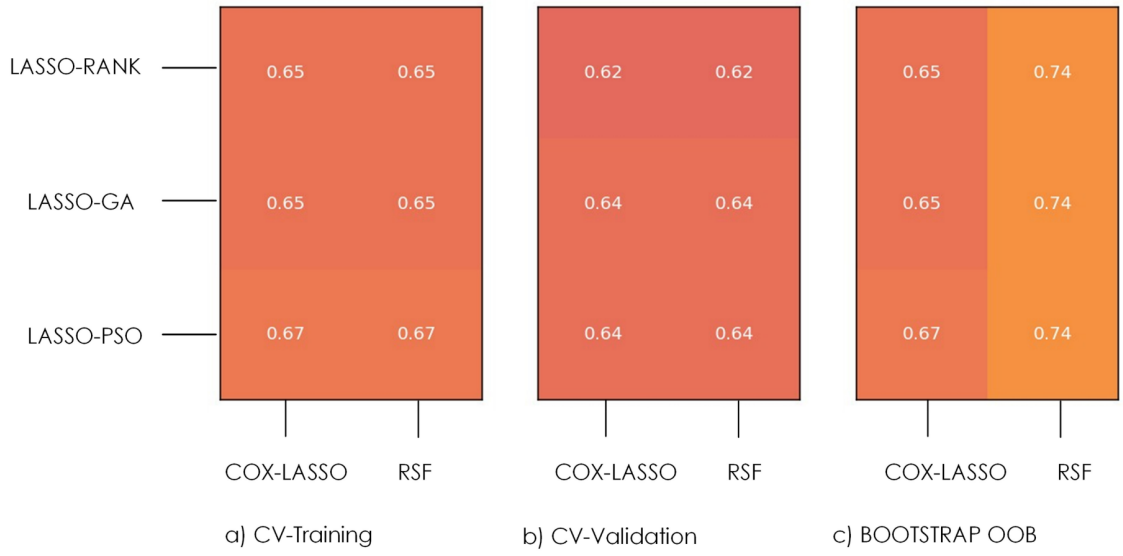


Figure 3.3 The C-index values for each model on each feature selection method and their corresponding ML algorithms were assessed for GBM time-to-event analysis. The results are presented as follows: (a) CV training performance, (b) CV validation performance, and (c) OOB evaluation.

Within the discovery cohort, the radiomic model using Cox-LASSO, built using a set of 10 RFs (Table 3.4), demonstrated notable predictive strength, as evidenced by a C-index of 0.64 (95% CI: 0.60–0.68). From the radiomic model, Dzm_zdnu_3D, a texture feature from 10 RFs achieved the highest hazard ratio (HR) of 1.15 (95% CI: 0.87–1.76) shown in Table 3.5. This feature calculates the distribution uniformity of zone frequencies across spatial distances.

Table 3.5 Univariate and Multivariate Cox regression analysis for Discovery and Holdout test datasets.

		Discovery cohort	Holdout test cohort	Discovery cohort		Holdout test cohort	
Model	Variable	HR [95% CI]	HR [95% CI]	p-Value	C-Index	p-Value	C-Index
Clinical Model	Age	1.32 [1.16-1.52]	1.78 [1.44-2.58]	3×10^{-2}	0.59 [0.55-0.64]	7×10^{-5}	0.71 [0.62-0.79]
Radiomic Model	morph_pca_maj_axis (TC label)	1.08 [0.77-1.47]	2.06 [0.85-8.72]	4×10^{-4}	0.64 [0.60-0.68]	2×10^{-2}	0.61 [0.52-0.72]
	morph_pca_flatness (TC label)	1.13 [0.81, 1.42]	1.39 [0.62, 3.74]				
	morph_comp_1 (TC label)	0.92 [0.57, 1.51]	2.81 [0.50, 15.15]				
	morph_vol_dens_aee (TC label)	0.89 [0.55, 1.45]	0.41 [0.07, 2.13]				
	morph_area_dens_aee (TC label)	1.12 [0.83, 1.70]	2.38 [0.26, 12.77]				
	ngl_dc_entr_3D (FLAIR sequence, WT label)	0.91 [0.76, 1.07]	0.41 [0.15, 0.84]				
	dzm_zdnu_3D (FLAIR sequence, TC label)	1.15 [0.87, 1.76]	1.14 [0.56, 2.51]				
	szm_lgze_3D (FLAIR sequence, ET label)	0.88 [0.45, 1.15]	0.51 [0.004, 27.53]				
	szm_lgze_3D (FLAIR sequence, TC label)	0.93 [0.69, 1.35]	1.23 [0.08, 14.69]				
	stat_skew (T1 sequence, TC label)	0.86 [0.65, 1.07]	1.06 [0.72, 2.47]				

Lower scores indicate a more homogeneous distribution of zones, whereas higher scores highlight localised clustering, indicative of increased intratumoral heterogeneity. Half of the selected features were extracted from two MRI sequences (FLAIR: 4/10; T1: 1/10), while the remaining half were shape-based features (Morphological: 5/10). These RFs exhibited low inter-feature correlations (Spearman's $\rho < 0.3$), indicating their relatively independent contributions to the model. In the hold-out test dataset, the radiomic model achieved a moderate C-index (0.61, 95% CI: 0.52–0.72). Morph_comp_1, a shape-based feature, from 10 RFs achieved the highest HR of 2.81 (95% CI: 0.50–15.15) shown in Table . This feature evaluates the similarity between the targeted ROI and a perfect sphere, helping as an indicator of morphological compactness.

In the training (discovery) dataset, the clinical–radiomic model, which integrates both the clinical feature (Age) and 10 RFs, demonstrated the best C-index (0.67, 95% CI: 0.63–0.70). As reported in Table 3.6, this model achieved a C-index of 0.71 (95% CI: 0.61–0.79) in the hold-out test dataset. Two RFs were identified as highly critical for defining high-risk groups, as their values exceeded 2.0 for HR. Specifically, szm_lgze_3D (FLAIR sequence, ET label) exhibited an HR of 2.46 (95% CI: [0.004–27.53]), while morph_pca_flatness (TC label) demonstrated an HR of 2.20 (95% CI: [1.00–6.30]). The szm_lgze_3D, a texture feature, is to calculate the presence and distribution of zones with lower grey-level intensities in ROI (ET label). The latter radiomic feature, morph_pca_flatness, is a shape-based feature described as the inverse ratio of the major to the least axis length. Its value approaches 1 as ROI (TC label) becomes closer to a perfect sphere, reflecting a higher level of shape uniformity. Additionally, the clinical-radiomic model achieved a C-index of 0.64 in the external validation dataset.

With a Kaplan–Meier curve cut-off of 0.012 (Appendix Table B- 3), the statistical significance of survival differentiation between stratified risk groups (low- and high-risk) was assessed with the use of the log-rank test. Significant differences were found in the training (discovery) dataset (p-values of 1×10^{-8}), in the hold-out test dataset (2×10^{-4}), and the external validation dataset (1×10^{-2})

Table 3.6 Multivariate Cox regression analysis for Clinical-Radiomic Model.

		Discovery cohort	Holdout test cohort	Discovery cohort		Holdout test cohort	
Model	Variable	HR [95% CI]	HR [95% CI]	p-Value	C-Index	p-Value	C-Index
Clinical-Radiomic Model	Age	1.36 [1.17-1.67]	1.92 [1.23-4.29]	1×10^{-8}	0.67 [0.63-0.70]	2×10^{-4}	0.71 [0.61-0.79]
	morph_pca_maj_axis (TC label)	1.23 [0.89-1.67]	1.72 [0.68-9.95]				
	morph_pca_flatness (TC label)	1.17 [0.81, 1.47]	2.20 [1.00-6.30]				
	morph_comp_1 (TC label)	0.90 [0.58, 1.53]	1.60 [0.52-8.76]				
	morph_vol_dens_aee (TC label)	0.94 [0.58, 1.53]	0.36 [0.08-1.91]				
	morph_area_dens_aee (TC label)	1.04 [0.74, 1.59]	1.98 [0.48-8.57]				
	ngl_dc_entr_3D (FLAIR sequence, WT label)	0.87 [0.70, 1.04]	0.54 [0.16-1.17]				
	dzm_zdnu_3D (FLAIR sequence, TC label)	1.14 [0.88, 1.73]	1.07 [0.37-2.18]				
	szm_lgze_3D (FLAIR sequence, ET label)	0.85 [0.42, 1.12]	2.46 [0.004-27.53]				
	szm_lgze_3D (FLAIR sequence, TC label)	0.99 [0.74, 1.38]	0.54 [0.08, 14.69]				
	stat_skew (T1 sequence, TC label)	0.88 [0.67, 1.11]	0.92 [0.55, 2.20]				

shown in Figure 3.4. The Kaplan–Meier plots demonstrated the model's consistent ability to differentiate low- and high-risk groups across both datasets. P-values demonstrated significant differentiation among low and high-risk groups, a finding that was consistently replicated in the external validation cohort. For each future, the permutation importance (feature importance) was demonstrated in Appendix Table B- 3. Further details, including feature weights and KM curve threshold, were provided in Table B- 3. Feature importance analysis and feature weights of the final model highlighted the clinical feature, Age, as the most influential predictor.

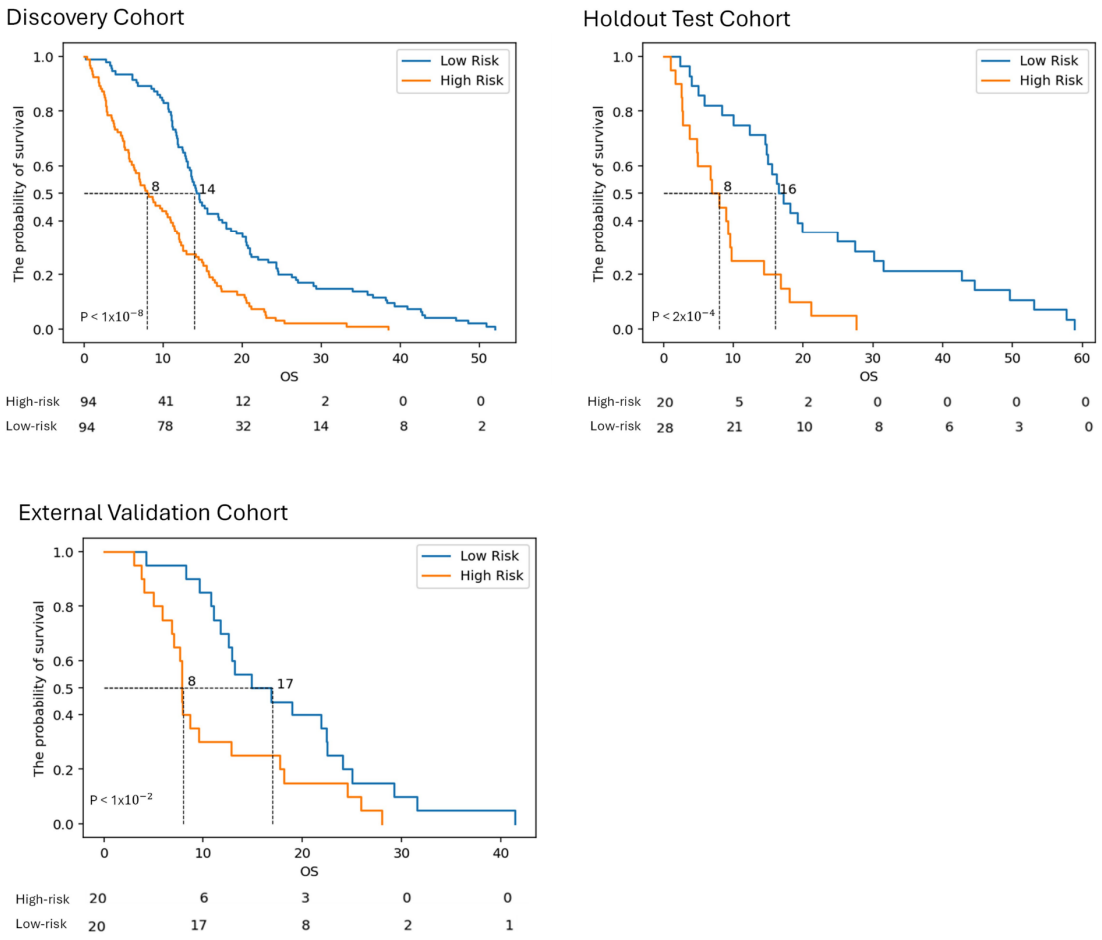


Figure 3.4 Kaplan–Meier curves display survival differences in the discovery, hold-out test and external validation cohorts, categorised into low- and high-risk groups by the Cox–LASSO model. The small p-values suggest strong statistical reliability in distinguishing between risk groups.

Among RFs, morph_pca_maj_axis demonstrated a predominant influence. This shape-based feature characterises the maximum axial extent of the ROI-encompassing ellipsoid (TC label), computed via principal component analysis and corresponding to the major eigenvalue (λ_{major}). Both age and morph_pca_maj_axis demonstrated an increased likelihood of a patient being stratified into the high-risk group. Their combined influence suggests that older age and a greater major axis length of the ROI-encompassing ellipsoid are significant predictors of poor prognosis and greater tumour aggressiveness (shown in Figure 3.5). As expected, both age and tumour size emerged as important predictors.

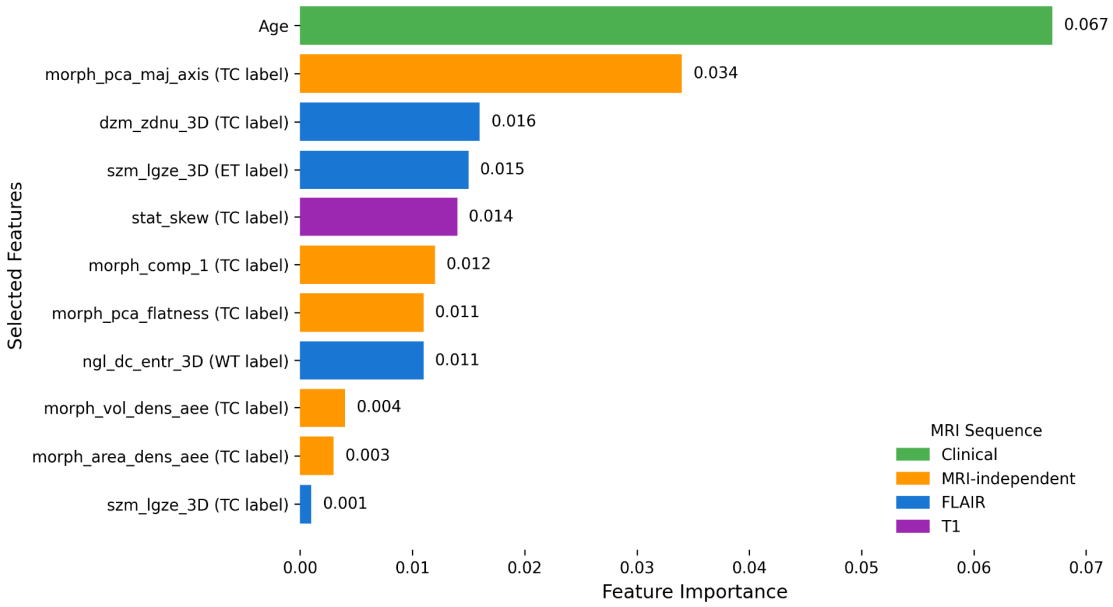


Figure 3.5 The feature importance for the final clinical-radiomic model.

3.4 Discussion

This study presents a risk-stratification model that incorporates a clinical feature and RFs to categorise GBM patients into low- and high-risk groups using preoperative MRI. The model development process followed rigorously established radiomics guidelines, emphasising interpretability, structured methodology, reproducibility, and generalisability. The feature selection process identified 10 RFs, sourced differentially from FLAIR and T1 MRI sequences. MRI sequence-independent morphological features predominated the selected feature set (5/10 RFs), with texture-based features derived from FLAIR comprising the second largest

group (4/10 RFs). The contribution from the T1 sequence was limited to a single first-order feature. With the exception of two texture-based features derived from ROIs of ET and WT labels, all RFs were extracted utilising the TC label. The clinical–radiomic model achieved a C-index of 0.71 in the hold-out test dataset, demonstrating statistically significant stratification between low- and high-risk groups. In the external validation dataset, the model attained a C-index of 0.64, with a statistically significant log-rank test p-value, further supporting its predictive robustness. Research reproducibility was ensured through multiple measures: employment of open-access datasets for both primary and external validation analyses, public release of code implementations, and detailed documentation of radiomic feature extraction and processing protocols. A comparative analysis of this study's results against prior research was presented in Table 3.7. To ensure consistency in comparison, only studies that utilised RFs (deep features and engineered features) and clinical information were included, while those relying on subjective assessments (e.g., VASARI features) or RFs with low reproducibility risk prior to the first IBSI study [66] were excluded. While Fathi et al. [219] exemplifies typical single-centre research constraints, Gomaa et al.'s study [217] uniquely distinguishes itself through its implementation of external validation protocols, enabling robust methodological comparisons. To minimise study limitations, we expanded the patient cohort by integrating multi-centre datasets, including BraTS 2020 for internal study and RHUH-GBM for external validation. While this approach enhances the study's robustness, potential biases may still exist, underscoring the need for additional validation in different clinical environments to strengthen its clinical applicability. Verduin et al. [218] reported a C-index of 0.69 with a model that included five RFs and five clinical features. Fathi et al. [219] achieved a C-index of 0.70 using a model composed of 24 RFs and three clinical features. Concordance indices of 0.71, 0.67, and 0.62 were achieved by Gomaa et al. [217] for the UPenn-GBM, UCSF, and RHUH-GBM cohorts, respectively, using a methodology that incorporated an unlimited number of deep features alongside four clinical parameters.

Table 3.7 The comparison of the proposed study with recent radiomics studies.

Study	ML Model	Log-rank Test Significance	C-index (Dataset)	Feature Details	IBSI 1,2 Standardised	Feature Limitations (3-10 features)
Verduin et al. [218]	Multimodel Cox regression	Yes	0.69 (Maastricht UMC + and Radboudumc)	5 RFs (engineered), 5 Clinical Features (age, sex, EOR, Adjuvant treatment, MGMT)	Partially (not aligned with IBSI 2)	Yes
Fathi et al. [219]	Cox-PH	Yes	0.7 (UPenn-GBM)	24 RFs (deep features, engineered features), 3 Clinical Features (age, sex, EOR)	No	No
Gomaa et al. [217]	Transformer-based DL model	Yes	0.71(UPenn-GBM), 67 (UCSF),62 (RHUH-GBM)	No info for number of Deep features, 4 Clinical Features (age, sex, MGMT, kps)	No	No
Al-Tashi et al. [208]	Swarm DeepSurv (SI-based DeepSurv)	No (0.14>0.05)	0.61 (TCGA-GBM)	49 RFs, Clinical Feature (Not provided)	Yes	No
Proposed	Cox-LASSO	Yes	0.71 Brats 2020, 0.64 (RHUH-GBM)	10 RFs (engineered), 1 Clinical Feature (Age)	Yes	No
Proposed (10 limit)	Cox-LASSO	Yes	0.71 Brats 2020, 0.63 (RHUH-GBM)	7 RFs (engineered), 1 Clinical Features (Age)	Yes	Yes

Similarly, Verduin et al. [218] attained a C-index of 0.69; their approach utilised convolutional filters that did not adhere to IBSI guidelines for convolutional filter implementation [236]. Al-Tashi et al. [208] developed a time-to-event analysis model using SI algorithms with a DL-based survival model and 49 RFs, reporting a

C-index of 0.61 and a non-significant p-value of 0.14 ($p > 0.05$). Unlike most radiomics studies that rely on Cox regression, Al-Tashi et al. along with Gomaa et al., explored alternative modelling approaches: SwarmDeepSurv, transformer-based DL models for survival analysis, which are still a challenge for clinical implementation due to their "black box" nature.

Our study achieved C-indices of 0.71 (BraTS 2020) and 0.64 (RHUH-GBM) using only 10 RFs extracted exclusively from FLAIR and T1 MRI sequences, along with the clinical feature Age. This performance matched or exceeded that of previous studies while using a minimal feature set. It is important to highlight that our study, which employs interpretable models (Cox-LASSO) and reproducible RFs in accordance with IBSI guidelines [66], demonstrated superior performance on the same external validation dataset (RHUH-GBM) when compared to the methodology of Gomaa et al. [217]. By systematically testing feature selection techniques and ML models, our novel SI-based feature selection method, LASSO-PSO, identified 10 RFs with the use of a clinical feature. Among the 10 RFs, morph_pca_maj_axis, a modality-independent morphological feature, exhibited the highest generalisability in the validation set.

This feature selection method enhanced the model's adaptability, enabling robust predictions across different healthcare settings, from small clinics to large research institutions, and ensuring generalisability across diverse datasets. Analysis of feature weights and importance (shown in Table B- 3) indicated that high-risk patients tend to exhibit greater axis length (the major eigenvector from PCA analysis), as extracted by the morph_pca_maj_axis feature, from a morphological feature family. Moreover, the zone distance non-uniformity (ZDNU: dzm_zdnu_3D), derived from the Gray Level Distance Zone Matrix (GLDZM), measures the distribution uniformity of zones across distances. Higher values indicated greater tumour heterogeneity, correlating with an increased likelihood of aggressive tumour types and high-risk patient classification. The two most influential RFs were quantified using the TC label. Among these features, dzm_zdnu_3D was derived from the FLAIR MRI sequence, emphasising its role in characterising tumour

heterogeneity. In this study, LASSO-PSO outperformed both LASSO-RANK and LASSO-GA in terms of identifying an optimal feature subset for radiomic model development. This result is consistent with the findings of a detailed review of SI-based feature selection methods conducted by Rostami et al. [221], who similarly found the PSO algorithm to be effective for feature selection in comparison to other SI algorithms, especially in selecting the minimum number of features.

To the best of our knowledge, this is the first SI-based feature selection approach with traditional ML models to demonstrate statistically significant risk group stratification (time-to-event analysis). Furthermore, the proposed model demonstrated superior or comparable performance despite utilising a restricted set of clinical and RFs. To ensure compliance with van Timmeren et al.'s guideline [67], we excluded RFs presenting low feature importance (<0.01, shown in Figure 3.5), namely, szm_lgze_3D from FLAIR MRI sequence and TC label, and *morph_vol_dens_aee* and *morph_area_dens_aee*, both extracted from TC label from the initial set of 10 RFs. This modified model retained discriminative capacity, yielding concordance indices of 0.71 and 0.63 for BraTS and RHUH-GBM cohorts, respectively. While the predefined limitation of 3–10 features played an important role in shaping the feature pool, SI-based feature selection may achieve greater performance if a larger set of RFs is considered, a prospect that merits exploration in future studies.

Within the domain of medical research, significant challenges arise from data scarcity and imbalance, attributable to the infrequent occurrence of certain conditions, limited patient cohorts, and gaps in clinical records, all of which obstruct the collection of extensive datasets [189]. In response to clinical data challenges, our study incorporated only the clinical variable of age due to incomplete clinical information across datasets. While this choice represents a limitation, it enabled us to maximise the patient cohort. Furthermore, the utilisation of only two MRI sequences, FLAIR and T1, enhances the feasibility of clinical implementation by reducing dependency on multiple MRI sequences. The present research addressed potential retrospective sampling bias through the utilisation of multi-institutional

data (BraTS 2020 dataset), further strengthened by external validation (RHUH-GBM dataset). Future model optimisation pathways encompass the integration of expanded clinical parameters (age, genetic markers, MGMT, survival metrics, KPS) and comprehensive omics data (genomic, pathomics etc.). The present study intentionally excluded DL-based RFs from the analytical framework, acknowledging their potential to improve survival prediction accuracy but prioritising reproducibility and interpretability, which currently restricts their clinical use [237]. To reduce the risk of overfitting, our workflow integrated hyperparameter optimisation alongside data resampling techniques (bootstrap). Within this optimised framework, the performance of the risk-stratifying model was thoroughly evaluated using both a hold-out test dataset and an external validation dataset. Also, this study achieved a score of 96.8% from METhodological RadiomICs Score (METRICS) [238], which is a quality score for radiomic research. As illustrated in Appendix Figure B- 8, future radiomic research needs a thorough evaluation of two critical metrics from the current results: the robustness analysis (item #14) and uni-parametric imaging comparison (item #23).

3.5 Conclusions

We developed and validated an interpretable clinical–radiomic model in this study, leveraging a novel SI-based hybrid feature selection method to stratify patients with GBM diagnosis into high- and low-risk categories based on OS. This study implemented an SI-based hybrid feature selection method in combination with a traditional ML model, Cox-LASSO, while following radiomic research guidelines. Additionally, it addresses important clinical challenges, enhances model interpretability and minimising feature number, underscoring its potential for clinical applicability and reproducibility. This framework, based on the previously specified criteria, achieved better predictive performance than that recently reported in prior studies by Poursaeed et al. [216]. Our model utilises 10 independent RFs from the FLAIR and T1 MRI sequences, along with the clinical feature, age, as predictive variables. With the recent standardisation of convolutional filters under the IBSI consensus guidelines [236], we plan to integrate standardised convolutional filters for engineered feature extraction in the future. In

addition, our aim is to be actively engaged in research that explores the potential of SI-based feature selection and model development to elevate predictive performance.

Future research directions encompass two key trajectories: first, the integration of comprehensive clinical parameters (age, sex, KPS, MGMT etc.) and expanded imaging modalities (PET, CT, ultrasound etc.) into the analytical framework; second, the optimisation of engineered feature extraction through strategic reduction of both MRI sequence requirements and label (ROI) numbers. Future performance optimisation frameworks include both imaging advancement through diffusional and functional MRI sequence integration and molecular characterisation expansion via multi-omics biomarker incorporation (genomic, transcriptomic, and metabolomic datasets). The integration of additional imaging modalities, supported by larger, multi-institutional external validation cohorts, has the potential to enhance the accuracy and reliability of predictive models, thereby enhancing clinical decision-making in GBM treatment and management. All RFs, code are publicly accessible through the repository at https://git.cardiff.ac.uk/c21099143/si-feature-selection-for-radiomics.

4. Reproducible Radiomic Analysis for Overall Survival Prediction in Glioblastoma Multiforme

4.1 Introduction

In this chapter, we examined a minimum set of engineered RFs to develop and validate traditional ML models that offer reproducibility and interpretability for prognostic assessment of OS in GBM patients. The research prioritised identifying robust RFs derived from MRI sequences and a single clinical ROI, leveraging multi-institutional retrospective data to enhance practical implementation and clinical adoption of radiomic methodologies.

Previous studies have demonstrated the efficacy of MRI-based radiomic analysis in stratifying into high and low-risk groups (time-to-event analysis) for GBM patients [239], [240], [241], [242]. However, proposed radiomic models frequently employed an excessive number of features (n>10) [243], leading to an increase of overfitting issues and interpretability challenges [67]. This approach conflicts with established radiomic guidelines that highlight the importance of the interpretability and generalisability of radiomic models [67]. For the time-to-event analysis of GBM patients, radiomic models that successfully provide risk-based stratification require RFs that are stable, reproducible, and easy to interpret (n=<10). This has been achieved in other types of cancers by following established guidelines [215], [244]. In this study, the aim was to close this gap by focusing on reproducible, stable RFs and interpretable ML models. The model building focused on a minimal set of RFs to stratify GBM patients into high- and low-risk groups based on OS data. We address real-world data challenges under a different preprocessing pipeline for STORM_GLIO compared to the BraTS 2020 dataset, along with limitations like relying on a single region of interest, specifically GTV in neuro-oncology radiotherapy planning, and the limited MRI sequences and varying acquisition parameters across patients.

4.2 Material and Methods

4.2.1 Study Population

In this research, radiomic signatures were developed and validated using data from 289 GBM patients. Two datasets were utilised: (1) the publicly available BraTS Challenge 2020 [211], [212], [213], which included 236 GBM cases, and (2) a local dataset named STORM_GLIO. The STORM_GLIO dataset is a retrospective collection of patients diagnosed with GBM and treated at our institution between April 2014 and April 2018, comprising 53 eligible cases out of 108 patients. Both datasets included four preoperative MRI sequences: T1, T1ce, T2, and FLAIR, following the guidelines of the RANO working group [93]. Additionally, both datasets included information on OS and patient age.

4.2.2 Study Design

The study design is illustrated in Figure 4.1. OS was evaluated through a time-to-event analysis, which measured the time interval (in days) between the patient's initial pathology-based diagnosis and either their death (denoted as censored = 1) or their last confirmed living date (denoted as censored = 0). The study population was split into two groups utilising random allocation: 80% were assigned to the training dataset and 20% to the validation dataset to train a model for predicting time-to-event outcomes.

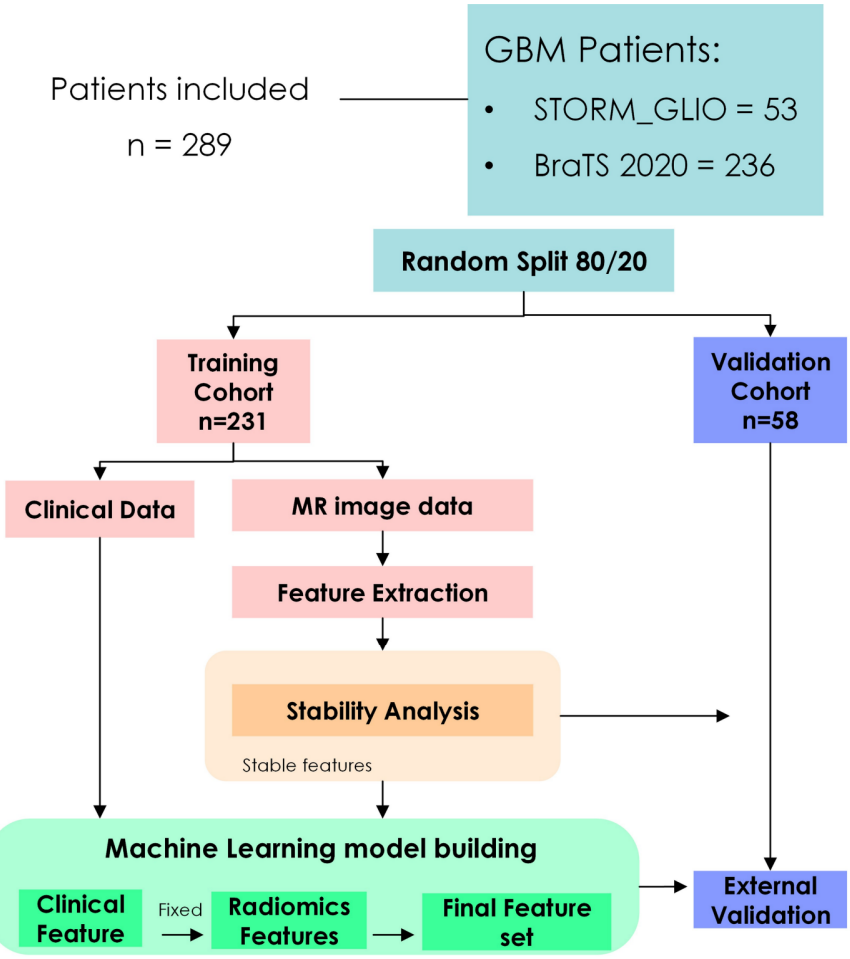


Figure 4.1 The study workflow: dataset splitting, feature extraction, stability analysis, model building, and model validation.

The process began with collecting clinical parameters from the training group. In radiotherapy planning, GTV is defined as the visible or palpable extent of the malignant tumour [91]. Duman et al. demonstrated a remarkable similarity between GTV and TC, emphasising their potential interchangeability [53], [92], [245]. Building on this finding, RFs were obtained from the pre-treatment MRI scans, specifically targeting the voxels within GTV for STORM_GLIO and the TC for BraTS dataset as defined by expert manual contouring. Next, these RFs were combined with the collected clinical parameters, potentially enhancing the generalisability of the radiomic models. The development and internal validation of risk-stratification model signatures were performed using the training dataset, followed by subsequent validation using the independent validation dataset.

4.2.3 Image Pre-Processing and Feature Extraction

The BraTS dataset, including MRI scans from 19 different institutions, was acquired using a variety of clinical protocols and scanners. To maintain consistency and quality, the images underwent a series of pre-processing steps. First, the MRI scans were converted from DICOM to NIfTI format to facilitate further processing. Next, N4 bias correction was applied temporarily to the scans as a preparatory step for registration [81] The MRI sequences, namely T1, T2, and FLAIR were then registered to the MRI T1ce sequence, and subsequently, the T1ce sequence was registered to the SRI24 anatomical atlas [223]. This pre-processing process provided co-registered, resampled volumes with uniform $1 \times 1 \times 1$ mm^3 isotropic voxel dimensions.[223]. This pre-processing process provided co-registered, resampled volumes with uniform $1 \times 1 \times 1$ mm^3 isotropic voxel dimensions.

To further refine the images with the skull-stripping step, a pre-trained DL model was performed for brain tissue extraction from all scans. The extracted images with variable intensity values were then normalised using intensity Z-scoring. All pre-processing steps were executed using CaPTk [74]. The resulting images had a fixed voxel resolution of $1 \times 1 \times 1$ mm^3 and a matrix size of $240 \times 240 \times 155$.

For the image pre-processing of the STORM_GLIO dataset, we employed a similar approach to that used in the curation of the BraTS 2020 dataset with consideration of our clinical settings. Specifically, our pre-processing pipeline involved two key steps: (1) skull stripping using the HD-BET algorithm [75], and (2) rigid registration of all sequences to align with the T1ce modality, a workflow that has been previously validated [53], [245]. Additionally, unlike the BraTS dataset, we did not perform registration to the SRI24 atlas, as it was not compatible with our clinical requirements.For the image pre-processing of the STORM_GLIO dataset, we employed a similar approach to that used in the curation of the BraTS 2020 dataset with consideration of our clinical settings. Specifically, our pre-processing pipeline involved two key steps: (1) skull stripping using the HD-BET algorithm [75], and (2) rigid registration of all sequences to align with the T1ce modality, a workflow that has been previously validated [53], [245]. Additionally, unlike the BraTS dataset, we

did not perform registration to the SRI24 atlas, as it was not compatible with our clinical requirements.

Following this, the MRI scans were uniformly resampled using B-splines to an isotropic voxel size of 1 × 1 × 1 mm^3. Before resampling, the size of the MRI scans was varied in size, as detailed in Table 4.1. Our image pre-processing pipeline and settings were guided by the recommendations of IBSI [66].

Table 4.1 Selection of relevant MRI acquisition parameters (average, standard deviation) for the scans included the STORM_GLIO dataset.

	T1	T1ce	T2	FLAIR
Thickness (mm)	4.77 +/-0.47	4.76 +/-0.47	4.74 +/-0.56	4.81 +/-0.39
TR (ms)	489 +/-96	494 +/-98	5627 +/-1856	8084 +/-1832
Echo Time (ms)	11 +/-2	11 +/-2	97 +/-8	112 +/-27
Inversion Time (ms)	0 +/-0	0 +/-0	0 +/-0	2217 +/-259
Field Strength (T)	1.54 +/-0.24	1.5 +/-0	1.54 +/-0.24	1.54 +/-0.24
Rows	426 +/-145	424 +/-146	546 +/-185	475 +/-219
Columns	417 +/-147	415 +/-148	527 +/-198	458 +/-232
Pixel spacing (mm)	0.62 +/-0.19	0.62 +/-0.19	0.48 +/-0.14	0.59 +/-0.21
Slice Spacing (mm)	5.99 +/-0.73	5.98 +/-0.74	6.27 +/-0.96	6.34 +/-0.72
SAR	1.09 +/-0.77	1.07 +/-0.76	0.91 +/-0.53	0.69 +/-0.67

In the BraTS 2020 challenge, three different tumour regions were defined: ET, TC, which encompasses both ET and NCR regions, and WT, including ET, NCR, and oedema regions. In contrast, the STORM_GLIO dataset included manually delineated GTV segmentation, defined as the visible extent of malignant growth [91]. Previous validations [53], [92], [245] have established that GTV and TC are analogous regions, indicating potential suitability for radiomic analysis. Therefore, both contours were treated as interchangeable for radiomic analysis.

Utilising the scans of the four MRI sequences associated with each patient, a total of 660 RFs (4 × 165) were extracted using the MATLAB version of SPAARC (https://www.spaarc-radiomics.io). These RFs are a broad range of numerical indicators that capture various aspects of tumour characteristics, including shape, texture, and intensity patterns. All features were standardised according to IBSI guidelines [66] and extracted using a 3D approach. The image pre-processing

settings and the names of the collected RFs are demonstrated in Appendix Figure C-1.

To assess the robustness of the RFs against variations in acquisition parameters and patient positioning, we employed image augmentation techniques similar to those used by Zwanenburg et al. [246]. In this research, we performed rotations (−4°, −2°, 0°, 2°, 4°) and volume changes (−20%, −10%, 10%, 20%) to the GTVs in the training cohort (detailed in Appendix C, Figure C- 5). A set of 20 variant images per patient was generated, which were used for feature stability analysis. To evaluate the consistency of each feature across these variations, we computed the intra-class correlation coefficient (ICC) with a 95% CI. Any feature with an ICC below 0.75 at the lower bound of the 95% CI was considered unstable and was removed from the feature set used in the model-building process. The same exclusion criteria were conducted for the features extracted from the validation cohort.

4.2.4 Identifying a Clinical and Radiomic Signature

For OS analysis, three feature selection methods were utilised, which enhanced the model's generalisability and mitigate overfitting issue. Our approach to identifying a clinical and radiomic signature involved a four-step process: (i) feature pre-processing, (ii) feature selection, (iii) hyperparameter optimisation for the ML models, and (iv) model building with internal validation. The detailed workflow of feature selection is demonstrated in Figure 4.2.

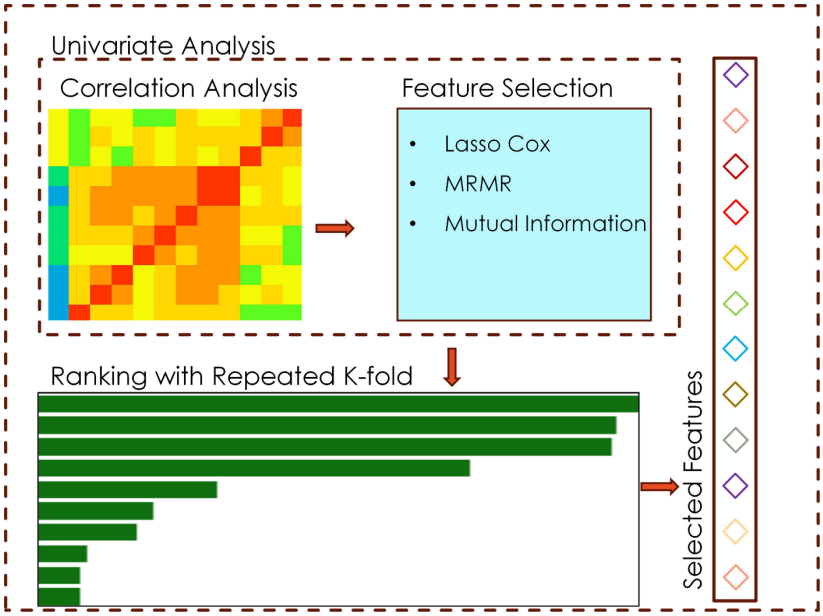


Figure 4.2 Feature Selection Workflow: Correlation analysis using Spearman and Pearson methods, feature selection through Lasso Cox, mRMR, and Mutual Information, ranking features over multiple iterations, and feeding the model with the selected features.

To ensure the reliability of our results, we used three-fold CV with 33 repetitions on the training data for all steps except (iv) model building, following the approach utilised by Kim et al. [247].

(i) In the feature pre-processing step, we applied the Yeo-Johnson transformation to align the feature distributions with a normal distribution [248]. Subsequently, features were z-score normalised. Both the transformation and normalisation processes were applied to the training dataset. The derived parameters during the process of the training dataset were used to normalise the features in the validation dataset. This ensured features exhibited consistent and similar distributional properties between the two datasets, an important requirement for robust model training and validation.

(ii) Building on the methodology of feature selection established by Leger et al. [230], the study conducted three distinct methods for feature selection: mutual information (MutInfo) [249], mRMR [250], and Lasso [229]. Upon completion of the feature selection process, three prognostic models were employed: Cox-LASSO, gradient boosting survival (GBS), and RSF [233]. These models are purpose-built for

time-to-event analysis, providing diverse analytical approaches that can potentially increase the accuracy and robustness of risk stratification.

(iii) To address overfitting, hyperparameter tuning was performed using bootstrap sampling of the training datasets for each model.

(iv) To comply with the radiomic guidelines and meet the minimum requirement of three features for radiomic analysis, including clinical information (age) [67], the two features collected from each of the 99 CV runs were counted and ranked based on their frequency of occurrence.

The prognostic models, built using three features, were evaluated on 200 bootstraps of the entire training dataset to assess their stratification performance using the C-index. This workflow was applied to construct prognostic models on the training dataset, and the developed models were then tested on the validation dataset.

4.2.5 Statistical Analysis

A comparative analysis of the survival distributions in the training and validation datasets was performed using the log-rank test. To evaluate whether significant differences existed in the distribution of categorical variables within the clinical data between the training and validation cohorts, the $\chi 2$ test was employed. Continuous variables, on the other hand, were assessed using the Mann-Whitney U test to determine if any notable differences were present. The prognostic models generated risk scores that were analysed using KM curve survival analysis. The median risk score was used as the threshold (cut-off) to categorise patients into high- and low-risk groups. The resulting KM curve was then assessed using the log-rank test to determine its statistical significance in stratifying the risk groups.

To further validate the prognostic models, the C-index was calculated to assess their risk stratification performance, which indicates how well the models can predict patient outcomes. To calculate the 95% CI for the C-index, 200 bootstraps were performed on both the training and validation cohorts [234]. In addition to the C-index, the integrated Area Under the Curve (iAUC) was calculated, offering a more

detailed understanding of their predictive abilities over time [251]. Unlike the conventional area under the curve (AUC), which provides a static evaluation of model performance, the iAUC provides a dynamic evaluation, capturing the models' performance as it changes over time. Moreover, the iAUC was calculated at 11 months for all models, as specified by our clinicians, to provide a more detailed understanding of the model's performance at this critical time point.

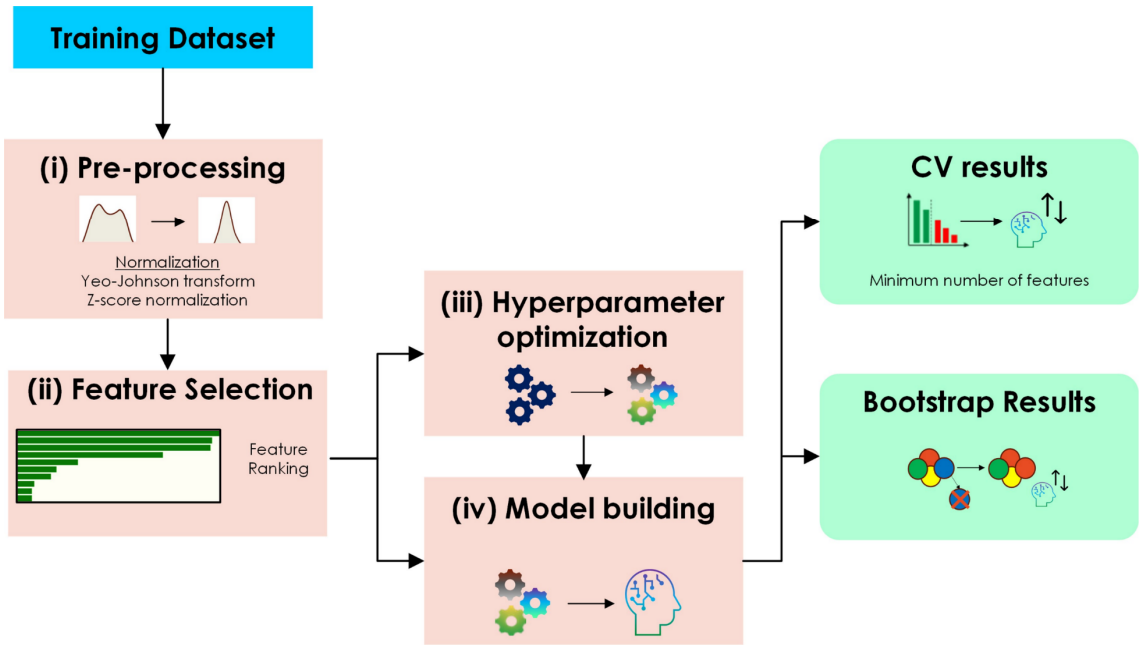


Figure 4.3 Overview of the framework used for feature selection and hyperparameter optimisation.

All statistical and survival analyses were conducted utilising Python software version 3.9. A p-value of less than 0.05 was considered statistically significant, indicating that the observed differences were unlikely to occur by chance. The image preprocessing and statistical analysis workflow are illustrated in Figure 4.3, providing a visual representation of the steps involved in the analysis. Finally, permutation feature importance was calculated using the scikit-learn library (version 1.3.2) to determine the relative importance of each feature in the models.

4.3 Results

The clinical attributes of the training and validation cohorts are presented in Table 4.2, which highlights the median OS of 11.9 months and 12.3 months for the

respective cohorts. Notably, the OS data between the two cohorts did not exhibit a statistically significant difference (p = 0.48, Table 4.2). Following a robustness analysis, 523 stable RFs remained out of the initial 660. Each robust radiomic feature for all four MRI sequences was listed in Appendix (Figure C- 2 and Figure C-3). Appendix Figure C- 4 illustrates the robustness of feature families as a percentage across all MRI scans in total.

Table 4.2 Characteristics of clinical variables for training and validation datasets.

Variable	Training dataset Median (range)	Validation dataset Median (range)	Statistical Cohort Comparison
Age (years)	61.1 [18.98–86.27]	63.4 [31.0–86.65]	U: 0.65, p-value: 0.74
OS (months)	11.9 [0.17–58.9]	12.3 [0.7666–57.7]	U: 0.63, p-value: 0.48
OS < 11-month (%)	43.7% (101/231)	39.7% (23/58)	χ2: 0.19, p-value: 0.66

Additionally, all RFs demonstrated a weak correlation with age, as evidenced by correlation coefficients below 0.3 (Spearman < 0.3). After excluding RFs with high correlation coefficients (Spearman > 0.95), a total of 227 RFs remained. These robust RFs were then utilised to perform feature selection using a three-fold CV setting with 33 repetitions, resulting in a total of 99 runs. Subsequently, a pool of 37 RFs was collected through LASSO feature selection. The top two RFs were selected from this feature set due to their high frequency of occurrence. To further refine the prognostic model, 200 bootstrapping iterations were applied to the entire training cohort to select the hyperparameters for each three-feature model, which included the top two RFs and age. Details of the selected hyperparameters and predefined settings can be found in Appendix Table C- 1. The radiomic model in the training cohort yielded optimal results using only two RFs: morph_av (morphological, occurrence: 31%) and dzm_zdnu_3D (texture, occurrence: 16%). These two RFs, derived from the FLAIR modality, exhibited a weak correlation with each other (Spearman < 0.6), indicating their complementary nature. The model demonstrated a C-index of 0.60 (95% CI: 0.54–0.66) and a HR of 2.72 (95% CI: 1.66–4.46), suggesting its potential for predicting patient outcomes.

The comprehensive analysis and rigorous feature selection process employed in this study aimed to identify the most informative RFs and optimal hyperparameters for developing a robust prognostic model. By leveraging the strengths of Lasso-Cox

feature selection and bootstrapping, the study sought to minimise the impact of overfitting and ensure the generalisability of the model to unseen data. A clinical-radiomic signature that integrated age, and two RFs was developed to develop prognostic models for the training cohort. The top two RFs, identified through feature selection methods, are presented in Table 4.3.

Table 4.3 The selected feature names are shown for each feature selection method. Each feature is displayed with its dependent modality in parentheses, except for "morph_av," which is a modality-independent feature.

Feature Selection Method		
Lasso	MutInfo	mRMR
morph_av dzm_zdnu_3D (FLAIR sequence)	szm_glnu_3D (T1ce sequence) stat_p10 (T2 sequence)	dzm_zdnu_3D (FLAIR sequence) szm_glnu_3D (T1ce sequence)

Additionally, the Lasso feature selection method was the most successful, producing optimal RFs that necessitate the minimum number of MRI sequences. Furthermore, the Cox-LASSO model, recognised for the highest interpretability among the ML models by Luo et al. [235], demonstrated a C-index of 0.64 for internal validation, as illustrated in Figure 4.4c. Morph_av (IBSI: 2PR5) is a shape-based feature that provides a surface-to-volume ratio, offering insights into tumour morphology. Dzm_zdnu_3D (IBSI: V294), on the other hand, is a texture feature that quantifies the association between spatial location and grey level value by measuring the size of homogeneous zones (groups) within a specified distance. This feature captures the distribution of such zone counts across various distances, providing a comprehensive understanding of tumour texture. The feature is derived from GLDZM.

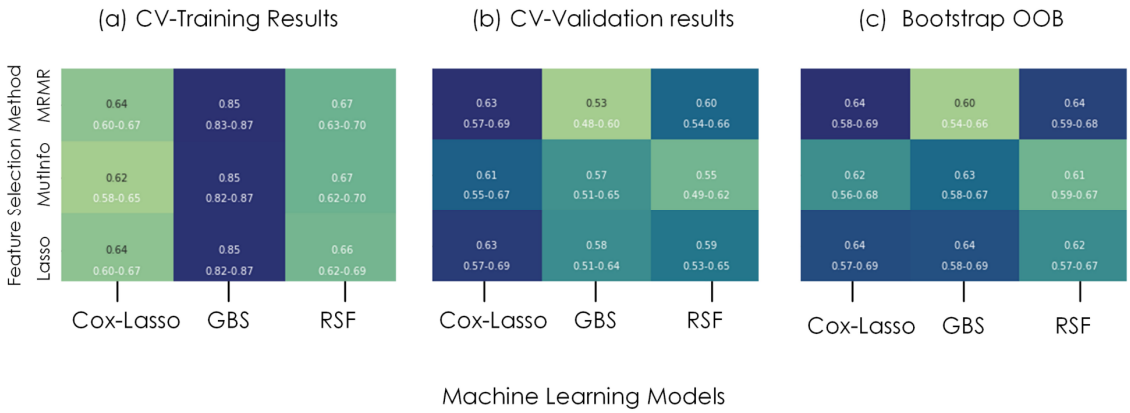


Figure 4.4 C-index of models based on each feature selection method and each corresponding ML algorithm for the prognosis of GBM. (a) CV-Training results. (b) CV-Validation results. (c) Bootstrap OOB.

The pairing of morph_av and dzm_zdnu_3D demonstrates the importance of considering both morphological and textural characteristics for accurate prognosis. The feature importance of each feature in the ML models was demonstrated in Table 4.4.

Table 4.4 Permutation feature importance: Permutation feature importance was conducted test for 200 repetitions

Feature Selection Method	Feature Names	ML Model (Cox-LASSO)
Lasso	morph_av	0.001
	dzm_zdnu_3D (FLAIR)	0.07
	Age	0.06
mRMR	dzm_zdnu_3D (FLAIR)	0.03
	szm_glnu_3D (T1ce)	0.02
	Age	0.06
Mutational Information	stat_mean (T1)	0.01
	cm_joint_entr_3D_comb (FLAIR)	0.02
	Age	0.09

The radiomic model achieved the highest C-index of 0.62 (95% CI: 0.54–0.71) and a HR of 2.97 (95% CI: 0.8–10.99) in the validation dataset, as detailed in Table 4.5. On the other hand, the combined clinical–radiomic model, utilising a clinical feature alongside RFs, achieved the top C-index of 0.63 (95% CI: 0.56–0.74) within the training dataset, as demonstrated in Table 4.6. This model had a C-index of 0.69 (95% CI: 0.62–0.75) in the validation dataset.

Table 4.5 Univariate Cox regression analysis.

Univariate Cox Regression Analysis							
Dataset	Model	Variable	HR [95% CI]	p-value	C-index	iAUC	11m-iAUC
Training	Clinical model	Age	1.32[1.15–1.50]	0.010	0.59 [0.53–0.64]	0.67	0.62
	Radiomic model	RFs Risk Score	2.72[1.66–4.46]	0.007	0.60 [0.54–0.66]	0.67	0.63
Validation	Clinical Model	Age	1.63 [1.23–2.16]	0.006	0.63 [0.56–0.68]	0.66	0.67
	Radiomic model	RFs Risk Score	2.97 [0.8–10.99]	0.290	0.62 [0.54–0.71]	0.79	0.78

Table 4.6 Multivariate Cox regression analysis.

Multivariate Cox Regression Analysis							
Dataset	Model	Variable	HR [95% CI]	p-value	C-index	iAUC	11m-iAUC
Training	Clinical–radiomic Model	Age	1.30 [1.14–1.49]	6×10^{-5}	0.63 [0.56–0.74]	0.68	0.69
		morph_av	1.02 [0.87–1.20]				
		dzm_zdnu_3D	1.36 [1.13–1.62]				
Validation	Clinical-radiomic Model	Age	1.60 [1.21–2.13]	7×10^{-5}	0.69 [0.62–0.75]	0.78	0.81
		morph_av	1.58 [1.08–2.29]				
		dzm_zdnu_3D	1.89 [1.19–3.01]				

Table 4.7 Feature weights and cut-off value. The weight of each feature and the cut-off value for risk-stratification into low and high-risk groups.

		Weights
Feature Name	morph_av	-0.107019
	dzm_zdnu_3D (FLAIR)	0.208010
	Age	0.340629
Cut-off Value	Median value of risk scores	0.015

For the KM curve, the cut-off value was set at 0.015 and feature weights were calculated, as shown in Table 4.7. In the training dataset, the log-rank p-value was 6×10^{-5}. For the same cut-off value, the log-rank p-value was 7×10^{-5} in the validation dataset (refer to Figure 4.5a, b). The KM plots effectively highlight the reliable capability of the model to separate between high- and low-risk groups across both datasets. The clear separation of survival curves, along with the highly significant p-values, highlights the model's capability and potential predictive strength for diverse, previously unseen patient populations. This robust predictive performance indicates that the model might be an effective tool for adjusting prognoses and developing personalised treatment strategies based on individual risk profiles. On the other hand, differences between the training and validation KM curves highlighted the need for future studies to incorporate larger unseen cohorts and additional clinical variables. Increasing both sample size and clinical variables would enable broader patient characteristics to be captured, thereby improving model robustness and generalisability.

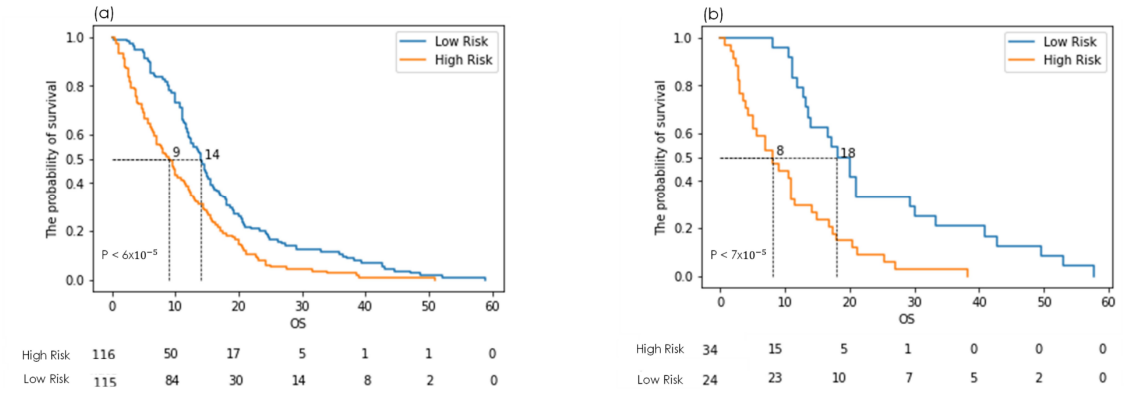


Figure 4.5 Kaplan–Meier plots showing differences between a) training and b) validation datasets stratified into high or low-risk groups by the Cox–Lasso model. The small p-values indicate a highly reliable differentiation between the risk groups.

At 11 months, the iAUC of the prognostic model using only two RFs was 0.63 for the training dataset and 0.78 for the validation dataset. The iAUC of the model with just the age information reached 0.62 in the training dataset and 0.67 in the validation dataset. The clinical–radiomic model, which combined age with two RFs, achieved an iAUC of 0.69 in the training dataset and 0.81 in the validation dataset. As presented in Table 4.6, the HR highlights the most significant influence from the GLDZM-based feature, with a value of 1.89 in the validation dataset. Both age and morphology features show nearly equivalent effects, with values of 1.60 and 1.58, respectively. Figure 4.6 offers a visual representation of the risk groups using example cases, including MRI FLAIR images and a 3D tumour mesh of the relevant patient.

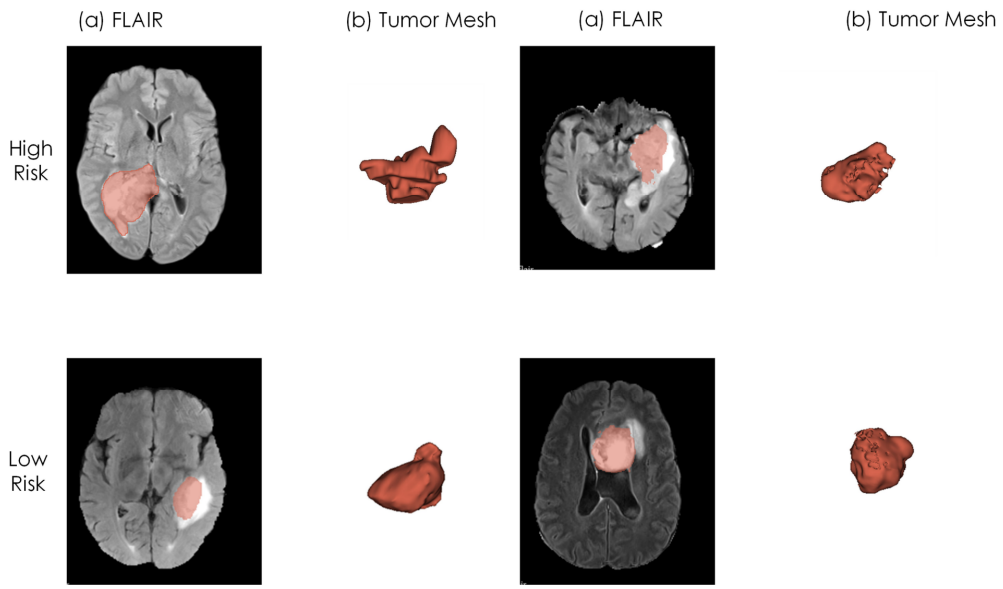


Figure 4.6 The visualisation of risk groups (first row: high risk, second row: low risk) for median OS values of each group. For each case, a transverse slice from the FLAIR scan (a) is coupled with a 3D mesh of the tumour. (a) FLAIR (b) Tumour Mesh.

4.4 Discussion

This study presents the development of a clinical-radiomic prognostic model aimed at stratifying GBM patients into high- and low-risk categories utilising preoperative MRI scans. There is important factor that the transition from the RTSTRUCT format to a mask can impact radiomic analysis when employing various software platforms [252]. To ensure consistency, we utilised a single software platform (Python) for generating masks within the STORM_GLIO dataset. Through robustness analysis of RFs, feature selection methods identified two RFs derived exclusively from the MRI FLAIR sequence. The clinical-radiomic model demonstrated a C-index of 0.69 upon validation, with significant differences detected between the stratified risk groups.

A comprehensive evaluation of our study's findings with existing research is provided in Table 4.8, where we applied rigorous inclusion criteria to select studies that solely employ RFs, concentrate on GBM (Grade 4), and explore time-to-event (overall survival) outcomes. Studies not adhering to these criteria were excluded. The table highlights potential biases, particularly concerning patient sample size, such as limited cohorts and single-centre studies, which might affect the validity and reliability of results. For example, although Hajianfar et al. [242] reported the

highest C-index, their study involved the smallest patient cohort. To address this, we endeavoured to maximise our patient cohort across multiple centres; however, potential biases persist in our study. Notably, Cepeda et al. [240] developed a model using multiple MRI sequences and 10 RFs, achieving a C-index of 0.61 and an iAUC of 0.77. Similarly, Tixier et al. [239] reported an AUC of 0.75 with 57 RFs. Verma et al. reported similar results (AUC = 0.78) using a more extensive set of over 300 RFs, derived from multiple MRI modalities [241]. Furthermore, Hajianfar et al. reported a C-index of 0.77 [242] utilising convolutional filters not standardised by IBSI at the time of publication [236]. Our study demonstrated a comparable C-index and achieved the highest iAUC at 11 months. We achieved this result by leveraging the largest patient cohort, a minimal set of RFs, RFs derived exclusively from the MRI FLAIR modality, and a single ROI (GTV). Through our experiments with various feature selection methods and ML algorithm combinations, we found that incorporating age, along with two RFs, a modality-independent morphology feature (morph_av) and a GLDZM feature (dzm_zdnu_3D) from the MRI FLAIR modality, yielded the most desirable performance in terms of generalisability on the validation set. This approach enhances the model's performance across diverse healthcare settings, from small local clinics to large research hospitals, offering reliable predictions and valuable insights from various data sources.

Table 4.8 The comparison of recent similar studies with our study.

References	No. of patients	MRI Sequence	Region of Feature Extraction	Extracted Feature number	Selected Feature Number	Feature number guideline (3–10)	ML model	Validation method	IBSI guideline	Performance metrics
Tixier et al. [239]	234	T1	Gd-ET, NEC, NET, TC	88	57	No	Lasso	Five-fold CV	Yes	AUC: 0.75
Cepeda et al. [240]	203	T1ce, T1, T2, FLAIR	Tumour, Peritumoural	15,720	10	Yes	Random Forest Survival	Five-fold CV	Partially (Convolutional Filters)	iAUC: 0.77 C-index 0.61
Verma et al. [241]	150	T1ce, T2, FLAIR	ET, NCR	3792	316	No	-	Five-fold CV	Partially (Convolutional Filters)	AUC: 0.78
Hajianfar et al. [242]	119	FLAIR, T1ce	ET, TC, NEC, ED	4471	-	No	Cox Boost	Three-fold CV Bootstrap	Partially (Convolutional Filters)	C-index: 0.77
Our Study	289	FLAIR	GTV (TC)	689	2 (without Age)	Yes	Cox-LASSO	Three-fold CV 33 repetitions Bootstrap	Yes	C-index: 0.69 iAUC: 0.81

As illustrated in Figure 4.6, high-risk patients are characterised by irregular boundaries and a non-smooth, irregular shape, indicated by a higher surface area to volume ratio (morph_av). Additionally, ZDNU from GLDZM, which measures zone size and distance variability in a 3D image (dzm_zdnu_3D), is elevated in high-risk patients, reflecting a greater degree of heterogeneity in textural patterns (in Figure 4.6a). This suggests that even seemingly homogeneous regions can exhibit significant zone size variations at different distances. Combining RFs with age resulted in improved outcomes compared to using clinical information alone for GBM, as also demonstrated by Cepeda et al. [240]. While integrating clinical (age) and radiomic data can enhance model performance, it may cover the importance of clinical factors. Without a rigorous feature selection and model-building approach, models risk overfitting the training data and struggling with new datasets with poor prognostic prediction, underscoring the importance of integrating clinical and radiomic features in a balanced approach. In the medical field, challenges such as data sparsity, scarcity, and imbalance arise due to the limited availability of data on rare diseases, small patient cohorts, and missing clinical information, hindering the collection of comprehensive datasets [189]. In our research, we experienced similar difficulties in gathering comprehensive data, including a variety of MRI sequences (T1, T1ce, T2, and FLAIR) and a range of clinical parameters such as age, genetic information, survival metrics, and Karnofsky performance status. Acknowledging these clinical limitations, we maximised the patient cohort by collecting limited clinical data and extracting radiomics features from a minimal yet informative set of MR sequences. By adopting this strategy, we aimed to align the trade-offs between data availability and model performance. Moreover, the risk of bias associated with the retrospective dataset was mitigated by assembling a multi-centre patient cohort with the largest feasible sample size.

The enhancement of our model's performance could be achieved by incorporating additional labels beyond the GTV, such as the multiple regions of feature extraction utilised in prior research, as illustrated in Table 4.8. Additionally, employing DL-based features has the potential to improve outcomes in survival analysis. However,

this study deliberately excluded deep RFs due to their limited reproducibility and interpretability, which pose significant constraints for clinical applications [237]. Previous research did not emphasise essential criteria, such as employing a singular ROI or enhancing interpretability by minimising the number of RFs. In contrast, our approach aligns with the recommendations of van Timmeren et al. [67], which suggest for limiting the number of features in radiomic model construction to a range between 3 and 10. To address the issue of overfitting, we developed a workflow incorporating hyperparameter optimisation and data resampling. The prognostic model's results were reported on an independent validation dataset using this workflow. However, larger unseen cohorts and additional clinical variables are needed to enhance the generalisability and reliability of radiomic models.

4.5 Conclusions

This research presents the development and validation of a clinical-radiomic model for the stratification of GBM patients based on OS. Notably, this is the first study to employ MRI-based RFs in accordance with IBSI guidelines, while addressing crucial clinical challenges, interpretability, and robustness analysis in GBM contexts. Our approach demonstrates superior performance compared to previous studies, such as that by Tabassum et al. [243]. The model incorporates two independent RFs derived from the FLAIR MRI sequence alongside patient age. With the recent standardisation of convolutional filters under the IBSI guidelines, future work will explore their application to RFs. Additionally, future research directions include to leverage DL features to improve model performance, with an emphasis on ensuring their interpretability. This may involve the use of multimodal foundation models, integrating further clinical parameters such as age, sex, and Karnofsky performance status, or incorporating multi-modality imaging data (e.g., PET, CT). Further avenues for performance improvement cover the integration of diffusional or functional MRI sequences and the acquisition of more comprehensive clinical datasets, including omics data such as genomics, transcriptomics, and metabolomics. These enhancements, supported by larger patient cohorts, are anticipated to yield more accurate and reliable models.

5. Region Focused Selection+: A Clinically Adaptable Strategy for Brain Tumour Segmentation

5.1 Introduction

In this chapter, we explored the challenges associated with deploying state-of-the-art DL-based automated segmentation models in diverse clinical settings. Although promising results have been achieved with these models on standardised datasets such as BraTS datasets, their performance may be degraded when applied to data acquired with varying imaging parameters. This study examined an integrated approach combining refined ROI selection, alternative normalisation methods, and resource allocation strategies to enhance segmentation flexibility and clinical applicability.

Enhancing the adaptability of DL models for brain tumour segmentation across varied clinical scenarios is the primary aim of this study while tackling the practical challenges of time and memory limitations inherent in real-world applications. To overcome these challenges, we introduced RFS+, which represents a significant departure from the region-focused selection (RFS) [92]. RFS+ moves beyond the limitations of RFS, which relies on Z-score normalisation for TC/GTV and uses tumour regions and labels as input masks, by introducing a broader framework that incorporates multi-class, multi-label, and binary class segmentation approaches, avoiding reliance on a single strategy. We employed RFS+ to train a U-Net [88] model on the BraTS training dataset, with the aim of improving the robustness and generalisability of DL models. Ensemble learning is utilised by identifying the top three models from the training dataset for each tumour region (ET, TC, and WT) to produce a unified segmentation result. Z-score intensity normalisation has been the predominant preprocessing technique in the majority of brain tumour segmentation studies to date [253]. Our research takes a broader approach by investigating several intensity normalisation methods and assessing their influence on segmentation performance using DSC metric.

Related Works

In the domain of medical imaging, DL has gained notable advancements and widespread application. This trend is evident in the BraTS challenge, where DL approaches are prevalent, and CNNs constitute the core of the top-performing models [86]. The first-place achievement in the BraTS 2018 competition underscores the significance of encoder-decoder architectures in brain tumour segmentation, with the winning model featuring an asymmetric encoder-decoder structure [180]. Among the architectural paradigms in the BraTS competition, U-Net-based architectures, leveraging encoder-decoder pathways, have maintained supremacy in recent years. For instance, the 2019 first-place model featured a two-stage cascaded U-Net architecture, securing the top spot on the leaderboard [181]. The leading model in 2020 was a 3D U-Net, referred to as nnU-net, which functioned as a self-configuring framework and excelled without requiring substantial alterations. The nnU-net framework continued its success in 2021, achieving first place with an expanded U-Net architecture. Furthermore, the 2021 winning model incorporated innovative methodological improvements to deliver superior results [152].

In the work of Magadza et al. [253], CNN architectures are systematically grouped into four subcategories: single pathway, dual pathway, cascaded architectures, and U-Net architectures. According to Pereira et al. [90], a single-pathway architecture is defined by its simplicity, utilising small kernels in their layers and maintaining a single processing path throughout the network. In dual-pathway architectures, two distinct processing paths are utilised within the same network, allowing for the concurrent extraction of global contextual features (e.g., anatomical brain location) and detailed local visual features [254]. Among the different types of cascaded architectures, the input cascade approach is most commonly implemented. This approach involves using the output of one CNN as the input to another, effectively adding an extra image channel for subsequent stages of architectures [255]. Another significant strategy within cascaded architectures is hierarchical segmentation, in which a multi-class segmentation task is decomposed into a sequence of binary segmentation stages. By utilising the hierarchical structure of tumour sub-regions, this design addresses the issue of class imbalance present in the dataset. This

approach generally starts with the segmentation of WT, followed by the generation of a bounding box based on the WT segmentation to inform the next stage. In subsequent stages, TC and ET regions are segmented in sequence. Although this approach results in longer training and inference times, its effectiveness has been demonstrated in successful applications, such as those reported by Wang et al. [179]. Models utilising a binary segmentation approach, benefiting from enhanced memory efficiency, have achieved remarkable outcomes [177]. U-Net-based [88] architectures have proven highly effective, as demonstrated by their adoption in the top-performing models of the 2020 and 2021 BraTS challenges [152], [182]. By employing a multi-label segmentation strategy, these models allowed for overlapping class representations, avoiding the need to treat each class as entirely separate. This study provides a comparison of these approaches, highlighting their respective mask representations.

Despite their potential, both feature extraction methodologies and transformer-based architectures in clinical integrations remain in early developmental stages [168], [183], having yet to surpass established performance benchmarks in BraTS competitions over the past three years [152], [182], [256]. Although a wide range of U-Net variants has been developed, this study contributes a thorough assessment of how ensemble learning methodologies influence fundamental U-Net structures. The architectural composition of DeepMedic [257] features a multi-scale 3D CNN design, contrasting markedly with the RFS+ framework's alternative structural approach. These architectural variations may have a substantial impact on the models' performance in brain tumour segmentation. With regard to balancing computational resources and model effectiveness, the Cascade U-Net [181] implements a hierarchical framework to enhance segmentation accuracy. By comparison, RFS+ emphasises computational efficiency, a feature particularly salient for clinical implementation. The framework's distinct capacity to accommodate various clinical scenarios distinguishes it from the Cascade U-Net model, which demonstrates more constrained adaptability.

Through the implementation of advanced optimisation strategies, RFS+ achieves superior learning efficiency compared to the deep supervision mechanisms characteristic of 3D Deeply Supervised Networks (3D-DSN) [257]. Moreover, RFS+ manifests exceptional adaptability to multiple MRI protocols and clinical environments, suggesting substantial benefits over 3D-DSN in heterogeneous medical contexts. This research utilised an array of architectures including 2D, 2.5D [245], 3D U-Net, and nnU-net [182], with specific attention to nnU-net based on its established clinical reliability. The consistent dominance of nnU-net and related U-Net configurations in recent BraTS competitions [152], [182], [256] illustrates their clinical performance supremacy relative to alternative approaches such as DeepMedic [168]. While the predecessor RFS framework [92] operated within the constraints of multi-class, multi-label, and binary segmentation methodologies for TC and GTV, RFS+ extends this capability by incorporating specialised normalisation strategies optimised for each target region. Through the implementation of this holistic methodology, the study seeks to advance the adaptability, computational efficiency, and clinical applicability of segmentation models for brain tumour analysis.

5.2 Material and Methods

The methodological framework encompasses architectural specifications, pre-processing protocols, and model hyperparameter optimisation procedures. Additionally, it characterises the two distinct datasets utilised: the BraTS 2021 collection (including training and validation subsets) and the institution-specific STORM_GLIO collection.

5.2.1 The Proposed Strategy: RFS+

The workflow for RFS, depicted in Figure 5.1 a), utilises Z-score normalisation as its singular normalisation technique, encompassing three distinct segmentation strategies optimised specifically for TC and GTV analysis.

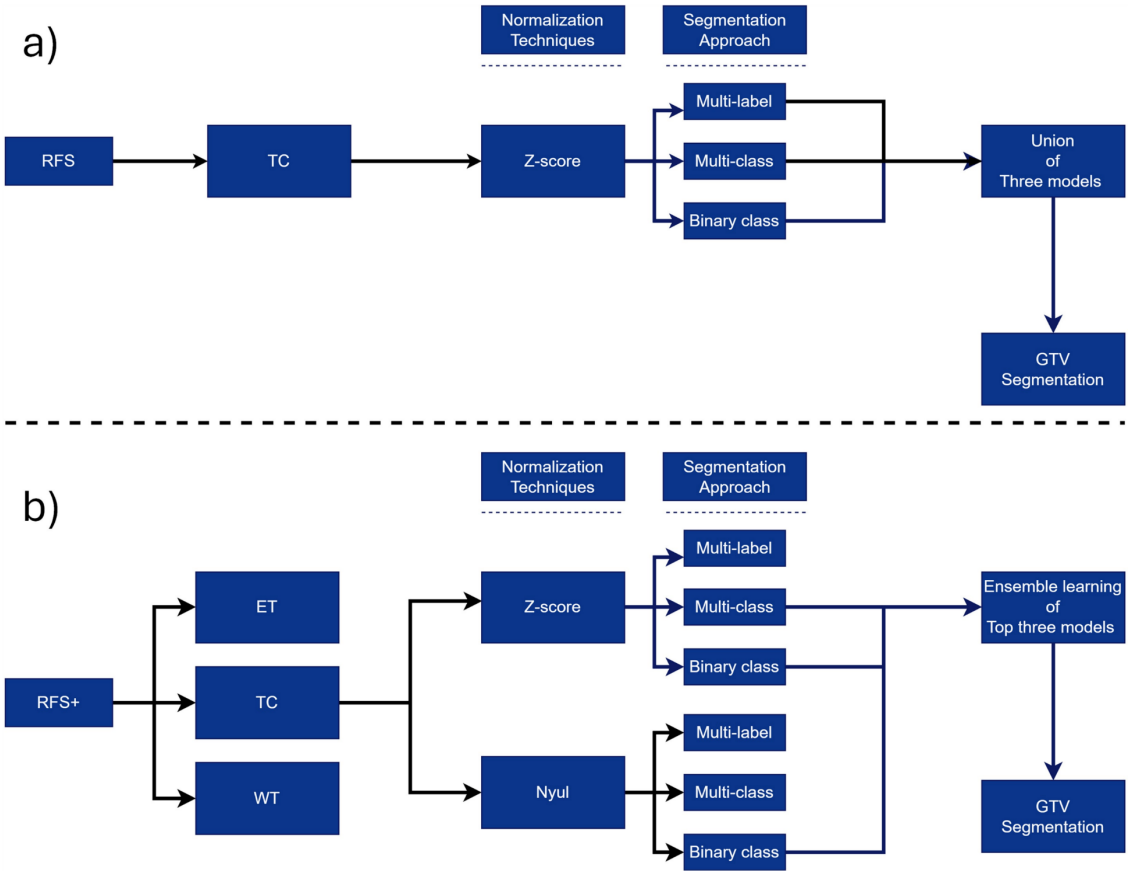


Figure 5.1 The proposed strategies: (a) RFS strategy and (b) RFS+ strategy on GTV segmentation.

Overlapping regions from the segmentation models are integrated into a unified structural representation. The enhanced methodology, RFS+, illustrated in Figure 5.1 b), provides an adaptable framework for implementing various DL models in brain tumour segmentation. This refined approach addresses multiple tumour regions, including ET, TC, and WT with comprehensive region-specific protocols detailed in Appendix (Figure D- 1, Figure D- 2, Figure D- 3). RFS+ distinguishes itself from its predecessor through two key advancements: the integration of ensemble learning methodologies and the implementation of region-adapted normalisation techniques. The framework's architecture is anchored by two core components:

1. Normalisation Techniques: Multiple normalisation approaches are applied as part of the pre-processing of MRI data.

2. Segmentation Approaches: Three unique segmentation methods are used, pairing normalisation techniques (e.g., Z-score) with segmentation strategies (e.g., multi-class segmentation) to target defined regions

The specific segmentation target (TC or GTV) is determined through the integration of selected normalisation techniques and segmentation approaches. Earlier investigations [92] have established the viability of transferring models trained on TC contours to GTV segmentation tasks. The selection of optimal models was identified on DSC metrics, evaluating performance on the training dataset (with 15% of the data reserved as unseen) for TC/GTV segmentation. The superior performing architectures comprise:

- Multi-class segmentation incorporating Z-score normalisation
- Binary-class segmentation utilising Z-score normalisation
- Binary-class segmentation employing Nyul normalisation

The outputs from these three models were fused by utilising ensemble learning techniques. A detailed schematic representation of each segmentation strategy, its requisite inputs, and the encompassing RFS+ architecture is presented in Figure 5.2. The main figures prioritise the visualisation of TC/GTV; in contrast, Appendix includes detailed, high-resolution illustrations of individual segmentation workflows (Figure D- 4, Figure D- 5, Figure D- 6).

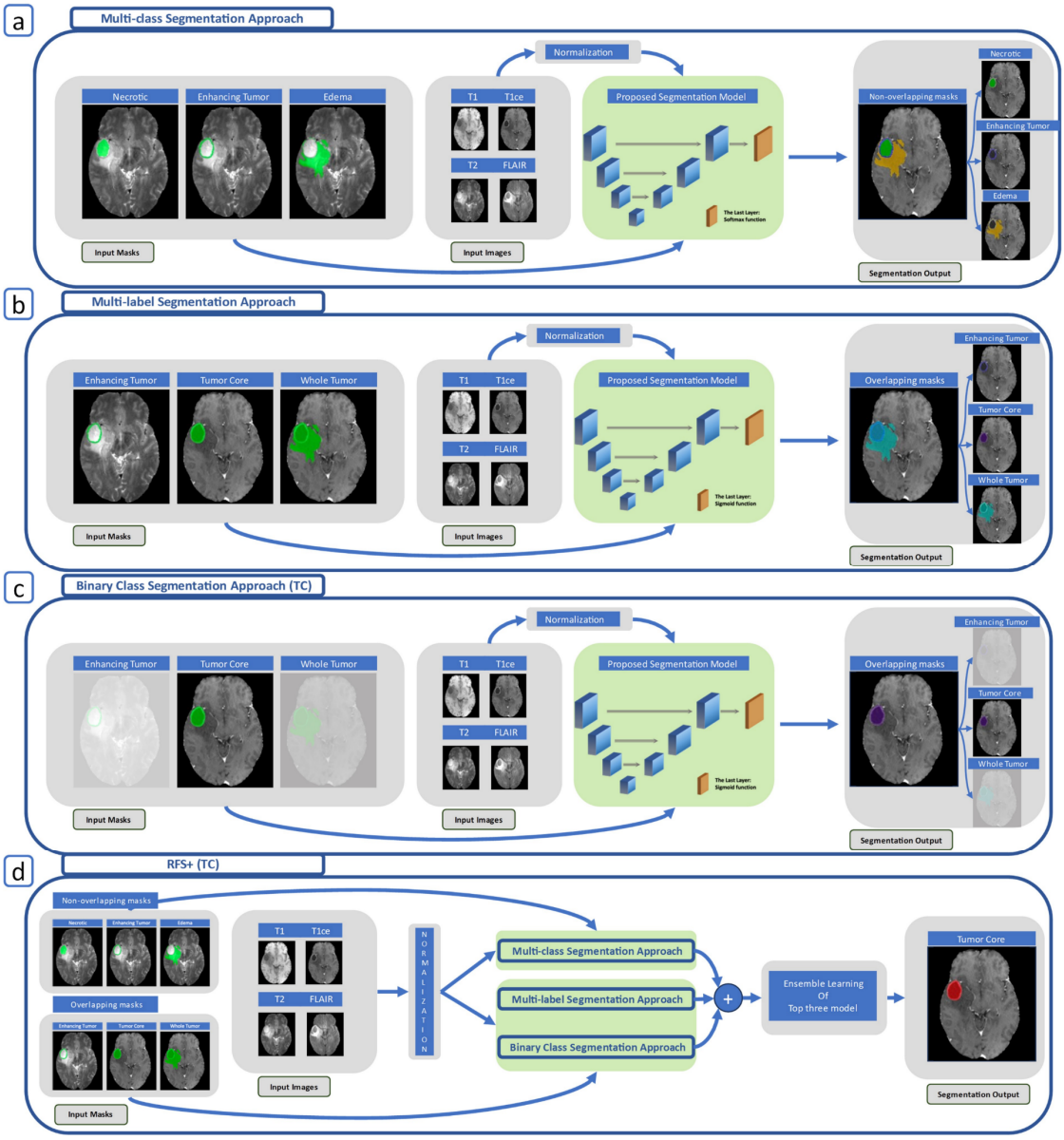


Figure 5.2 Different Segmentation Approaches: (a) Multi-class segmentation, (b) multi-label segmentation, (c) binary class segmentation, and (d) RFS+ for TC/GTV segmentation.

5.2.2 Normalisation of MRI Scans

To address MRI scanner-dependent intensity variation, the implementation of dual normalisation strategies was employed: Z-score normalisation and piecewise linear histogram matching (Nyul). These specific approaches were extracted from a more comprehensive array of normalisation techniques as detailed by Reinhold et al. [258], having demonstrated superior efficacy in DL applications. Appendix (Table D- 1) contains comparative analyses of alternative normalisation methodologies.

5.2.3 Network Architectures

5.2.3.1 Segmentation Approaches

The segmentation methodology incorporates three different approaches: multi-class, multi-label, and binary class segmentation. The architectural distinction lies in mask configuration, with non-overlapping masks applied in multi-class segmentation, while overlapping masks were implemented in both binary class and multi-label approaches, as shown in Figure 5.3.

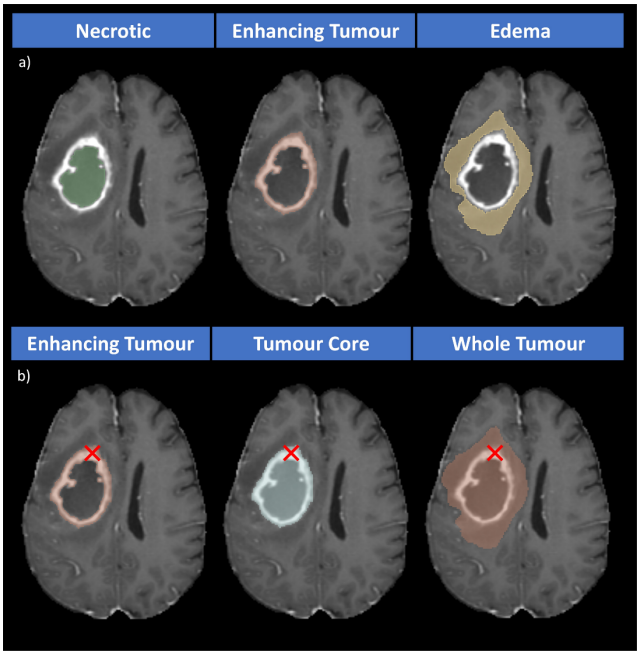


Figure 5.3 Different masks: (a) The non-overlapping masks (the input for the multi-class approach) and (b) the overlapping masks (the input for the binary class and the multi-label approaches; the red cross shows an example of overlapping pixels for ET tissue).

The implementation encompassed three U-net architectural variants: 2D, 2.5D, and 3D configurations. While Figure 5.4 delineates the 2D U-net configuration, the 3D U-net implementation aligns with the architectural specifications outlined by Çiçek et al. [107].

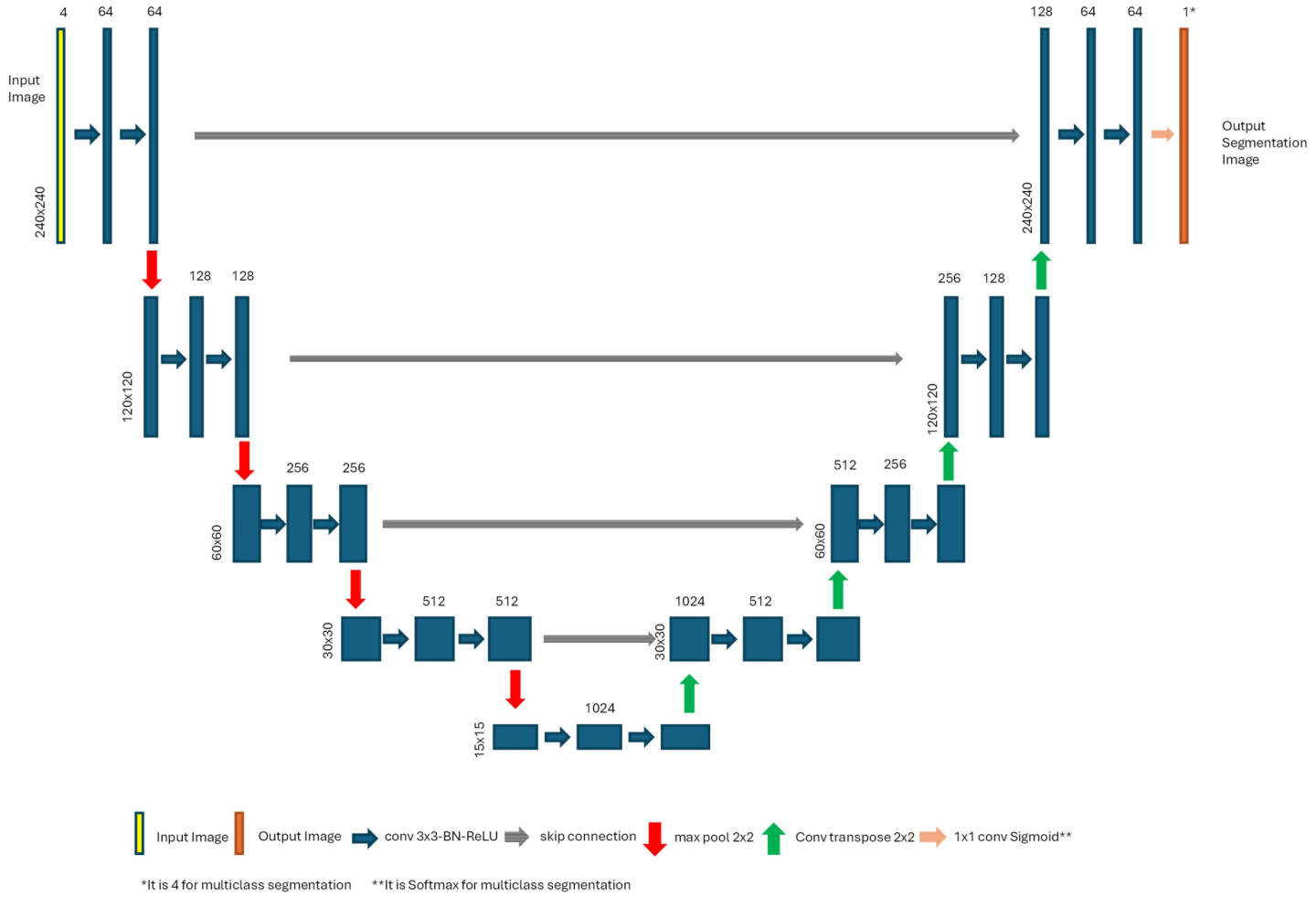


Figure 5.4 The proposed 2D UNET model.

The 2.5D U-net architecture [245] expands upon the 2D U-net model through multi-channel input processing. The design concatenates triplets of consecutive slices (prior, present, and subsequent) from the imaging volume, enabling effective utilisation of all MRI modality inputs. Figure 5.5 demonstrates the implementation of this three-channel approach across modalities.

The 2D U-net architecture was adapted into three variants, with modifications solely in the final layer's activation function. The implementation strategy employed:

- Sigmoid activation for binary class segmentation tasks
- Sigmoid activation for multi-label segmentation scenarios
- Softmax activation for multi-class segmentation applications

These functional choices were specifically tailored to optimise performance for each segmentation paradigm and its corresponding mask structure.

Input dimensionality was structured as follows:

The 2D U-net processed inputs of size 240 × 240 × 4, where the four channels represented aligned T1, T1ce, T2, and FLAIR modalities. In contrast, the 2.5D U-net expanded the channel dimension to 240 × 240 × 12, incorporating triplets of adjacent slices for each modality. Both implementations maintained identical spatial dimensions of 240 × 240 pixels. Four sequential blocks form the encoder path of the U-net models, with each block employing two convolutional layers followed by batch normalisation to normalise feature maps and enhance training stability. Each convolution is followed by a ReLU activation function. The feature maps undergo progressive downsampling through max-pooling operations, with the number of channels expanding from 64 at the input, doubling after each pooling stage (128, 256, 512), ultimately reaching 1024 channels at the bottleneck layer. The decoder architecture maintains symmetry with the encoder through a series of blocks, each implementing a convolutional transpose layer for upsampling, followed by two convolutional operations.

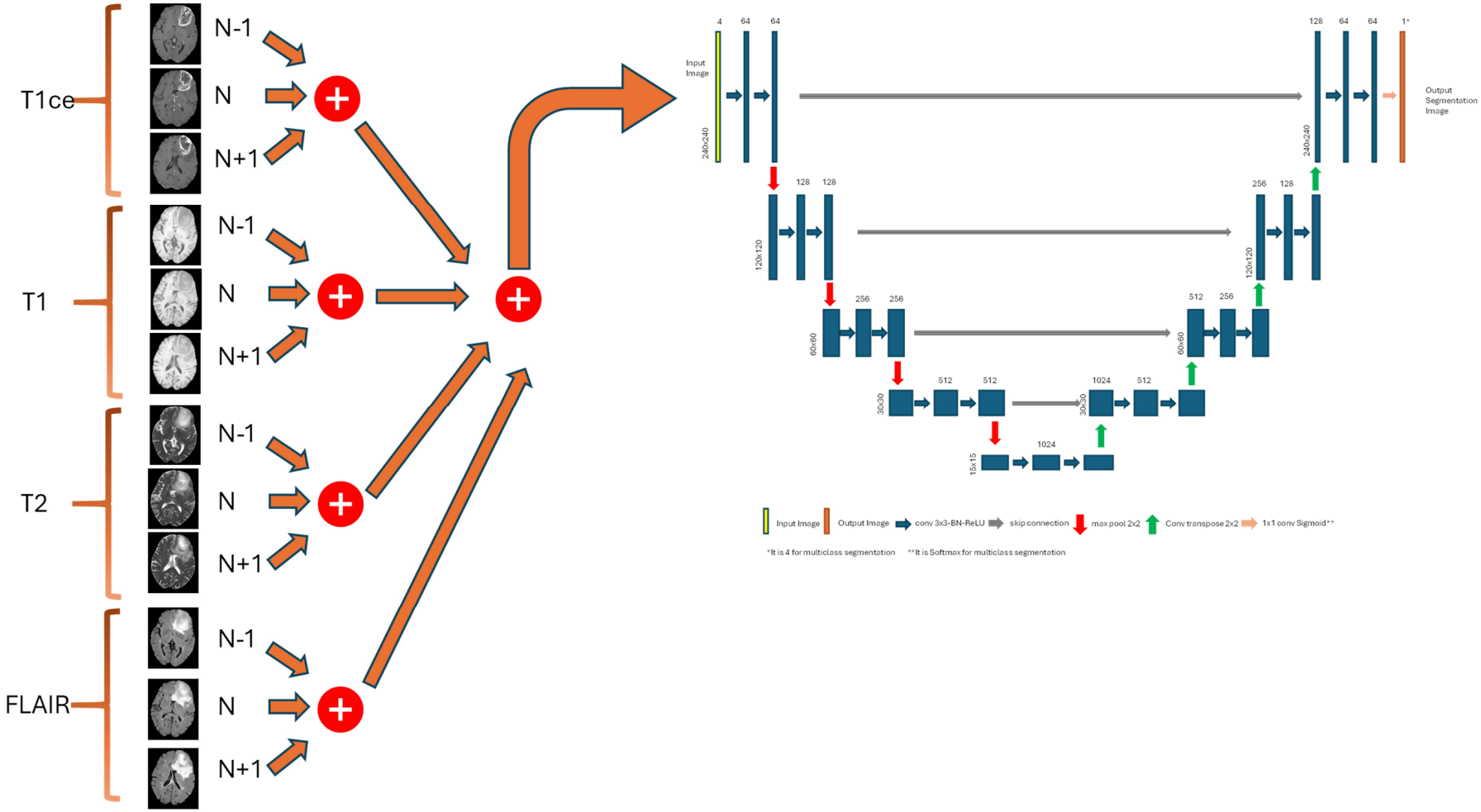


Figure 5.5 Three Channel Method: An example of each modality based on the 3-channel method of the 2.5D UNET model.

The architecture implements skip connections that bridge corresponding encoder and decoder levels, facilitating the preservation of fine spatial details and enabling the effective fusion of feature maps across the network. The network concludes with an output layer generating a 240 × 240 pixel segmentation mask, precisely corresponding to the spatial dimensions of the input imagery. As visualised in Figure 5.2a, the multi-class segmentation implementation handles mutually exclusive classifications, with class structures detailed in Figure 5.3a. The architecture employs a softmax activation in its final layer, specifically designed to process non-overlapping segmentation masks where each pixel belongs to exactly one class. The multi-label (Figure 5.2b) and binary class (Figure 5.2c) segmentation approaches handle non-mutually exclusive classifications, with their overlapping class relationships demonstrated in Figure 5.3b. The architectural design employs a sigmoid function in the output layer, allowing pixels to simultaneously belong to multiple classes.

Both the 3D U-Net model and the nnU-net (implemented as DynU-Net in MONAI [259]) maintain architectural consistency across all three segmentation paradigms: multi-class, multi-label, and binary class. This uniformity encompasses the configuration of final layers and channel structures throughout all segmentation variants. All model variants utilise a standardised input patch size of 192 × 192 × 128, representing an intentional departure from the original 128 × 128 × 128 configuration in the 3D models. This dimensional adjustment prioritises the capture of global contextual information, ultimately enabling more time-efficient optimisation of the DSC performance metric. The implementation of larger patch dimensions facilitates the processing of greater data volumes within each iteration cycle. This enhancement manifests in reduced computational overhead through two mechanisms: decreased total iteration requirements for processing the complete dataset while simultaneously accelerating the training convergence process. This design modification exemplifies the balance between computational efficiency and model robustness. While facilitating faster training and enhanced global context capture, the larger patch size potentially introduces generalisation limitations and

heightened overfitting tendencies. The original nnU-net patch dimensions, though potentially more robust, demanded extended training periods that contradicted this study's primary efficiency objectives. The prioritisation of computational efficiency extended to the enhanced nnU-net implementation. The comparative performance implications of these efficiency-focused architectural decisions are thoroughly examined in this chapter's results section.

5.2.3.2 Loss Function

The selection of loss functions for the U-Net architectures is fundamentally determined by the input mask characteristics, particularly the distinction between mutually exclusive and non-mutually exclusive class structures (as illustrated in Figure 5.3). This framework employs binary cross-entropy for multi-label and binary class segmentation scenarios, while multi-class cross-entropy is specifically implemented for multi-class segmentation tasks. This methodological distinction ensures appropriate loss function application for each segmentation approach.

Non-overlapping input masks in multi-class segmentation tasks (depicted in Figure 5.3a) necessitate the implementation of the softmax activation function for each class category. This choice directly aligns with the mutually exclusive nature of the class distributions. The implementation of softmax activation ensures probabilistic normalisation, with class probabilities summing to 1. This mathematical property ensures that probability increase for any single class inherently requires compensatory probability reduction across the remaining classes. Consistent with the interdependent class structure, the implemented multi-class cross-entropy loss quantifies distributional divergence between predictions and true labels. The mathematical formulation for N classes at each pixel position is:

$$CE = -\sum_{c=1}^{N} y_{o,c} \, log\left(p_{o,c}\right) \qquad (5.1)$$

In this formulation, $y_{o,c}$ serves as a binary truth indicator (0 or 1) specifying the correct class assignment for each pixel-class pair (pixel o, class c), while $p_{o,c}$ represents the corresponding predicted class probability. The overall loss metric is calculated by averaging values across all pixels in the image. The resulting multi-

class segmentation method provides improved analytical precision, specifically addressing uneven class distributions. This approach is particularly effective for advanced segmentation tasks, specifically when processing hierarchical input masks that capture three tumour classifications: ET, TC, and WT regions.

For binary class segmentation, each mask undergoes separate processing via the sigmoid function. Similarly, multi-label segmentation scenarios, characterised by overlapping input masks (visualised in Figure 5.3b), utilise the same sigmoid function methodology. Through independent class prediction processing, the sigmoid function yields separate class probabilities. This approach facilitates effective multi-label analysis by converting the task into separate binary classification tasks. The final loss measurement represents the average sum of losses from each label classification. At the pixel level, these approaches define the binary cross-entropy loss according to the following formulation:

$$BCE = -[y\,log(p) + (1-y)\,log(1-p)]\,(y)\,log(1-p) \qquad (5.2)$$

The variable y denotes the true pixel state (assigned 1 for object regions and 0 for background areas), whereas p expresses the estimated probability of the pixel being part of the object. The binary cross-entropy loss for the entire image is derived by calculating the mean of the pixel-wise losses. This mathematical framework, characteristic of binary class segmentation, offers both computational efficiency and analytical simplicity. The approach demonstrates particular efficacy in scenarios with single region input masks per training sample. Within the multi-label segmentation approach, the final loss is obtained by calculating the mean of region-specific loss values. While offering enhanced analytical flexibility and capability compared to binary segmentation, this method presents significant trade-offs through its computational intensity, resource consumption, and challenges in managing label interdependencies. The different segmentation methods have their own unique strengths and weaknesses. Where traditional methods typically implement a uniform segmentation strategy across tumour regions, our proposed region-focused ensemble learning model represents a departure from this conventional paradigm. By combining optimised segmentation methods, each

utilising customised normalisation, the ensemble model harnesses method-specific advantages to achieve superior tumour region delineation. The novel methodology yields precise segmentation outcomes, particularly when addressing the challenges posed by complex, heterogeneous tumour morphologies.

5.2.4 Dataset

Model development and validation are conducted using two different datasets, with dedicated sets for training and testing procedures. The BraTS 2021 dataset, a retrospective collection of brain tumour MRI scans from multiple institutions [178], [211], [212], was used as the first dataset in this study. The dataset contained 1251 training samples, 219 validation samples, and 570 testing samples at the time of our analysis. The dataset encompasses four MRI modalities: T1, T1ce, T2, and FLAIR, which together enable a detailed analysis of brain tumours. The tumour boundaries were carefully manually annotated by neuro-radiologists to ensure precise delineation.

As illustrated in Figure 5.3, the BraTS competition categorises GBM into three distinct tumour sub-regions:

1. NCR, assigned label 1,
2. ED, assigned label 2, and
3. ET, assigned label 4.

The combination of these sub-regions results in three clinically significant regions:

Label 4 designates the ET region, distinguished by its hyperintense appearance in T1ce images relative to T1 scans. Comprising labels 1 and 4, TC manifests hypo-intensity in T1ce sequences and constitutes an essential region for surgical excision [178]. WT, formed by the combination of labels 1, 2, and 4, presents distinctive hyper-intense features in FLAIR images, supporting thorough tumour assessment. The BraTS dataset is pre-processed using skull-stripping and co-registration to align with a standardised anatomical template. Each modality is subsequently resampled to an isotropic resolution of 1 mm^3, yielding a voxel matrix size of 240 × 240 × 155.

This study also utilises the STORM_GLIO dataset, a local dataset gathered in Wales from April 2014 to April 2018. The dataset consists of 108 glioblastoma patients, but only 53 patients have complete imaging data, including all four modalities (T1, T1ce, T2, and FLAIR), similar to the BraTS dataset. The adoption of DICOM formatting in STORM_GLIO, coupled with its non-standardised resolution and matrix size across patients and between MRI sequences, distinguishes it from BraTS and presents a challenge for analysis and processing of the dataset. The heterogeneity of the STORM_GLIO dataset, in terms of resolution and modality specifications, presents an additional challenge for preprocessing. For a detailed examination of the dataset, including its resolution and modality specifications, refer to Appendix (Table D- 5).

5.2.4.1 Data Pre-Processing

Dataset preparation began with 3D MRI scans (240 × 240 × 155) from BraTS 2021, followed by a 70/15/15 split for training, validation, and testing respectively, prior to generating either 2D slices (240 × 240) or 3D patches (192 × 192 × 128), as depicted in Figure 5.6.

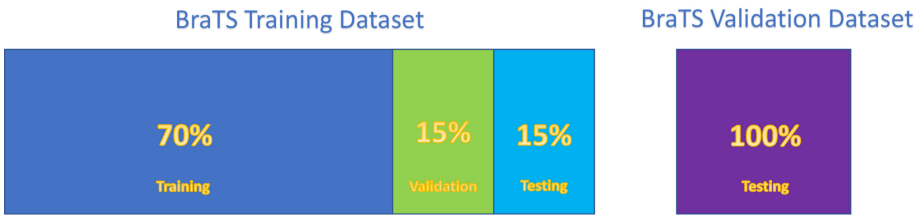


Figure 5.6 The use of the BraTS training and validation datasets.

The training dataset, reduced by 10% compared to the reference model [152], yields 155 slices per patient case for 2D U-Net implementations, while the 2.5D U-Net implementation employs a 3-channel slice extraction approach. The slice selection mechanism captured three consecutive images: central (N), preceding (N-1), and subsequent (N+1) slices. Synthetic black slices compensated for missing adjacent slices at volume boundaries to preserve the 3-channel structure. Patient-specific patches (192 × 192 × 128) were utilised for 3D U-Net processing, a dimensional choice that balanced comprehensive spatial information with processing efficiency.

Through this preprocessing framework, the dataset underwent targeted preparation to meet the distinct input specifications of both 2D and 3D U-Net variants. Derived from clinical practice, STORM_GLIO's MRI scans exhibited variable dimensions and utilise patient-specific coordinate systems, requiring mandatory registration procedures prior to analytical processing. The BraTS dataset provided pre-processed MRI scans, with T1ce modality registration aligned to the SRI24 atlas [223], establishing a unified coordinate system across modalities. The BraTS dataset provided pre-processed MRI scans, with T1ce modality registration aligned to the SRI24 atlas [223], establishing a unified coordinate system across modalities. The preprocessing pipeline further included comprehensive skull stripping and detailed sub-tumour class segmentation. The alignment of STORM_GLIO to BraTS standards involved executing a modified BraTS preprocessing pipeline, with two key adjustments. Integration of the CaPTk [73], [74] served as a crucial step in achieving format compatibility. Two key deviations from the standard pipeline were implemented: skull extraction was performed using the more advanced HD-BET [75] tool instead of CaPTk's default method, and SRI24 atlas registration was excluded to maintain ground truth data fidelity due to the risk of ground truth deformation. Two key deviations from the standard pipeline were implemented: skull extraction was performed using the more advanced HD-BET [75] tool instead of CaPTk's default method, and SRI24 atlas registration was excluded to maintain ground truth data fidelity due to the risk of ground truth deformation. Through these pipeline modifications, DL model segmentations could be effectively converted to RTSTRUCT, ensuring practical clinical deployment capabilities. The complete MRI acquisition parameters unique to the STORM_GLIO dataset can be found detailed in Appendix (Table D- 5) for comprehensive technical reference.

5.2.4.2 Implementation Details

Implementation of the training protocol for 2D and 2.5D U-Net configurations encompassed a comprehensive 100-epoch cycle utilising 16-sample batches, while employing Adam optimisation methodology [260] with learning rate established at 0.0001. The training dataset underwent augmentation procedures including multidirectional image rotation and dual-axis flipping transformations with all

computational processes executed on specialised hardware consisting of an NVIDIA RTX 3070 graphics processing unit with 8 GB of dedicated memory, alongside an Intel i7-11700 processor and 32 GB of system RAM. Implementation of 3D U-Net and nnU-net model training required adjustment of key parameters, extending the epoch count to 150 and reducing batch size to 4 samples, while preserving the Adam optimisation methodology with an unchanged learning rate of 0.0001. The computational infrastructure for these experimental procedures comprised an NVIDIA RTX 3090 graphics processor featuring 24 GB of dedicated memory, supported by an Intel i7-11700 processing unit and 32 GB system RAM, with software deployment performed in a Linux operating system environment leveraging Python 3.9.13 and PyTorch version 1.10 frameworks. The experimental procedures were compared with the integration of the BraTS 2021 challenge-winning Docker image containing the extending nnU-net framework [152], which incorporates a comprehensive suite of 10 distinct model architectures and advanced post-processing protocols. To ensure optimal segmentation accuracy, the STORM_GLIO dataset was resampled to a consistent isotropic resolution of 1 mm^3, resulting in a uniform matrix size of 240 × 240 × 155, which was applied to all 3D models.

After segmentation, the outputs were reverted to their original voxel size and dimensions, facilitating a direct comparison with the reference ground truth data. Quantitative assessment of segmentation accuracy was conducted using the DSC as the principal evaluation metric. The DSC calculation determines how well the predicted segmentation masks (Y_pred) match the expert-annotated ground truth masks (Y_true) by measuring their spatial correspondence [261]. Values of the DSC metric are bounded between 0 and 1, where unity indicates optimal segmentation performance with exact correspondence between ground truth and algorithmic output. By convention, cases where both the reference and predicted masks are devoid of tumour pixels are assigned a maximum score of 1 [182]. The segmentation performance assessment incorporated multiple complementary metrics beyond the DSC, including sensitivity (true positive rate), specificity (true negative rate), and the 95th percentile Hausdorff distance (HD95) for target region evaluation. These

measurements were calculated utilising the SegmentationMetrics library (version 1.0.1) in Python.

$$DSC = \frac{2|Y_{true,pos} \cap Y_{pred,pos}|}{|Y_{true,pos}| + |Y_{pred,pos}|} \quad (5.3)$$

$$Sensitivity = \frac{|Y_{true,pos} \cap Y_{pred,pos}|}{|Y_{true,pos}|} \quad (5.4)$$

$$Specificity = \frac{|Y_{true,neg} \cap Y_{pred,neg}|}{|Y_{true,neg}|} \quad (5.5)$$

$$HD95 = 95th\ percentile\ of\ \{max_{a \in Y_{true,pos}} min_{b \in Y_{pred,pos}} ||a - b||_2, max_{b \in Y_{pred,pos}} min_{a \in Y_{true,pos}} ||a - b||_2\} \quad (5.6)$$

5.3 Results and Discussion

We conducted a systematic evaluation of the proposed approach through three distinct experimental protocols. The preliminary phase concentrated on architectural comparison and selection; wherein various candidate models were evaluated using the BraTS 2021 dataset to determine the optimal network configuration. We evaluated and compared multiple models to select the most effective architecture for implementing the RFS+ strategy. The second experimental phase employed the BraTS validation dataset to conduct a comparative assessment between our proposed U-net architecture and the state-of-the-art extended nnU-net, which achieved superior performance in the BraTS 2021 challenge. The third experimental phase evaluated the efficacy of the RFS+ strategy by comparing the performance of top-ranked architectures against two baselines: the extended nnU-net and the equivalent models without RFS+ implementation.

5.3.1 Model Selection Using the BraTS 2021 Dataset

We describe herein the experimental methodology, evaluation protocols, and comparative assessments conducted on the proposed DL models, highlighting the various intensity normalisation strategies employed in their respective implementations. Model performance evaluation was conducted using dual

datasets, BraTS 2021 and STORM_GLIO, facilitating a methodical selection process. The quantitative outcomes of these comparative analyses are summarised in Table 5.1. TC segmentation performance is critical since GTV delineation depends on it. In our results, the baseline 2D U-Net (with Z-score normalisation and multi-class approach) achieved the best TC score. Specifically, 3D U-net and nnU-net achieved moderate performance due to aiming time-efficiency with a bigger patch size.

Table 5.1 Single Model Comparison of DSC Scores for the 2D, 2.5D, 3D U-NET, and nnU-net with Z-score normalisation and multi-class approach on the BraTS 2021 training dataset.

Model	ET	TC	WT
nnU-net	83.96	88.34	92.53
3D U-net	83.21	87.55	91.67
2.5D U-net	84.34	88.55	91.64
2D U-net	84.99	89.71	91.65

The segmentation performance on the STORM_GLIO dataset was evaluated using models pre-trained on BraTS data, implementing a multi-class segmentation framework with Z-score intensity normalisation. The quantitative results are presented in Table 5.2.

Table 5.2 Comparison of the models with multi-class approach on STORM_GLIO.

Models	GTV
nnU-net	77.45
3D U-net	75.74
2.5D U-net	70.35
2D U-net	78.43

The 2D U-net configuration exhibited exceptional performance, achieving superior DSC measurements for both TC and GTV delineation compared to alternative architectures. The observed superiority of the 2D U-net architecture can be primarily attributed to its exclusive focus on individual slice processing, whereas the 2.5D and 3D variants, which incorporate volumetric information, demonstrated reduced accuracy potentially due to heterogeneous slice characteristics. For example, 2.5D U-Net yielded poor performance, highlighting the challenges posed by real-world data. Since the 2D U-Net yielded robust performance with time-

efficient limitations, detailed 2D U-Net variants from three segmentation approaches were trained.

Table 5.3 Segmentation Approach Comparison of DSC scores for binary class, multi-label, and multi-class approaches of 2D U-net with several intensity normalisation techniques on the BraTS 2021 training dataset.

Intensity Norm. Tech	Segmentation Approach	ET	TC	WT
Nyul	multi-class	79.44	79.53	88.98
	multi-label	83.52	88.78	92.05
	binary class	84.21	89.42	90.30
Z-score	multi-class	84.99	89.71	91.65
	multi-label	82.29	87.27	92.24
	binary class	85.19	89.48	92.18

A comparative analysis of various intensity normalisation approaches applied to the 2D U-net architecture is presented in Table 5.3, quantified through DSC measurements on the BraTS 2021 training dataset. For TC delineation, optimal performance was achieved by three distinct configurations: the Z-score normalisation with multi-class segmentation achieved a DSC of 89.71%, followed by Z-score with binary classification at 89.48%, and Nyul normalisation with binary classification at 89.42%. A weighted average ensemble learning strategy was implemented, leveraging their individual strengths within the proposed methodological framework.

5.3.2 Benchmarking the RFS+ Method: A Comparative Analysis

The BraTS 2021 validation dataset served as the benchmark for comparative analysis between the developed U-net variants and both conventional and extended implementations of the nnU-net architecture. Z-score normalisation was uniformly applied across all architectural variants of DL models, with their comparative performance metrics on the BraTS validation cohort presented in Table 5.4.

Table 5.4 Comparison of the BraTS validation dataset based on online evaluation.

Models	DSC(ET) (%)	DSC(TC) (%)	DSC(WT) (%)
Extended nnU-net	84.51	87.81	92.75
nnU-net	78.65	85.96	91.67
3D U-net	78.89	81.05	91.16
2.5D U-net	78.80	84.23	90.90
2D U-net	77.45	82.14	90.82

Performance analysis revealed the extended nnU-net [152] architecture achieved superior segmentation results compared to both the proposed U-net variants and standard nnU-net implementations across all anatomical regions under evaluation. The enhanced performance metrics can be traced to the extended nnU-net's architectural modifications, which were deliberately designed to capitalise on the uniform size and resolution parameters inherent to the BraTS training dataset. The consistent matrix dimensions and resolution parameters shared between the BraTS validation and training datasets enabled the extended nnU-net to leverage its specialised modifications, resulting in elevated DSC scores across all tumours sub-regions. The computation of DSC scores requires online submission and evaluation, as the ground truth segmentations for the BraTS validation dataset are maintained privately by the challenge organisers. The comparative DSC metrics presented in Table 5.4 encompass segmentation performance across U-net variants, standard nnU-net, and extended nnU-net implementations, thereby contextualising the methodology's effectiveness relative to state-of-the-art standards.

5.3.3 Ablation Study

Table 5.5 Ablation study on U-net.

	Z-Score Normalisation			Nyul Normalisation			Combined Method		GTV DSC (%)
	Multi-class	Multi-label	Binary	Multi-class	Multi-label	Binary	Union	Ensemble	
Base U-net (Multi-class)	Yes								78.43
Multi-label		Yes							77.91
Binary			Yes						78.22
Base U-net (Multi-class)				Yes					77.61
Multi-label					Yes				78.20
Binary						Yes			78.91
RFS	Yes	Yes	Yes				Yes		78.51
RFS+(only Z-score normalisation)	Yes	Yes	Yes					Yes	78.69
Proposed RFS+	Yes		Yes			Yes		Yes	79.22

Various segmentation methodologies and normalisation techniques were tested to improve ensemble learning performance. To examine the contribution of each segmentation approach, an ablation study was undertaken, systematically removing and re-evaluating each component to quantify its impact on the final outcome. The multi-class segmentation U-net was selected as the baseline model (designated as "base U-net"), a choice informed by the distinct, non-overlapping characteristics of the masks in the training dataset. A detailed comparative analysis is presented in Table 5.5, evaluating the performance of multiple U-net variants (differing in their segmentation methodologies and normalisation procedures) using the STORM_GLIO dataset within the RFS+ framework, enabling a thorough assessment of their relative strengths and weaknesses. The base U-net entries in Table 5.5 establish the reference DSC metrics against which the performance of other model variations can be quantitatively assessed. The table was organised to showcase various combinations of segmentation and normalisation methods, with the rightmost column dedicated to the evaluation of GTV segmentation accuracy, as

measured by the DSC metrics, enabling a detailed examination of this key performance metric. A normalisation-dependent trend emerges from the results, where Z-score normalisation maximises DSC for the base U-net, and Nyul normalisation optimises performance when used in conjunction with the binary class model, highlighting the importance of normalisation technique selection.

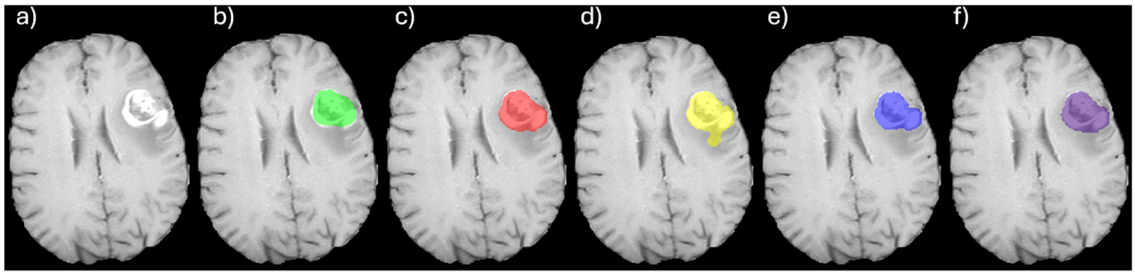


Figure 5.7 Predictions of models with different segmentation approaches on STORM_GLIO. a) T1ce, b) ground truth, c) 2D U-net Nyul/binary class, d) 2D U-net Z-score/multi-class, e) 2D U-net Z-score/binary, and f) 2D U-net with RFS+.

Figure 5.7 illustrates the variable performance of different segmentation and normalisation techniques across distinct features. The lack of a consistently superior configuration emphasises the importance of developing task-specific approaches for each segmentation objective. Implementation of the standard RFS technique, which combines model segmentations using a union-based fusion, yielded a small improvement in performance, with the DSC metric, from 78.43% to 78.51%, as described in [245]. The RFS+ method represents an advancement over the traditional RFS approach by leveraging weighted ensemble learning, which allocates weights to models according to their accuracy on the BraTS training dataset. The combination of weighted ensemble learning and Z-score normalisation did not result in a significant improvement in segmentation accuracy (RFS+ with only Z-score 78.69%) compared to previous results (single models with Z-score normalisation: Multi-class (78.43%), Multi-label (77.91%), Binary class (78.22%), and RFS (78.51%)). For the RFS+ implementation with Nyul normalisation, we employed a selection strategy that utilised the three models achieving superior DSC metrics in BraTS training dataset evaluation. Evaluation of the RFS+ methodology on the STORM_GLIO dataset revealed outstanding results, with a DSC of 79.22% from a DSC of 78.43% for GTV segmentation, highlighting the effectiveness of this

approach in accurately delineating tumour volumes. The study's outcomes emphasise the importance of model diversification and ensemble learning in achieving enhanced performance, as the strategic integration of multiple normalisation methods and segmentation parameters, coupled with ensemble methodology, results in marked improvements in the accuracy and robustness of U-net models.

5.3.4 Validation of RFS+ on a Local Dataset

This section evaluates the performance of top-performing U-net and nnU-net models in segmenting GTV using a local dataset, comparing the standard and RFS+-enhanced versions of these models, as well as an expanded nnU-net configuration. To establish a comparative framework, we implemented the RFS algorithm [92] with a 2D U-net model, and conducted a performance evaluation using a suite of quantitative metrics, comprising DSC, HD95, sensitivity, and specificity, with the detailed results presented in Table 5.6, allowing for a thorough comparison with other models.

Table 5.6 Comparison of recent models: base models (with Z-score normalisation and Multiclass approach), the proposed models, and the state-of-the-art model on the GTV label. Upper arrows indicate that a higher value is preferable, while lower arrows indicate that a lower value is most favourable.

Models	Details	DSC ↑	HD95 ↓	Sensitivity ↑	Specificity ↑
nnU-net-Large	Extended nnU-net [152]	79.09	7.80	74.07	99.97
nnU-net	The base model	77.83	10.72	74.65	99.95
	RFS+	78.30	8.20	73.59	99.97
2D U-net	The base model	78.43	8.80	77.24	99.94
	RFS [92]	78.51	11.33	78.48	99.93
	RFS+	79.22	8.10	76.93	99.95

Analysis of Table 5.6 demonstrates superior performance of the RFS+-enhanced 2D U-net model compared to the extended nnU-net implementation. This enhanced performance was attained through the strategic fusion of the three highest-performing models identified during prior experimentation. The enhanced performance of the RFS+-integrated 2D U-net was achieved through the simultaneous application of three segmentation methodologies, multi-class, multi-

label, and binary class, combined with both Nyul and Z-score normalisation approaches. In comparison to the extended nnU-net, which employed a single approach with multi-label segmentation and Z-score normalisation, the RFS+ ensemble methodology demonstrated enhanced performance, characterised by higher DSC scores in GTV segmentation and better generalisation properties, outperforming reference implementations. The RFS+ approach yielded about a 1% relative improvement in DSC over the standard RFS method, as determined by quantitative analysis, and this advancement in boundary delineation accuracy, measured by HD95, holds considerable promise for improving the effectiveness of therapeutic planning, encompassing both surgical and radiotherapy treatments, by enabling more precise targeting and treatment of tumours. A comparison of HD95 metrics revealed that RFS+ and the extended nnU-net exhibited minimal differences (8.1 vs 7.8), indicating that RFS+ maintains high clinical accuracy standards while offering improved computational efficiency, which supports the validity of our approach. The empirical evidence demonstrates RFS+'s capability to enhance tumour segmentation accuracy, with relevance for clinical workflows demanding high-precision boundary definition. The model's performance characteristics make it especially suitable for applications where accurate tumour delineation is crucial.

The successful deployment of computer-aided systems in clinical practice requires optimal sensitivity to detect tumour tissue accurately and comprehensively, a necessity that is heightened in situations where automated analyses play a critical role in informing diagnostic processes and evaluating treatment efficacy. By improving sensitivity from 74.07% to 76.93%, our RFS+ method was able to identify more tumour tissue than the state-of-the-art model, resulting in fewer false negatives and a more comprehensive analysis of tumour regions. The enhanced sensitivity demonstrated by the RFS+ method has important clinical implications, especially in the context of therapeutic planning, where accurate tissue differentiation is essential for precise radiotherapy administration and optimal surgical approach determination. The RFS+ approach demonstrates a comparable result in specificity, reaching 99.97% versus 99.95% for the top-performing model, while sustaining its resource-efficient profile, thereby fulfilling the objective of

maintaining a high level of healthy tissue recognition, which is essential for ensuring both clinical safety and feasible implementation in real-world settings. A comparative analysis of tumour segmentation performance is presented in Figure 5.7, which displays the outputs of traditional 2D U-net architectures alongside those of RFS+-integrated models, revealing distinct differences in their ability to accurately delineate tumour regions and highlighting the advantages of the RFS+ approach.

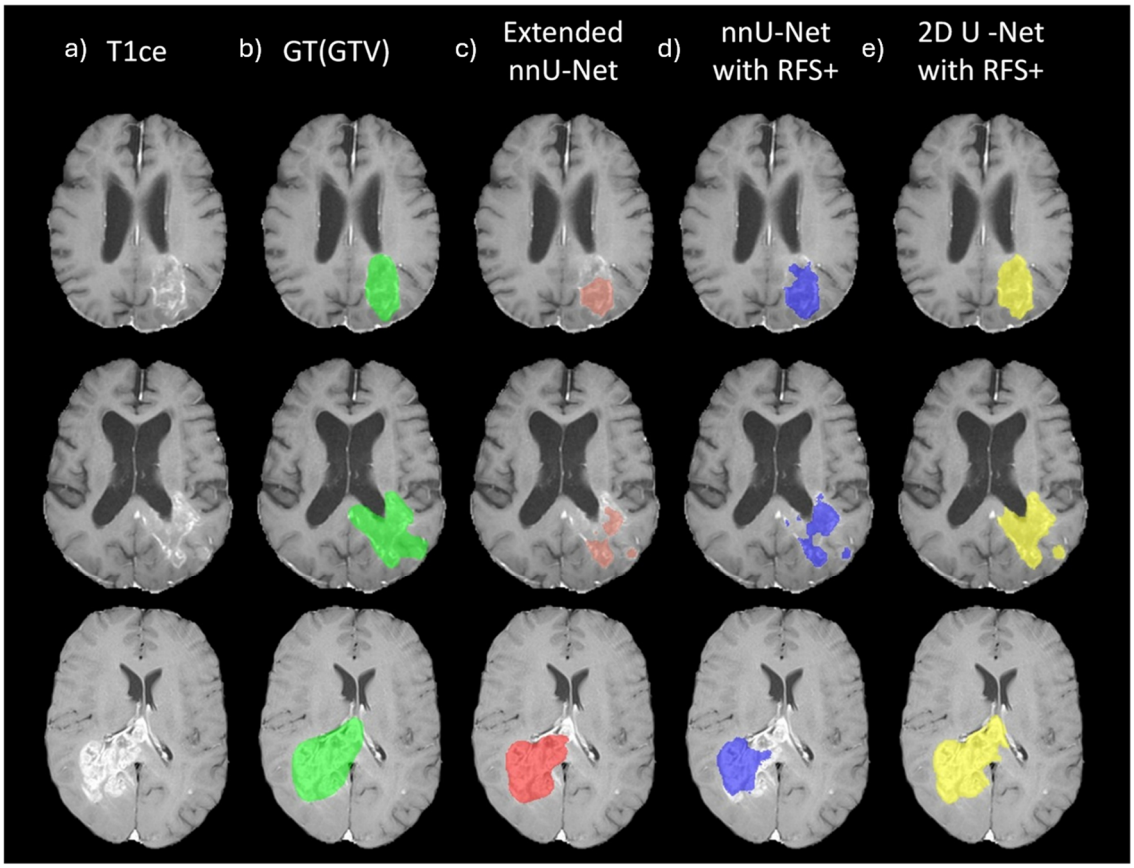


Figure 5.8 Predictions of models on STORM_GLIO. a) T1ce, b) ground truth (GTV), c) extended nnU-net model, d) nnU-net with RFS+, and e) 2D U-net with RFS+.

The performance of the baseline models was heterogeneous and task-dependent, with significant variability in accuracy observed between TC and GTV segmentation, indicating that the effectiveness of each approach is highly contextual and that no single methodology consistently outperformed the others. By combining the benefits of different tumour segmentation methods, RFS+ creates a comprehensive ensemble learning method that overcomes the limitations of individual approaches,

resulting in improved tumour mapping capabilities and more effective detection outcomes. The quantitative comparison presented in Figure 5.8 reveals the superiority of the RFS+-augmented 2D U-net over the extended nnU-net in terms of comprehensive tumour tissue segmentation, with the visual results validating the RFS+ model's enhanced performance in detecting tumour tissue. The local dataset presented methodological challenges due to its heterogeneous nature, characterised by variations in image resolution between patients and inconsistencies in matrix dimensions within individual patients across different imaging modalities. In the presence of dataset heterogeneity, the RFS+ framework demonstrated superior adaptability and outperformed the winning model, validating its robust architectural design and ability to effectively process a wide range of imaging parameters and scan qualities. The development of the RFS+ methodology marks a substantial advancement in the field of neuro-oncological image analysis, as it enables the optimisation of DL models for clinical deployment, which is a crucial step in overcoming the challenges of applying medical AI in translational settings. Benchmarking results against the extended nnU-net model demonstrate that RFS+ offers considerable computational advantages, including a 10% reduction in training data needs, a 67% decrease in memory usage, and a 92% decrease in training duration, as shown in Appendix (Table D- 2, Table D- 3, Table D- 4), which collectively indicate a substantial optimisation of computational resources. The reduction in resource requirements achieved by RFS+ is a pivotal development for the clinical adoption of DL models, as it facilitates the integration of cutting-edge segmentation technologies into existing healthcare infrastructure, enabling the widespread deployment of advanced diagnostic tools in clinical settings.

On local test data, the RFS+ framework attained a DSC of 79.22%, indicating its strong capacity for precise tumour delineation in clinically relevant settings, and highlighting its suitability for application in real-world medical imaging contexts. The ensemble architecture of RFS+ combines the predictions of multiple model variants, each with its own normalisation strategy, to create a robust and accurate segmentation model that leverages the complementary learning patterns of its

constituent models, thereby reducing the impact of individual model biases and improving the detection of segmentation boundaries through a collective and synergistic learning process. The multi-method approach of RFS+ provides the flexibility to handle the complexities and variations of in real-world hospital scan data, making it a valuable asset for clinicians working in diverse settings. A systematic and rigorous comparative analysis of the RFS+ strategy against prominent benchmark methodologies is crucial to validate its effectiveness and identify areas for improvement, ensuring that its contributions are accurately contextualised within the broader literature and providing a foundation for future research and development. The architectural differences between DeepMedic's multi-scale paradigm and RFS+'s domain-adaptive framework result in unique strengths, as DeepMedic's approach is well-suited for integrating features across multiple spatial scales, while RFS+ exhibits superior resilience to variations in dataset characteristics, yet DeepMedic's sophisticated feature integration capabilities may offer advantages in scenarios characterised by complex tumour morphologies. While the Cascade U-net framework's architecture is well-suited for achieving high-precision boundary delineation, particularly in regions with intricate tumoural structures, its cascading approach incurs considerable computational costs, which may limit its practical implementation in resource-constrained settings, despite its potential advantages over RFS+ in certain scenarios. The 3D-DSN framework's deep supervision mechanisms facilitate the capture of features at various scales, allowing for more detailed segmentations that may surpass those achieved by RFS+, and the architecture's capacity to utilise intermediate layers for nuanced feature detection may confer advantages that warrant thorough consideration in comparison to the RFS+ methodology, particularly in applications where subtle feature is critical. The robust generalisation capabilities of nnU-net, as evidenced by its performance across diverse medical image segmentation tasks, underscore its methodological reliability, but RFS+ offers distinct advantages through its efficient computational profile and demonstrates notable efficacy in particular dataset scenarios, particularly in resource-constrained environments where localised implementations are common. With its autonomous configuration and adaptive architectural design, the nnU-net framework exhibits exceptional

ability to generalise, achieving robust performance across multiple medical segmentation challenges and varied data scenarios, which highlights its potential as a reliable and effective solution for diverse clinical applications. The design of nnU-net is guided by principles of scalability and adaptability, which has established it as a leading methodology in medical image analysis, and comprehensive validation studies have provided strong evidence of its exceptional robustness and ability to generalise across diverse imaging protocols and pathological conditions. Although the RFS+ methodology has shown encouraging preliminary results, it still lacks the extensive empirical validation and automated optimisation capabilities that would establish its robustness and generalisability, and therefore, a thorough investigation of its empirical validation and automated optimisation capabilities is necessary to determine its full potential and limitations. The scalability of the RFS+ framework is called into question by the robust performance of nnU-net on large and complex datasets, suggesting that the RFS+ framework may struggle with highly heterogeneous datasets or those requiring extensive computational resources, and highlighting the importance of addressing these potential limitations to ensure the framework's widespread applicability and effectiveness in real-world medical imaging scenarios.

The RFS+ methodology has demonstrated notable performance in specific dataset contexts, including localised applications, but the current validation framework is largely restricted to comparisons with nnU-net, indicating a potential gap in the evaluation process that could be addressed by conducting more extensive comparative analyses across a range of methodologies to further establish the methodology's relative merits and limitations. A more comprehensive understanding of RFS+'s capabilities and limitations can be achieved through future research that undertakes extensive comparative analyses, incorporating a wide range of current methodological approaches, which would enable a more nuanced evaluation of its performance metrics and provide valuable insights into its relative strengths and weaknesses within the context of contemporary solutions. Clinical datasets, marked by substantial heterogeneity in terms of imaging protocols, patient characteristics, and pathological conditions, pose a fundamental challenge to the

generalisability of DL models, including RFS+-based approaches, highlighting the need for careful consideration of these factors in the development and validation of such architectures. The findings of this study, which focused on GTV/TC segmentation using ET and NCR labels from a local dataset, demonstrate that the effectiveness of intensity normalisation techniques is highly context-dependent, varying significantly across different datasets and label categories, and thus, highlighting the importance of continued research to address this critical issue and develop more robust and generalisable normalisation strategies.

5.4 Conclusions

The current research investigates the advantages of employing the RFS+ approach for brain tumour segmentation in MRI, particularly in the context of clinical datasets with heterogeneous characteristics, such as varying resolution parameters and matrix dimensions, to assess its effectiveness and generalisability in complex and diverse imaging environments. The assessment of model performance yielded dataset-specific results, with the extended nnU-net showing excellence in BraTS validation metrics and the proposed 2D U-net with RFS+ integration achieving outstanding results in the local dataset analysis, particularly in GTV segmentation tasks, where it attained a DSC accuracy of 79.22%, underscoring the need for careful model selection and optimisation for specific datasets. The superior performance of the model in local dataset analysis can be attributed to its ability to effectively manage variability in imaging parameters, such as non-standardised resolutions and diverse matrix sizes, which is a common feature of real-world datasets, but not typically seen in more standardised datasets like BraTS, highlighting the model's ability to generalise to diverse imaging conditions. These results demonstrate the vital importance of adapting intensity normalisation methodologies to specific ROI, such as TC and GTV, and show that customised approaches can significantly enhance segmentation performance, emphasising the need for nuanced and region-aware normalisation strategies. In the BraTS 2021 validation cohort, the U-net model employing a multi-class approach yielded varied performance metrics for different tumour regions, with DSC scores ranging from 77.45% for ET to 82.87% for TC and 90.82% for WT, demonstrating the model's ability to adapt to distinct tumour

characteristics and segmentation challenges. A comparative analysis demonstrated the superiority of the RFS+ strategy in local dataset applications, where its ensemble learning framework and heterogeneous normalisation techniques conferred unique operational benefits, outperforming other models and establishing its effectiveness in real-world scenarios. The RFS+ methodology demonstrated substantial computational efficiencies over the extended nnU-net, with empirical measurements showing decreased resource requirements, specifically a 10% reduction in training data needs, a 67% decrease in memory consumption, and a 92% decrease in computational time, as reported in Appendix (Table D- 2, Table D-3 and Table D- 4). The capabilities of the RFS+ methodology, as demonstrated in this study, suggest that it can significantly enhance the accuracy of brain tumour segmentation, especially in the presence of variable clinical imaging protocols and data acquisition parameters, which often pose significant challenges in medical image analysis. The RFS+ framework's ability to balance superior segmentation performance, low computational overhead, and versatility in clinical applications marks it as a promising advancement in medical image processing. It has the potential to significantly benefit the field by providing medical imaging professionals with a reliable, efficient, and widely applicable tool. To fully establish the RFS+ strategy's broad effectiveness, further validation studies are required, involving diverse patient cohorts and extensive comparative evaluations against state-of-the-art segmentation frameworks, to confirm its generalisability and superiority in various clinical scenarios. Future research will seek to harness the efficiency gains of RFS+ while expanding its scope to achieve superior segmentation performance, with plans to conduct rigorous benchmarking exercises against state-of-the-art methodologies in diverse clinical settings and applications, ultimately validating its effectiveness and versatility as a reliable tool for medical image analysis. To ensure that models can perform effectively in a wide range of real-world clinical settings, it is crucial to develop sophisticated data augmentation techniques that are tailored to the unique characteristics and challenges of medical imaging applications.

6. Conclusions and Future Works

6.1 Clinical Data Curation for Glioblastoma Multiforme Brain Tumour Analysis

In Chapter 2, a preprocessing pipeline was developed to align with clinical requirements for GBM MRI datasets, prioritising automatic tumour segmentation and radiomic analysis. This involved optimising the BraTS preprocessing pipeline through a publicly available standardised dataset and a local GBM dataset (STORM_GLIO). Qualitative assessments revealed significant deformation, particularly at boundary regions, suggesting a potential impact on subsequent radiomic analyses. These findings highlight the importance of further research on how preprocessing methods affect the reproducibility of radiomic studies, with particular emphasis on ROI deformation. Additionally, Whybra and Spezi [252] highlighted variations in contour handling across different software platforms, which may affect the computation of engineered RFs. Integrating these results with the insights from Chapter 2 could amplify their significance, warranting future research to explore the direct effects of such pre-processing discrepancies on radiomic outcomes. On the other hand, precise brain extraction techniques remain essential for achieving accurate automated tumour segmentation [71], [79], thus highlighting their significance in radiomic analytical frameworks and clinical integration protocols [69]. Although brain extraction tools contribute up to 15.7% to automated tumour segmentation accuracy [79], exploring automated segmentation techniques that eliminate dependence on brain extraction tools remains an open and active area of research, requiring further external validation studies. HD-BET outperformed CaPTk for skull-stripping, benefiting from GPU acceleration, achieved a 20-fold speed increase compared to CPU-based execution, thus making it more suitable for demanding clinical workflows. The proposed preprocessing pipeline improved automated tumour segmentation performance, producing outputs in RTSTRUCT, a DICOM-compliant format. Automated DL-based tumour segmentation, trained on extensive BraTS-format datasets, including TC labels, rather than the limited availability of traditional radiotherapy GTV labels, may enhance segmentation accuracy in clinical settings. Automated contouring

software has the potential to reduce contouring time by up to 80% for various types of cancer, providing reliable and reproducible contours that are essential for planning radiation therapy [262]. Further research is essential to evaluate the clinical acceptability of outputs from auto-contouring studies for radiotherapy planning [263].

In Chapter 4, the proposed pipeline served as the foundation for preprocessing STORM_GLIO and conducting subsequent radiomic analyses. The cumulative effect of multiple preprocessing steps lowered the DSC score of automated tumour segmentation below 75%, indicating that while the proposed pipeline provided an improvement, it remains below the performance of the recent BraTS-winning models, which achieved 87.81% average DSC on external validation [264]. To address this, Chapter 5 explores the replacement of outdated segmentation models like DeepMedic with state-of-the-art alternatives to improve GTV/TC segmentation performance for clinical settings. Future studies could aim at optimising its integration within radiomic workflows, with the goal of minimising segmentation-related artefacts and enhancing the integration of PACS (Picture Archiving and Communication System). Addressing these challenges can help avoid common pitfalls in radiomic studies, thus facilitating a smoother integration into clinical practice [265]. With the growing use of multicentre datasets in radiomic studies, harmonisation of imaging data continues to be a key research focus. Notable efforts, such as ComBat harmonisation [67], [70], [83], [266] for multicentre MRI-based radiomics features, as well as comprehensive reviews on radiomic methodology and standardisation across imaging protocols, underscore the challenges involved. However, the field still lacks universally accepted standards, highlighting the need for further investigations to promote accuracy and reproducibility in MRI-based radiomic analyses. Balancing the trade-off between non-standardised (i.e., non-reproducible) and over-standardised (i.e., potentially information-losing) MRI data and RFs, both prior to and following feature extraction, remains an important challenge [267]. This balance must be carefully managed before and after feature extraction to ensure the reliability and analytical value of radiomic studies.

6.2 A Novel Hybrid Feature Selection Method for Radiomic-Based Overall Survival Analysis in Glioblastoma Multiforme

In Chapter 3, we introduced a novel SI-based feature selection methodology designed to enable interpretable analysis of OS prediction via traditional ML frameworks. Due to clinical data scarcity in the medical domain [189], the study was limited to include solely the clinical feature of Age. Drawing inspiration from the radiomic analysis guidelines articulated by van Timmeren [67], a hybrid feature selection strategy was formulated. Among the validated models, the Cox regression model with the PSO-enhanced LASSO feature selection method showed the most consistent performance across datasets. Additionally, the development of a novel SI-based feature selection method that achieves state-of-the-art performance while maintaining a high degree of interpretability [235], a critical requirement for clinical applicability. To the best of our knowledge, this study constitutes the first development of a model with an SI-based feature selection method that achieves statistically significant risk stratification while prioritising interpretability without compromising predictive performance. Shape-based (morphological) features, followed by FLAIR-derived texture features, were the most influential predictors after age variable. Selected RFs primarily originated from TC regions. These findings highlight the superior performance of shape-based and FLAIR-derived texture features. Thus, this study suggests that future research prioritises TC as the primary ROI to improve model generalisability**.** Future research could focus on reducing the number of MRI sequences and ROIs, further standardising datasets, and broadening external validation studies. Another promising direction involves the optimisation of model hyperparameters to enhance performance by utilising SI-based methods for traditional ML models. Additionally, the integration of convolutional filter-based RFs [236] and DL-derived features presents opportunities for further improvement [268], due to an increase in complex pattern recognition. Lastly, the incorporation of more comprehensive clinical data alongside multi-omics datasets, encompassing pathomics and genomics [269], [270], holds the potential to significantly advance the predictive capacity of future models. As the number and variety of input features increase, the complexity of developing models also increases. This might require the

integration of DL-based features and the use of DL-based models. A key area of ongoing research focuses on improving the interpretability of DL models, which is vital for enhancing the reliability and clinical generalisability of these models [57], [269].

6.3 Development of a robust and interpretable clinical-radiomic model for predicting overall survival in Glioblastoma Multiforme

In Chapter 4, this study investigated radiomic analysis for GBM patients, leveraging varied preprocessing approaches, including the proposed pipeline in Chapter 2. Following the radiomic guideline [67], this study developed a robust radiomic model addressing existing clinical challenges and limitations. The model's distinctive approach encompasses reduced MRI sequence dependency, single ROI utilisation, and preliminary evidence supporting GTV-TC interchangeability [245]. For GBM OS analysis, this radiomic study encompassed the most extensive patient cohort reported in the literature, following the IBSI guidelines. Reproducibility analyses resulted in the selection of only two robust RFs. The final clinical-radiomic model comprises a single clinical variable (patient age) and two robust RFs: a shape-based radiomic feature and a texture-based radiomic feature derived from the FLAIR MRI sequence. The developed clinical-radiomic model exhibited promising predictive performance in the holdout test cohort, with notable results compared to current literature. Future improvements in performance may be achieved through the expansion of the clinical dataset and the incorporation of multi-omics data, such as pathomics and genomics [269], [270]. The validated interchangeability between GTV and TC contours underscores the model's compatibility with established clinical workflows, facilitating seamless integration into existing practice protocols. While using multi-ROI may enhance the characterisation of tumour heterogeneity, it also might introduce variabilities, such as including both inter- and intra-tumour segmentation variability [67], [271], affecting feature consistency and extraction reliability. A simplified single-ROI method centred on the TC/GTV region might facilitate improved reproducibility and consistent feature extraction. The identification of a suitable ROI method needs careful consideration, balancing

reproducibility and feature consistency with clinical utility to enhance outcomes in radiomic analysis. This thorough investigation might enhance the development of clinical decision-support systems for GBM treatment and management, demonstrating considerable potential for future clinical implementation [265].

6.4 An efficient DL-based automated tumour segmentation model complained with clinical settings

Chapter 5 introduced RFS+, a novel approach for automated tumour segmentation, aiming to mitigate performance drops of state-of-the-art DL models on clinically heterogeneous data [183]. Domain shift, which results from differences between source and target domains, such as scanner type and patient demographics, can reduce segmentation model performance on unseen datasets [272]. The issue emphasises the urgent need for models capable of generalising across domains, which remains an ongoing challenge [273], and draws attention to the equally important task of standardising practices in medical imaging [183], [267]. RFS+ directly addressed limited generalisability by integrating diverse normalisation techniques and three segmentation approaches. The DL models, using RFS+, compared to a state-of-the-art model, the extending nnU-Net [152]. While DL models achieving superior results on standardised datasets, its efficacy dropped notably when applied to STORM_GLIO outlined in Chapter 2**.** In contrast, the proposed RFS+ approach demonstrated a notable improvement in performance for the nnU-net and conventional U-Net models. On the STORM_GLIO dataset, the 2D U-Net with RFS+ achieved the best segmentation results, matching extended nnU-Net performance with lower computational cost. RFS+ enhanced models showed clear improvements by leveraging intrinsic interrelationships between tumour regions. The methodology targeted limitations of resampling in suboptimal MRI acquisitions, with the 2D U-Net combined with RFS+ effectively reducing interpolation artifacts and supporting robust training on original data. However, the 2D model with RFS+ achieved equivalent segmentation performance with superior computational efficiency, indicating it could be explored further as an option for clinical viability. The deployment of these models in clinical workflows necessitates additional

external validation, as demonstrated in prior studies [183]. The significant role of data augmentation has been underscored by recent BraTS challenges [264]. Moreover, the extending nnU-Net, utilised in our study, has demonstrated performance comparable to recent state-of-the-art models [264], consistently contributing to winning ensemble strategies in the BraTS challenge since 2021 [256], [264]. Future research directions could focus on enhancing nnU-Net methodologies beyond GTV segmentation to facilitate comprehensive tumour region analysis, including ET, TC and WT. The clinical importance of U-Net variants [139], suggests that further refinement and adaptation could enhance segmentation accuracy across diverse datasets. Moreover, while a larger patch size for 3D models in our study was deliberately used to shorten training time, alternative configurations could further optimise performance.

In this thesis, the STORM_GLIO dataset was employed to optimise MRI data preprocessing for radiomic analysis while minimising deformation associated with resampling and registration steps. The pre-processed open-access and local datasets were used for radiomic analysis and automated tumour segmentation. Future improvements may involve the integration of additional clinical data, multi-omics data, and DL-derived features to further enhance radiomic model performance [269], [270]. Moreover, to improve reproducibility, an automated and computationally efficient model was developed, enabling the interchangeable use of GTV and TC segmentation while maintaining satisfactory segmentation accuracy. Future applications of the presented research extend to diverse neuro-oncological diseases, capitalising on foundational models [274] to enhance both segmentation precision and radiomic analytical frameworks. This integrative approach demonstrates significant potential to elevate model performance metrics, enhance cross-cohort generalisability [272], and facilitate the translation of research outcomes into clinical practice [265]. The findings of this research may contribute to the development of advanced clinical decision support systems, thereby advancing personalised therapeutic strategies for GBM and additional neuro-oncological disorders.

DL models continue to face challenges related to interpretability [69], [268], with ongoing efforts to improve transparency. This study shows that a traditional interpretable ML model can match the performance of DL methods. However, for clinical adoption, additional external validation and enhanced interpretability are required to ensure trust and applicability in medical practice [69], [267]. Additionally, reproducibility remains a critical concern in GBM radiomics research [71]. In addition to Combat harmonisation methods in radiomic research [83], physical phantom studies assessing the repeatability and reproducibility of RFs highlight the need for standardised methods in radiomic research for reliable clinical use across different clinical settings, demonstrating how acquisition settings and scanner differences affect feature stability [275], [276], [277], [278]. On the other hand, the recent introduction of more heterogeneous digital phantoms, specifically the ImSURE phantoms [279], facilitated an enhanced assessment of feature reproducibility via testing on five IBSI-standardised, open-access software packages. Analysis of the results indicated that only two software packages achieved a high percentage (>95%) of exact feature matches. DL-based methodologies, deep features, require more standardised applications and share generalisability issues with engineered RFs [70] while the IBSI contributed standardisation for ensuring the reproducibility of engineered RFs [236]. Additionally, studies such as TRIPOD [280] and METRICS [238] provided valuable tools for assessing radiomic research quality. Implementing rigorous statistical analysis, such as k-fold CV and bootstrap, further enhances the credibility and reliability of radiomics research [215], [230]. By helping to mitigate model overfitting, these approaches improve the generalisability of developed radiomic signatures to unseen data. The practical implementation of radiomic analysis in clinical settings is hindered by several key limitations, including a high dependency on ROIs [67]. One prominent example is the difficulty encountered in the automated tumour segmentation task of the BRATS challenge, specifically when dealing with MRI data acquired using low-quality parameters, as is often the case in regions with limited resources, such as Sub-Saharan Africa [226]. Besides lacking external validation, data quality challenges, interoperability limitations, incompatibility with established clinical workflows in both engineered and DL-based radiomics analysis, and the computational intensity

of DL-based radiomic analysis present substantial obstacles to its widespread clinical integration [265], [281]. Future research must, therefore, focus on developing radiomic solutions that address these challenges to support integration into standard clinical practice [265].

7. References

[1] Q. T. Ostrom, H. Gittleman, G. Truitt, A. Boscia, C. Kruchko, and J. S. Barnholtz-Sloan, 'CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2011-2015', *Neuro-Oncol.*, vol. 20, no. suppl_4, pp. iv1–iv86, Oct. 2018.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, 'Cancer statistics, 2019', *CA. Cancer J. Clin.*, vol. 69, no. 1, pp. 7–34, Jan. 2019.

[3] A. Verkhratsky, M. S. Ho, R. Zorec, and V. Parpura, 'The Concept of Neuroglia', in *NEUROGLIA IN NEURODEGENERATIVE DISEASES*, vol. 1175, A. Verkhratsky, M. S. Ho, R. Zorec, and V. Parpura, Eds, 2019, pp. 1–13.

[4] R. S. Snell, *Clinical Neuroanatomy*, 7th edn. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2010.

[5] M. Simons and K.-A. Nave, 'Oligodendrocytes: myelination and axonal support', *Cold Spring Harb. Perspect. Biol.*, vol. 8, no. 1, pp. a020479–a020479, 2016.

[6] M. N. Rasband, 'Glial contributions to neural function and disease', *Mol. Cell. Proteomics*, vol. 15, no. 2, pp. 355–361, 2016.

[7] F. He and Y. E. Sun, 'Glial cells more than support cells?', *Int. J. Biochem. Cell Biol.*, vol. 39, no. 4, pp. 661–665, 2007.

[8] M. Malhotra, A. Toulouse, B. M. D. C. Godinho, D. J. Mc Carthy, J. F. Cryan, and C. M. O'Driscoll, 'RNAi therapeutics for brain cancer: current advancements in RNAi delivery strategies', *Mol. Biosyst.*, vol. 11, no. 10, pp. 2635–2657, 2015.

[9] P. Wesseling, J. M. Kros, and J. W. M. Jeuken, 'The pathological diagnosis of diffuse gliomas: towards a smart synthesis of microscopic and molecular information in a multidisciplinary context', *Diagn. Histopathol.*, vol. 17, no. 11, pp. 486–494, 2011.

[10] H. S. Greenberg, W. F. Chandler, H. M. Sandler, and H. M. Sandler, *Brain Tumors*. Cary, UNITED STATES: Oxford University Press, Incorporated, 1999.

[11] B. D. Fox, V. J. Cheung, A. J. Patel, D. Suki, and G. Rao, 'Epidemiology of metastatic brain tumors', *Neurosurg. Clin.*, vol. 22, no. 1, pp. 1–6, 2011.

[12] D. N. Louis *et al.*, 'The 2021 WHO Classification of Tumors of the Central Nervous System: a summary', *Neuro-Oncol.*, vol. 23, no. 8, pp. 1231–1251, Aug. 2021.

[13] D. M. Gress *et al.*, 'Principles of cancer staging', *AJCC Cancer Staging Man.*, vol. 8, pp. 3–30, 2017.

[14] R. Stupp, 'European Organisation for Research and Treatment of Cancer Brain Tumor and Radiotherapy Groups; National Cancer Institute of Canada Clinical Trials Group, Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma', *N Engl J Med*, vol. 352, pp. 987–996, 2005.

[15] R. Stupp *et al.*, 'Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial', *Lancet Oncol.*, vol. 10, no. 5, pp. 459–466, May 2009.

[16] M. Price *et al.*, 'CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2017–2021', *Neuro-Oncol.*, vol. 26, no. Supplement_6, pp. vi1–vi85, 2024.

[17] M. T. C. Poon, C. L. M. Sudlow, J. D. Figueroa, and P. M. Brennan, 'Longer-term (≥ 2 years) survival in patients with glioblastoma in population-based studies pre-and post-2005: a systematic review and meta-analysis', *Sci. Rep.*, vol. 10, no. 1, pp. 11622–11622, 2020.

[18] M. Lara-Velazquez *et al.*, 'Advances in brain tumor surgery for glioblastoma in adults', *Brain Sci.*, vol. 7, no. 12, pp. 166–166, 2017.

[19] D. Garnier, O. Renoult, M.-C. Alves-Guerra, F. Paris, and C. Pecqueur, 'Glioblastoma Stem-Like Cells, Metabolic Strategy to Kill a Challenging Target', *Front. Oncol.*, vol. 9, 2019.

[20] D. E. Azagury *et al.*, 'Image-guided surgery', *Curr. Probl. Surg.*, vol. 52, no. 12, pp. 476–520, 2015.

[21] M. D. Jenkinson, D. G. Barone, M. G. Hart, A. Bryant, T. A. Lawrie, and C. Watts, 'Intraoperative imaging technology to maximise extent of resection for glioma', *Cochrane Database Syst. Rev.*, no. 9, 2017.

[22] C. Watts and N. Sanai, 'Surgical approaches for the gliomas', *Handb. Clin. Neurol.*, vol. 134, pp. 51–69, 2016.

[23] I. J. Gerard, M. Kersten-Oertel, K. Petrecca, D. Sirhan, J. A. Hall, and D. L. Collins, 'Brain shift in neuronavigation of brain tumors: A review', *Med. Image Anal.*, vol. 35, pp. 403–420, 2017.

[24] London: National Institute for Health and Care Excellence (NICE), 'Brain tumours (primary) and brain metastases in adults.', *NICE Guidel. No 99*, Jan. 2021.

[25] R. Stupp *et al.*, 'Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma', *N. Engl. J. Med.*, vol. 352, no. 10, pp. 987–996, 2005.

[26] V. M. Lu, T. R. Jue, and K. L. McDonald, 'Cytotoxic lanthanum oxide nanoparticles sensitize glioblastoma cells to radiation therapy and temozolomide: an in vitro rationale for translational studies', *Sci. Rep.*, vol. 10, no. 1, pp. 18156–18156, 2020.

[27] A. V. Krauze *et al.*, 'A Phase 2 Study of Concurrent Radiation Therapy, Temozolomide, and the Histone Deacetylase Inhibitor Valproic Acid for Patients With Glioblastoma', *Int. J. Radiat. Oncol.*, vol. 92, no. 5, pp. 986–992, 2015.

[28] A. R. Cabrera *et al.*, 'Radiation therapy for glioblastoma: Executive summary of an American Society for Radiation Oncology Evidence-Based Clinical Practice Guideline', *Pract. Radiat. Oncol.*, vol. 6, no. 4, pp. 217–225, July 2016.

[29] H. Murshed, *Fundamentals of radiation oncology: physical, biological, and clinical aspects*. Elsevier, 2024.

[30] S. Rockwell, I. T. Dobrucki, E. Y. Kim, S. T. Marrison, and V. T. Vu, 'Hypoxia and radiation therapy: past history, ongoing research, and future promise', *Curr. Mol. Med.*, vol. 9, no. 4, pp. 442–458, 2009.

[31] L. Juillerat-Jeanneret, 'The targeted delivery of cancer drugs across the blood–brain barrier: chemical modifications of drugs or drug-nanoparticles?', *Drug Discov. Today*, vol. 13, no. 23, pp. 1099–1106, 2008.

[32] D. Beier, J. B. Schulz, and C. P. Beier, 'Chemoresistance of glioblastoma cancer stem cells - much more complex than expected', *Mol. Cancer*, vol. 10, no. 1, pp. 128–128, 2011.

[33] R. C. Gimple, S. Bhargava, D. Dixit, and J. N. Rich, 'Glioblastoma stem cells: lessons from the tumor hierarchy in a lethal cancer', *Genes Dev.*, vol. 33, no. 11–12, pp. 591–609, 2019.

[34] B. Huang, X. Li, Y. Li, J. Zhang, Z. Zong, and H. Zhang, 'Current immunotherapies for glioblastoma multiforme', *Front. Immunol.*, vol. 11, pp. 603911–603911, 2021.

[35] Y. T. Lee, Y. J. Tan, and C. E. Oon, 'Molecular targeted therapy: Treating cancer with specificity', *Eur. J. Pharmacol.*, vol. 834, pp. 188–196, 2018.

[36] C. Mecca, I. Giambanco, R. Donato, and C. Arcuri, 'Targeting mTOR in glioblastoma: rationale and preclinical/clinical evidence', *Dis. Markers*, vol. 2018, no. 1, pp. 9230479–9230479, 2018.

[37] J. Rieger *et al.*, 'ERGO: A pilot study of ketogenic diet in recurrent glioblastoma Erratum in/ijo/45/6/2605', *Int. J. Oncol.*, vol. 44, no. 6, pp. 1843–1852, 2014.

[38] S. S. K. Yalamarty *et al.*, 'Mechanisms of resistance and current treatment options for glioblastoma multiforme (GBM)', *Cancers*, vol. 15, no. 7, pp. 2116–2116, 2023.

[39] NHS England, 'NHS England » Improving Outcomes through Personalised Medicine', Accessed: Dec. 01, 2024. [Online]. Available: https://www.england.nhs.uk/publication/improving-outcomes-through-personalised-medicine/

[40] A. K. Jha *et al.*, 'Radiomics: a quantitative imaging biomarker in precision oncology', *Nucl. Med. Commun.*, vol. 43, no. 5, pp. 483–493, 2022.

[41] I. Dagogo-Jack and A. T. Shaw, 'Tumour heterogeneity and resistance to cancer therapies', *Nat. Rev. Clin. Oncol.*, vol. 15, no. 2, pp. 81–94, 2018.

[42] N. Beig, K. Bera, and P. Tiwari, 'Introduction to radiomics and radiogenomics in neuro-oncology: implications and challenges', *Neuro-Oncol. Adv.*, vol. 2, no. Supplement_4, pp. iv3–iv14, Dec. 2020.

[43] C. Guy and D. ffytche, *An Introduction to the Principles of Medical Imaging*. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO., 2005, p. 420.

[44] E. Seeram, 'Computed tomography: physical principles and recent technical advances', *J. Med. Imaging Radiat. Sci.*, vol. 41, no. 2, pp. 87–109, 2010.

[45] C. Rincon-Guio, 'The role of computed tomography as a prognostic tool in traumatic brain trauma', *Hospital (Rio J.)*, 1971.

[46] M. Wasay *et al.*, 'Brain CT and MRI findings in 100 consecutive patients with intracranial tuberculoma', *J. Neuroimaging*, vol. 13, no. 3, pp. 240–247, 2003.

[47] D. E. Haines, M. A. (Mary A. Willis, and H. W. Lambert, *Neuroanatomy atlas in clinical context : structures, sections, systems, and syndromes*, 10th edition. Philadelphia ; Wolters Kluwer Health, 2019.

[48] T. Osborne, C. Tang, K. Sabarwal, and V. Prakash, 'How to interpret an unenhanced CT Brain scan. Part 1: Basic principles of Computed Tomography and relevant neuroanatomy', *South Sudan Med. J.*, vol. 9, no. 3, pp. 67–69, 2016.

[49] R. A. Pooley, 'AAPM/RSNA physics tutorial for residents - Fundamental physics of MR imaging', *RADIOGRAPHICS*, vol. 25, no. 4, pp. 1087–1099, 2005.

[50] K. E. Thomas, A. Fotaki, R. M. Botnar, and V. M. Ferreira, 'Imaging Methods: Magnetic Resonance Imaging', *Circ. Cardiovasc. Imaging*, vol. 16, no. 1, p. e014068, Jan. 2023.

[51] P. Suetens, *Fundamentals of medical imaging*. Cambridge university press, 2017.

[52] A. Berger, 'Magnetic resonance imaging', *BMJ*, vol. 324, no. 7328, pp. 35–35, Jan. 2002.

[53] A. Duman, O. Karakuş, X. Sun, S. Thomas, J. Powell, and E. Spezi, 'RFS+: A clinically adaptable and computationally efficient strategy for enhanced brain tumor segmentation', *Cancers*, vol. 15, no. 23, pp. 5620–5620, 2023.

[54] A. G. Rockall, A. Hatrick, P. Armstrong, and M. Wastie, *Diagnostic Imaging*. Hoboken, UNITED KINGDOM: John Wiley & Sons, Incorporated, 2013.

[55] T. C. Booth *et al.*, 'High-grade glioma treatment response monitoring biomarkers: a position statement on the evidence supporting the use of advanced MRI techniques in the clinic, and the latest bench-to-bedside developments. Part 2: spectroscopy, chemical exchange saturation, multiparametric imaging, and radiomics', *Front. Oncol.*, vol. 11, pp. 811425–811425, 2022.

[56] oncologymedicalphysics.com, 'MRI Design and Operation', [Online]. Available: https://oncologymedicalphysics.com/mri-design-and-operation/

[57] B. Taha *et al.*, 'Potential and limitations of radiomics in neuro-oncology.', *J. Clin. Neurosci.*, 2021.

[58] A. C. Tan, D. M. Ashley, G. Y. López, M. Malinzak, H. S. Friedman, and M. Khasraw, 'Management of glioblastoma: State of the art and future directions', *CA. Cancer J. Clin.*, vol. 70, no. 4, pp. 299–312, 2020.

[59] P. Lambin *et al.*, 'Radiomics: extracting more information from medical images using advanced feature analysis', *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[60] J. Müller Bark, A. Kulasinghe, B. Chua, B. W. Day, and C. Punyadeera, 'Circulating biomarkers in patients with glioblastoma', *Br. J. Cancer*, vol. 122, no. 3, pp. 295–305, 2020.

[61] R. J. Gillies, P. E. Kinahan, and H. Hricak, 'Radiomics: Images Are More than Pictures, They Are Data', *Radiology*, vol. 278, no. 2, pp. 563–577, Nov. 2015.

[62] A. Shaheen, S. T. Bukhari, M. Nadeem, S. Burigat, U. Bagci, and H. Mohy-ud-Din, 'Overall Survival Prediction of Glioma Patients With Multiregional Radiomics', *Front. Neurosci.*, vol. 16, 2022.

[63] X. Jia, L. Ren, and J. Cai, 'Clinical implementation of AI technologies will require interpretable AI models', *Med. Phys.*, no. 1, pp. 1–4, 2020.

[64] B. H. M. Van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, 'Explainable artificial intelligence (XAI) in deep learning-based medical image analysis', *Med. Image Anal.*, vol. 79, pp. 102470–102470, 2022.

[65] R. Berenguer *et al.*, 'Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters', *Radiology*, vol. 288, no. 2, pp. 407–415, 2018.

[66] A. Zwanenburg *et al.*, 'The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping', *Radiology*, vol. 295, no. 2, pp. 328–338, Mar. 2020.

[67] J. E. van Timmeren *et al.*, 'Radiomics in medical imaging-"how-to" guide and critical reflection.', *Insights Imaging*, 2020.

[68] A. Depeursinge, O. S. Al-Kadi, and J. R. Mitchell, *Biomedical texture analysis: fundamentals, tools and challenges*. Academic Press, 2017.

[69] P. Martin *et al.*, 'Challenges in Glioblastoma Radiomics and the Path to Clinical Implementation', *Cancers*, 2022.

[70] A. Carré *et al.*, 'Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics.', *Sci. Rep.*, 2020.

[71] H. Moradmand, S. M. R. Aghamiri, and R. Ghaderi, 'Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma.', *J. Appl. Clin. Med. Phys.*, 2020.

[72] R. T. Shinohara *et al.*, 'Statistical normalization techniques for magnetic resonance imaging', *NeuroImage Clin.*, vol. 6, pp. 9–19, 2014.

[73] S. Pati *et al.*, 'The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview', presented at the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, A. Crimi and S. Bakas, Eds, Cham: Springer International Publishing, 2020, pp. 380–394.

[74] C. Davatzikos *et al.*, 'Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome', *J. Med. Imaging*, vol. 5, no. 1, pp. 011018–011018, Jan. 2018.

[75] F. Isensee *et al.*, 'Automated brain extraction of multisequence MRI using artificial neural networks', *Hum. Brain Mapp.*, vol. 40, no. 17, pp. 4952–4964, Dec. 2019.

[76] S. M. Smith, 'Fast robust automated brain extraction', *Hum. Brain Mapp.*, vol. 17, no. 3, pp. 143–155, 2002.

[77] R. W. Cox, 'AFNI: software for analysis and visualization of functional magnetic resonance neuroimages', *Comput. Biomed. Res.*, vol. 29, no. 3, pp. 162–173, 1996.

[78] M. Larobina and L. Murino, 'Medical Image File Formats', *J. Digit. Imaging*, vol. 27, no. 2, pp. 200–206, 2014.

[79] B. M. Pacheco, G. de S. e Cassia, and D. Silva, 'Towards fully automated deep-learning-based brain tumor segmentation: Is brain extraction still necessary?', *Biomed. Signal Process. Control*, 2023.

[80] K. Fatania *et al.*, 'Intensity standardization of MRI prior to radiomic feature extraction for artificial intelligence research in glioma—a systematic review', *Eur. Radiol.*, 2022.

[81] N. J. Tustison *et al.*, 'N4ITK: Improved N3 Bias Correction', *IEEE Trans. Med. Imaging*, 2010.

[82] B. Baeßler, K. Weiss, and D. Pinto dos Santos, 'Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study', *Invest. Radiol.*, vol. 54, no. 4, 2019.

[83] E. Stamoulou, G. C. Manikis, M. Tsiknakis, and K. Marias, 'ComBat harmonization for multicenter MRI based radiomics features', 2021.

[84] R. N. Mahon, M. Ghita, G. D. Hugo, and E. Weiss, 'ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets', *Phys. Med. Biol.*, vol. 65, no. 1, pp. 015010–015010, 2020.

[85] R. T. H. Leijenaar *et al.*, 'The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis', *Sci. Rep.*, vol. 5, no. 1, pp. 11075–11075, 2015.

[86] M. Ghaffari, A. Sowmya, and R. Oliver, 'Automated Brain Tumor Segmentation Using Multimodal Brain Scans: A Survey Based on Models Submitted to the BraTS 2012–2018 Challenges', *IEEE Rev. Biomed. Eng.*, vol. 13, pp. 156–168, 2020.

[87] Y. Bengio, A. Courville, and P. Vincent, 'Representation learning: A review and new perspectives', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[88] O. Ronneberger, P. Fischer, and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation', *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 9351, pp. 234–241, 2015.

[89] W. J. Zabel *et al.*, 'Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring of Bladder and Rectum for Prostate Radiation Therapy', *Pract. Radiat. Oncol.*, vol. 11, no. 1, pp. e80–e89, 2021.

[90] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, 'Brain tumor segmentation using convolutional neural networks in MRI images', *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[91] D. Jones, 'ICRU Report 50—Prescribing, Recording and Reporting Photon Beam Therapy', *Med. Phys.*, vol. 21, no. 6, pp. 833–834, June 1994.

[92] A. Duman, P. Whybra, J. Powell, S. Thomas, X. Sun, and E. Spezi, 'PO-1620 Transferability of deep learning models to the segmentation of gross tumour volume in brain cancer', *Radiother. Oncol.*, vol. 182, pp. S1315–S1316, 2023.

[93] P. Y. Wen, S. M. Chang, M. J. Van den Bent, M. A. Vogelbaum, D. R. Macdonald, and E. Q. Lee, 'Response Assessment in Neuro-Oncology Clinical Trials', *J. Clin. Oncol.*, vol. 35, no. 21, pp. 2439–2449, June 2017.

[94] K. Kamnitsas *et al.*, 'Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation', *Med. Image Anal.*, vol. 36, pp. 61–78, 2017.

[95] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, 'nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation', *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[96] Y. Gao, Y. Jiang, Y. Peng, F. Yuan, X. Zhang, and J. Wang, 'Medical Image Segmentation: A Comprehensive Review of Deep Learning-Based Methods', *Tomography*, vol. 11, no. 5, pp. 52–52, 2025.

[97] X. Zhang, J. Wang, J. Wei, X. Yuan, and M. Wu, 'A Review of Non-Fully Supervised Deep Learning for Medical Image Segmentation', 2025.

[98] G. Qu *et al.*, 'Motion-artifact-augmented pseudo-label network for semi-supervised brain tumor segmentation', *Phys. Med. Biol.*, vol. 69, no. 5, pp. 055023–055023, 2024.

[99] H. Chen, J. An, B. Jiang, L. Xia, Y. Bai, and Z. Gao, 'WS-MTST: Weakly supervised multi-label brain tumor segmentation with transformers', *IEEE J. Biomed. Health Inform.*, vol. 27, no. 12, pp. 5914–5925, 2023.

[100] W. H. L. Pinaya *et al.*, 'Unsupervised brain imaging 3D anomaly detection and segmentation with transformers', *Med. Image Anal.*, vol. 79, p. 102475, July 2022.

[101] N. Otsu, 'A Threshold Selection Method from Gray-Level Histograms', *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[102] D. L. Pham, C. Xu, and J. L. Prince, 'Current Methods in Medical Image Segmentation1', *Annu. Rev. Biomed. Eng.*, vol. 2, no. Volume 2, 2000, pp. 315–337, Aug. 2000.

[103] M. Kass, A. Witkin, and D. Terzopoulos, 'Snakes: Active contour models', *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[104] B. Jiang *et al.*, 'Deep learning for brain tumor segmentation in multimodal MRI images: A review of methods and advances', *Image Vis. Comput.*, vol. 156, p. 105463, Apr. 2025.

[105] Q. Xia *et al.*, 'A comprehensive review of deep learning for medical image segmentation', *Neurocomputing*, vol. 613, p. 128740, Jan. 2025.

[106] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, 'U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications', *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[107] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, '3D U-Net: learning dense volumetric segmentation from sparse annotation', presented at the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 424–432.

[108] F. Milletari, N. Navab, and S.-A. Ahmadi, 'V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation', in *2016 Fourth International Conference on 3D Vision (3DV)*, Oct. 2016, pp. 565–571.

[109] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[110] O. Oktay *et al.*, 'Attention u-net: Learning where to look for the pancreas', *ArXiv Prepr. ArXiv180403999*, 2018.

[111] C. Zhou, C. Ding, X. Wang, Z. Lu, and D. Tao, 'One-Pass Multi-Task Networks With Cross-Task Guided Attention for Brain Tumor Segmentation', *IEEE Trans. Image Process.*, vol. 29, pp. 4516–4529, 2020.

[112] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, 'SA-UNet: Spatial Attention U-Net for Retinal Vessel Segmentation', in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp. 1236–1242.

[113] H. Huang *et al.*, 'Channel prior convolutional attention for medical image segmentation', *Comput. Biol. Med.*, vol. 178, p. 108784, Aug. 2024.

[114] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, 'Densely Connected Convolutional Networks', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 2261–2269.

[115] L. Yu *et al.*, 'Automatic 3D Cardiovascular MR Segmentation with Densely-Connected Volumetric ConvNets', in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds, Cham: Springer International Publishing, 2017, pp. 287–295.

[116] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, 'H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes', *IEEE Trans. Med. Imaging*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[117] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, 'UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation', *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, June 2020.

[118] H. Huang *et al.*, 'UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation', in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 1055–1059.

[119] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, 'DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[120] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, 'DenseASPP for Semantic Segmentation in Street Scenes', in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 3684–3692.

[121] X. Zhao, L. Zhang, and H. Lu, 'Automatic Polyp Segmentation via Multi-scale Subtraction Network', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds, Cham: Springer International Publishing, 2021, pp. 120–130.

[122] X. Zhao *et al.*, 'M^{2}SNet: Multi-scale in Multi-scale Subtraction Network for Medical Image Segmentation', Mar. 20, 2023, *arXiv*: arXiv:2303.10894.

[123] A. Vaswani *et al.*, 'Attention is all you need', *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[124] D. Karimi, S. D. Vasylechko, and A. Gholipour, 'Convolution-Free Medical Image Segmentation Using Transformers', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds, Cham: Springer International Publishing, 2021, pp. 78–88.

[125] Z. Liu *et al.*, 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9992–10002.

[126] H. Cao *et al.*, 'Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation', in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds, Cham: Springer Nature Switzerland, 2023, pp. 205–218.

[127] J. Chen *et al.*, 'TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation', Feb. 08, 2021, *arXiv*: arXiv:2102.04306.

[128] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, 'nnFormer: Interleaved Transformer for Volumetric Segmentation', Feb. 04, 2022, *arXiv*: arXiv:2109.03201.

[129] Y. Zhang, H. Liu, and Q. Hu, 'TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds, Cham: Springer International Publishing, 2021, pp. 14–24.

[130] F. Yuan, Z. Zhang, and Z. Fang, 'An effective CNN and Transformer complementary network for medical image segmentation', *Pattern Recognit.*, vol. 136, p. 109228, Apr. 2023.

[131] A. Hatamizadeh *et al.*, 'UNETR: Transformers for 3D Medical Image Segmentation', in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2022, pp. 1748–1758.

[132] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. Shahbaz Khan, 'UNETR++: Delving Into Efficient and Accurate 3D Medical Image Segmentation', *IEEE Trans. Med. Imaging*, vol. 43, no. 9, pp. 3377–3390, Sept. 2024.

[133] A. Gu and T. Dao, 'Mamba: Linear-Time Sequence Modeling with Selective State Spaces', May 31, 2024, *arXiv*: arXiv:2312.00752.

[134] Y. Liu *et al.*, 'VMamba: Visual State Space Model', Dec. 29, 2024, *arXiv*: arXiv:2401.10166.

[135] J. Ruan, J. Li, and S. Xiang, 'VM-UNet: Vision Mamba UNet for Medical Image Segmentation', Nov. 08, 2024, *arXiv*: arXiv:2402.02491.

[136] W. Liao, Y. Zhu, X. Wang, C. Pan, Y. Wang, and L. Ma, 'LightM-UNet: Mamba Assists in Lightweight UNet for Medical Image Segmentation', Mar. 11, 2024, *arXiv*: arXiv:2403.05246.

[137] J. Liu *et al.*, 'Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining', Mar. 06, 2024, *arXiv*: arXiv:2402.03302.

[138] J. Ma, F. Li, and B. Wang, 'U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation', Jan. 09, 2024, *arXiv*: arXiv:2401.04722.

[139] F. Isensee *et al.*, 'nnu-net revisited: A call for rigorous validation in 3d medical image segmentation', presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 488–498.

[140] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, 'SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation', Sept. 15, 2024, *arXiv*: arXiv:2401.13560.

[141] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, 'No New-Net', in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds, Cham: Springer International Publishing, 2019, pp. 234–244.

[142] M. Noori, A. Bahri, and K. Mohammadi, 'Attention-Guided Version of 2D UNet for Automatic Brain Tumor Segmentation', in *2019 9TH INTERNATIONAL CONFERENCE ON COMPUTER AND KNOWLEDGE ENGINEERING (ICCKE 2019)*, New York: IEEE, 2019, pp. 269–275.

[143] S. Alqazzaz, X. Sun, X. Yang, and L. Nokes, 'Automated brain tumor segmentation on multi-modal MR image using SegNet', *Comput. Vis. Media*, vol. 5, no. 2, pp. 209–219, June 2019.

[144] D. Maji, P. Sigedar, and M. Singh, 'Attention Res-UNet with Guided Decoder for semantic segmentation of brain tumors', *Biomed. Signal Process. Control*, vol. 71, pp. 103077–103077, 2022.

[145] R. Rajaragavi and S. Rajan, 'Optimized U-Net Segmentation and Hybrid Res-Net for Brain Tumor MRI Images Classification', *Intell. Autom. Soft Comput.*, vol. 32, no. 1, pp. 1–14, 2021.

[146] T. Ruba, R. Tamilselvi, and M. Parisa Beham, 'Brain tumor segmentation using JGate-AttResUNet – A novel deep learning approach', *Biomed. Signal Process. Control*, vol. 84, p. 104926, July 2023.

[147] N. Cinar, A. Ozcan, and M. Kaya, 'A hybrid DenseNet121-UNet model for brain tumor segmentation from MR Images', *Biomed. Signal Process. Control*, vol. 76, pp. 103647–103647, 2022.

[148] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, 'Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images', in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds, Cham: Springer International Publishing, 2022, pp. 272–284.

[149] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, 'TransBTS: Multimodal Brain Tumor Segmentation Using Transformer', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds, Cham: Springer International Publishing, 2021, pp. 109–119.

[150] Y. Xie, J. Zhang, C. Shen, and Y. Xia, 'CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S.

Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds, Cham: Springer International Publishing, 2021, pp. 171–180.

[151] J. Chen *et al.*, 'TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers', *Med. Image Anal.*, vol. 97, p. 103280, Oct. 2024.

[152] H. M. Luu and S.-H. Park, 'Extending nn-UNet for Brain Tumor Segmentation', presented at the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, A. Crimi and S. Bakas, Eds, Cham: Springer International Publishing, 2022, pp. 173–186.

[153] R. Zhou, J. Wang, G. Xia, J. Xing, H. Shen, and X. Shen, 'Cascade Residual Multiscale Convolution and Mamba-Structured UNet for Advanced Brain Tumor Image Segmentation', *Entropy*, vol. 26, no. 5, Art. no. 5, May 2024.

[154] L. Yang, Q. Dong, D. Lin, C. Tian, and X. Lü, 'MUNet: a novel framework for accurate brain tumor segmentation combining UNet and mamba networks', *Front. Comput. Neurosci.*, vol. 19, Jan. 2025.

[155] J. Xu, Y. Lan, Y. Zhang, C. Zhang, S. Stirenko, and H. Li, 'CDA-mamba: cross-directional attention mamba for enhanced 3D medical image segmentation', *Sci. Rep.*, vol. 15, no. 1, p. 21357, July 2025.

[156] H. Zhang *et al.*, 'A Survey on Visual Mamba', *Appl. Sci.*, vol. 14, no. 13, Art. no. 13, Jan. 2024.

[157] G. Litjens *et al.*, 'A survey on deep learning in medical image analysis', *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[158] F. Gao *et al.*, 'Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images', *Med. Image Anal.*, vol. 80, p. 102515, 2022.

[159] T. Islam, Md. S. Hafiz, J. R. Jim, Md. M. Kabir, and M. F. Mridha, 'A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions', *Healthc. Anal.*, vol. 5, p. 100340, June 2024.

[160] A. K. Upadhyay and A. K. Bhandari, 'Advances in Deep Learning Models for Resolving Medical Image Segmentation Data Scarcity Problem: A Topical Review', *Arch. Comput. Methods Eng.*, vol. 31, no. 3, pp. 1701–1719, Apr. 2024.

[161] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, 'Federated learning for medical image analysis: A survey', *Pattern Recognit.*, vol. 151, p. 110424, July 2024.

[162] L. El Jiani, S. El Filali, and E. H. Benlahmer, 'Overcome medical image data scarcity by data augmentation techniques: A review', in *2022 International Conference on Microelectronics (ICM)*, Dec. 2022, pp. 21–24.

[163] E. Christodoulou *et al.*, 'Confidence Intervals Uncovered: Are We Ready for Real-World Medical Imaging AI?', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, Eds, Cham: Springer Nature Switzerland, 2024, pp. 124–132.

[164] D. Sharma, Z. Shanis, C. K. Reddy, S. Gerber, and A. Enquobahrie, 'Active Learning Technique for Multimodal Brain Tumor Segmentation Using Limited Labeled Images', in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, Q. Wang, F. Milletari, H. V. Nguyen, S. Albarqouni, M. J. Cardoso, N. Rieke, Z. Xu, K. Kamnitsas, V. Patel, B. Roysam, S. Jiang, K. Zhou, K. Luu, and N. Le, Eds, Cham: Springer International Publishing, 2019, pp. 148–156.

[165] Q. Yang, R. Jing, and J. Mu, 'Multi-modal MR image segmentation strategy for brain tumors based on domain adaptation', *Computers*, vol. 13, no. 12, p. 347, 2024.

[166] L. Dai, T. Li, H. Shu, L. Zhong, H. Shen, and H. Zhu, 'Automatic Brain Tumor Segmentation with Domain Adaptation', in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds, Cham: Springer International Publishing, 2019, pp. 380–392.

[167] X. Du *et al.*, 'UM-Net: Rethinking ICGNet for polyp segmentation with uncertainty modeling', *Med. Image Anal.*, vol. 99, p. 103347, Jan. 2025.

[168] H. G. Pemberton *et al.*, 'Multi-class glioma segmentation on real-world data with missing MRI sequences: comparison of three deep learning algorithms', *Sci. Rep.*, vol. 13, no. 1, pp. 18911–18911, 2023.

[169] L. Zang *et al.*, 'A deep learning model based on Mamba for automatic segmentation in cervical cancer brachytherapy', *Sci. Rep.*, vol. 15, no. 1, p. 10152, Mar. 2025.

[170] Y. Hu *et al.*, 'Comparative analysis of U-Mamba and no new U-Net for the detection and segmentation of esophageal cancer in contrast-enhanced computed tomography images', *Quant. Imaging Med. Surg. Vol 15 No 3 March 03 2025 Quant. Imaging Med. Surg.*, 2025.

[171] P. M. Kazaj *et al.*, 'From Claims to Evidence: A Unified Framework and Critical Analysis of CNN vs. Transformer vs. Mamba in Medical Image Segmentation', Mar. 03, 2025, *arXiv*: arXiv:2503.01306.

[172] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, 'Medical image segmentation using deep learning: A survey', *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, Apr. 2022.

[173] S. Maqsood, R. Damasevicius, and F. M. Shah, 'An Efficient Approach for the Detection of Brain Tumor Using Fuzzy Logic and U-NET CNN Classification', presented at the Computational Science and Its Applications – ICCSA 2021, O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, and C. M. Torre, Eds, Cham: Springer International Publishing, 2021, pp. 105–118.

[174] Z. Liu, Q. Lv, Z. Yang, Y. Li, C. H. Lee, and L. Shen, 'Recent progress in transformer-based medical image analysis', *Comput. Biol. Med.*, vol. 164, pp. 107268–107268, 2023.

[175] L. Qin, H. Zhao, S. Zhang, and Z. Tang, 'Automated brain tumor segmentation using cascaded bootstrapping model', presented at the Eleventh International Conference on Graphics and Image Processing (ICGIP 2019), SPIE, 2020, pp. 356–360.

[176] A. Beers, K. Chang, J. Brown, E. Gerstner, B. Rosen, and J. Kalpathy-Cramer, 'Sequential neural networks for biologically informed glioma segmentation', presented at the Medical Imaging 2018: Image Processing, SPIE, 2018, pp. 807–816.

[177] N. Hashemi, S. Masoudnia, A. Nejad, and M.-R. Nazem-Zadeh, 'A Memory-efficient Deep Framework for Multi-Modal MRI-based Brain Tumor Segmentation', presented at the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2022, pp. 3749–3752.

[178] U. Baid *et al.*, 'The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification', *ArXiv Prepr. ArXiv210702314*, 2021.

[179] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, 'Automatic Brain Tumor Segmentation Using Cascaded Anisotropic Convolutional Neural Networks', presented at the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, Eds, Cham: Springer International Publishing, 2018, pp. 178–190.

[180] A. Myronenko, '3D MRI Brain Tumor Segmentation Using Autoencoder Regularization', presented at the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds, Cham: Springer International Publishing, 2019, pp. 311–320.

[181] Z. Jiang, C. Ding, M. Liu, and D. Tao, 'Two-Stage Cascaded U-Net: 1st Place Solution to BraTS Challenge 2019 Segmentation Task', presented at the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, A. Crimi and S. Bakas, Eds, Cham: Springer International Publishing, 2020, pp. 231–241.

[182] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, 'nnU-net for brain tumor segmentation', presented at the International MICCAI Brainlesion Workshop, Springer, 2020, pp. 118–132.

[183] A. Berkley *et al.*, 'Clinical capability of modern brain tumor segmentation models', *Med. Phys.*, vol. 50, no. 8, pp. 4943–4959, Aug. 2023.

[184] D. Leithner *et al.*, 'ComBat Harmonization for MRI Radiomics: Impact on Nonbinary Tissue Classification by Machine Learning', *Invest. Radiol.*, vol. 58, no. 9, 2023.

[185] P. Whybra, C. Parkinson, K. Foley, J. Staffurth, and E. Spezi, 'Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging', *Sci. Rep.*, vol. 9, no. 1, pp. 9649–9649, 2019.

[186] C. Piazzese, K. Foley, P. Whybra, C. Hurt, T. Crosby, and E. Spezi, 'Discovery of stable and prognostic CT-based radiomic features independent of contrast administration and dimensionality in oesophageal cancer', *PLoS One*, vol. 14, no. 11, pp. e0225550–e0225550, 2019.

[187] W. A. Noortman *et al.*, 'Multicollinearity and redundancy of the PET radiomic feature set', *Eur. Radiol.*, May 2025.

[188] Y. Suter *et al.*, 'Radiomics for glioblastoma survival analysis in pre-operative MRI: exploring feature robustness, class boundaries, and machine learning techniques', *Cancer Imaging*, vol. 20, no. 1, p. 55, Aug. 2020.

[189] F. Mandreoli, D. Ferrari, V. Guidetti, F. Motta, and P. Missier, 'Real-world data mining meets clinical practice: Research challenges and perspective', *Front. Big Data*, vol. 5, 2022.

[190] X. Zhao *et al.*, 'Deep learning signatures reveal multiscale intratumor heterogeneity associated with biological functions and survival in recurrent nasopharyngeal carcinoma', *Eur. J. Nucl. Med. Mol. Imaging*, vol. 49, no. 8, pp. 2972–2982, July 2022.

[191] W. Zhang, Y. Guo, and Q. Jin, 'Radiomics and its feature selection: a review', *Symmetry*, vol. 15, no. 10, p. 1834, 2023.

[192] Y. Wang, Z. Hu, and H. Wang, 'The clinical implications and interpretability of computational medical imaging (radiomics) in brain tumors', *Insights Imaging*, vol. 16, no. 1, p. 77, Mar. 2025.

[193] Q. Wan *et al.*, 'Comparative analysis of deep learning and radiomic signatures for overall survival prediction in recurrent high-grade glioma treated with immunotherapy', *Cancer Imaging*, vol. 25, no. 1, p. 5, 2025.

[194] A. Demircioğlu, 'Are deep models in radiomics performing better than generic models? A systematic review', *Eur. Radiol. Exp.*, vol. 7, no. 1, p. 11, Mar. 2023.

[195] A. Demircioğlu, 'Reproducibility and interpretability in radiomics: a critical assessment', *Diagn. Interv. Radiol.*, July 2025.

[196] M. R. Tomaszewski and R. J. Gillies, 'The biological meaning of radiomic features', *Radiology*, vol. 298, no. 3, pp. 505–516, 2021.

[197] A. Perniciano, A. Loddo, C. Di Ruberto, and B. Pes, 'Insights into radiomics: impact of feature selection and classification', *Multimed. Tools Appl.*, vol. 84, no. 26, pp. 31695–31721, Aug. 2025.

[198] H. Moradmand *et al.*, 'Graph feature selection for enhancing radiomic stability and reproducibility across multiple institutions in head and neck cancer', *Sci. Rep.*, vol. 15, no. 1, p. 27995, July 2025.

[199] H. Xu, C. Caramanis, and S. Mannor, 'Sparse algorithms are not stable: A no-free-lunch theorem', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 187–193, 2011.

[200] H. Zou and T. Hastie, 'Regularization and variable selection via the elastic net', *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[201] R. Muthukrishnan and R. Rohini, 'LASSO: A feature selection technique in predictive modeling for machine learning', in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Oct. 2016, pp. 18–20.

[202] K. Koyama, K. Kiritoshi, T. Okawachi, and T. Izumitani, 'Effective Nonlinear Feature Selection Method based on HSIC Lasso and with Variational Inference', in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, PMLR, May 2022, pp. 10407–10421.

[203] F. Li, Y. Yang, and E. Xing, 'From Lasso regression to Feature vector machine', in *Advances in Neural Information Processing Systems*, MIT Press, 2005.

[204] M. Yuan and Y. Lin, 'Model selection and estimation in regression with grouped variables', *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, 2006.

[205] S. Katoch, S. S. Chauhan, and V. Kumar, 'A review on genetic algorithm: past, present, and future', *Multimed. Tools Appl.*, vol. 80, no. 5, pp. 8091–8126, Feb. 2021.

[206] X. Song, Y. Zhang, D. Gong, and X. Sun, 'Feature selection using bare-bones particle swarm optimization with mutual information', *Pattern Recognit.*, vol. 112, p. 107804, 2021.

[207] T. M. Shami, A. A. El-Saleh, M. Alswaitti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, 'Particle swarm optimization: A comprehensive survey', *Ieee Access*, vol. 10, pp. 10031–10061, 2022.

[208] Q. Al-Tashi *et al.*, 'SwarmDeepSurv: swarm intelligence advances deep survival network for prognostic radiomics signatures in four solid cancers', *Patterns*, vol. 4, no. 8, 2023.

[209] X. Pan, C. Liu, T. Feng, and X. S. Qi, 'A multi-objective based radiomics feature selection method for response prediction following radiotherapy', *Phys. Med. Biol.*, vol. 68, no. 5, pp. 055018–055018, 2023.

[210] D. T. Do, M.-R. Yang, L. H. T. Lam, N. Q. K. Le, and Y.-W. Wu, 'Improving MGMT methylation status prediction of glioblastoma through optimizing radiomics features using genetic algorithm-based machine learning approach', *Sci. Rep.*, vol. 12, no. 1, pp. 13412–13412, 2022.

[211] B. H. Menze *et al.*, 'The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)', *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[212] S. Bakas *et al.*, 'Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features', *Sci. Data*, vol. 4, no. 1, pp. 170117–170117, 2017.

[213] S. Bakas *et al.*, 'Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge', 2018.

[214] S. Cepeda *et al.*, 'The Río Hortega University Hospital Glioblastoma dataset: A comprehensive collection of preoperative, early postoperative and recurrence MRI scans (RHUH-GBM)', *Data Brief*, vol. 50, pp. 109617–109617, 2023.

[215] A. Rabasco Meneghetti *et al.*, 'Definition and validation of a radiomics signature for loco-regional tumour control in patients with locally advanced head and neck squamous cell carcinoma', *Clin. Transl. Radiat. Oncol.*, vol. 26, pp. 62–70, 2021.

[216] R. Poursaeed, M. Mohammadzadeh, and A. A. Safaei, 'Survival prediction of glioblastoma patients using machine learning and deep learning: a systematic review', *BMC Cancer*, vol. 24, no. 1, pp. 1581–1581, 2024.

[217] A. Gomaa *et al.*, 'Comprehensive multimodal deep learning survival prediction enabled by a transformer architecture: A multicenter study in glioblastoma', *Neuro-Oncol. Adv.*, vol. 6, no. 1, pp. vdae122–vdae122, Jan. 2024.

[218] M. Verduin *et al.*, 'Prognostic and predictive value of integrated qualitative and quantitative magnetic resonance imaging analysis in glioblastoma', *Cancers*, vol. 13, no. 4, pp. 1–20, Feb. 2021.

[219] A. Fathi Kazerooni *et al.*, 'Clinical measures, radiomics, and genomics offer synergistic value in AI-based prediction of overall survival in patients with glioblastoma', *Sci. Rep.*, vol. 12, no. 1, pp. 8784–8784, 2022.

[220] A. Demircioğlu, 'Benchmarking feature selection methods in radiomics', *Invest. Radiol.*, vol. 57, no. 7, pp. 433–443, 2022.

[221] M. Rostami, K. Berahmand, E. Nasiri, and S. Forouzandeh, 'Review of swarm intelligence-based feature selection methods', *Eng. Appl. Artif. Intell.*, vol. 100, pp. 104210–104210, 2021.

[222] A. G. Tzalavra, I. Andreadis, K. V Dalakleidi, F. Constantinidis, E. I Zacharaki, and K. S Nikita, 'Dynamic contrast enhanced-magnetic resonance imaging radiomics combined with a hybrid adaptive neuro-fuzzy inference system-particle swarm optimization approach for breast tumour classification', *Expert Syst.*, vol. 39, no. 4, pp. e12895–e12895, 2022.

[223] T. Rohlfing, N. M. Zahr, E. V. Sullivan, and A. Pfefferbaum, 'The SRI24 multichannel atlas of normal adult human brain structure', *Hum. Brain Mapp.*, vol. 31, no. 5, pp. 798–819, 2010.

[224] S. Bakas, C. Sako, H. Akbari, M. Bilello, A. Sotiras, and G. Shukla, 'Multi-parametric magnetic resonance imaging (mpMRI) scans for de novo Glioblastoma (GBM) patients from the University of Pennsylvania Health System (UPENN-GBM)', *Cancer Imaging Arch. TCIA Public Access*, 2021.

[225] S. Thakur *et al.*, 'Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training', *Neuroimage*, vol. 220, pp. 117081–117081, 2020.

[226] M. Adewole *et al.*, 'The brain tumor segmentation (brats) challenge 2023: glioma segmentation in sub-saharan Africa patient population (brats-africa)', *ArXiv*, 2023.

[227] S. Bakas *et al.*, 'BraTS 2024 Cluster of Challenges (BraTS + Beyond-BraTS)', *Zenodo*, Apr. 2024.

[228] M. Aboian *et al.*, 'MICCAI 2025 Lighthouse Challenge: Brain Tumor Segmentation Cluster of Challenges (BraTS)', *Zenodo*, Oct. 2024.

[229] R. Tibshirani, 'The lasso method for variable selection in the Cox model', *Stat. Med.*, vol. 16, no. 4, pp. 385–395, 1997.

[230] S. Leger *et al.*, 'A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling', *Sci. Rep.*, vol. 7, no. 1, pp. 13206–13206, 2017.

[231] J. H. Holland, 'Genetic algorithms', *Sci. Am.*, vol. 267, no. 1, pp. 66–73, 1992.

[232] J. Kennedy and R. Eberhart, 'Particle swarm optimization', presented at the Proceedings of ICNN'95-international conference on neural networks, ieee, 1995, pp. 1942–1948.

[233] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, 'Random survival forests', *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, Sept. 2008.

[234] O. W. Mwangi, A. Islam, and O. Luke, 'Bootstrap Confidence Intervals for Proportions of Unequal Sized Groups Adjusted for Overdispersion', *Open J. Stat.*, vol. 5, no. 6, pp. 502–510, 2015.

[235] J.-C. Luo, Q.-Y. Zhao, and G.-W. Tu, 'Clinical prediction models in the precision medicine era: old and new algorithms', *Ann. Transl. Med.*, vol. 8, no. 6, 2020.

[236] P. Whybra *et al.*, 'The Image Biomarker Standardization Initiative: Standardized Convolutional Filters for Reproducible Radiomics and Enhanced Clinical Insights', *Radiology*, vol. 310, no. 2, pp. e231319–e231319, Feb. 2024.

[237] X. Zhang *et al.*, 'Deep Learning With Radiomics for Disease Diagnosis and Treatment: Challenges and Potential', *Front. Oncol.*, vol. 12, 2022.

[238] B. Kocak *et al.*, 'METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII', *Insights Imaging*, vol. 15, no. 1, pp. 8–8, 2024.

[239] F. Tixier, V. Jaouen, C. Hognon, O. Gallinato, T. Colin, and D. Visvikis, 'Evaluation of conventional and deep learning based image harmonization methods in radiomics studies', *Phys. Med. Biol.*, vol. 66, no. 24, pp. 245009–245009, 2021.

[240] S. Cepeda *et al.*, 'Predicting Short-Term Survival after Gross Total or Near Total Resection in Glioblastomas by Machine Learning-Based Radiomic Analysis of Preoperative MRI', *Cancers*, vol. 13, no. 20, 2021.

[241] R. Verma *et al.*, 'Stable and Discriminatory Radiomic Features from the Tumor and Its Habitat Associated with Progression-Free Survival in Glioblastoma: A Multi-Institutional Study', *Am. J. Neuroradiol.*, vol. 43, no. 8, pp. 1115–1115, Aug. 2022.

[242] G. Hajianfar *et al.*, 'Time-to-event overall survival prediction in glioblastoma multiforme patients using magnetic resonance imaging radiomics', *Radiol. Med. (Torino)*, vol. 128, no. 12, pp. 1521–1534, 2023.

[243] M. Tabassum, A. A. Suman, E. Suero Molina, E. Pan, A. Di Ieva, and S. Liu, 'Radiomics and Machine Learning in Brain Tumors and Their Habitat: A Systematic Review', *Cancers*, vol. 15, no. 15, 2023.

[244] J. Zhang *et al.*, 'Fully automatic classification of breast lesions on multi-parameter MRI using a radiomics model with minimal number of stable, interpretable features', *Radiol. Med. (Torino)*, vol. 128, no. 2, pp. 160–170, 2023.

[245] A. Duman, J. Powell, S. Thomas, X. Sun, and E. Spezi, 'Generalizability of Deep Learning Models on Brain Tumour Segmentation', Cardiff University Press, 2024.

[246] A. Zwanenburg *et al.*, 'Assessing robustness of radiomic features by image perturbation', *Sci. Rep.*, vol. 9, no. 1, pp. 614–614, Jan. 2019.

[247] J.-H. Kim, 'Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap', *Comput. Stat. Data Anal.*, vol. 53, no. 11, pp. 3735–3745, 2009.

[248] I.-K. Yeo and R. A. Johnson, 'A New Family of Power Transformations to Improve Normality or Symmetry', *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.

[249] B. Gierlichs, L. Batina, P. Tuyls, and B. Preneel, 'Mutual information analysis: A generic side-channel distinguisher', presented at the International Workshop on Cryptographic Hardware and Embedded Systems, Springer, 2008, pp. 426–442.

[250] F. Long, H. Peng, and C. Ding, 'Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 08, pp. 1226–1238, 2005.

[251] A. N. Kamarudin, T. Cox, and R. Kolamunnage-Dona, 'Time-dependent ROC curve analysis in medical research: current methods and applications', *BMC Med. Res. Methodol.*, vol. 17, no. 1, pp. 53–53, 2017.

[252] P. Whybra and E. Spezi, 'Sensitivity of standardised radiomics algorithms to mask generation across different software platforms', *Sci. Rep.*, vol. 13, no. 1, pp. 14419–14419, 2023.

[253] T. Magadza and S. Viriri, 'Deep learning for brain tumor segmentation: a survey of state-of-the-art', *J. Imaging*, vol. 7, no. 2, pp. 19–19, 2021.

[254] A. Casamitjana, S. Puch, A. Aduriz, and V. Vilaplana, '3D convolutional neural networks for brain tumor segmentation: A comparison of multi-resolution architectures', presented at the International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer, 2016, pp. 150–161.

[255] S. Hussain, S. M. Anwar, and M. Majid, 'Brain tumor segmentation using cascaded deep convolutional neural network', presented at the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, pp. 1998–2001.

[256] R. A. Zeineldin, M. E. Karar, O. Burgert, and F. Mathis-Ullrich, 'Multimodal CNN Networks for Brain Tumor Segmentation in MRI: A BraTS 2022 Challenge Solution', presented at the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, S. Bakas, A. Crimi, U. Baid, S. Malec, M. Pytlarz, B. Baheti, M. Zenk, and R. Dorent, Eds, Cham: Springer Nature Switzerland, 2023, pp. 127–137.

[257] T. Henry *et al.*, 'Brain Tumor Segmentation with Self-ensembled, Deeply-Supervised 3D U-Net Neural Networks: A BraTS 2020 Challenge Solution', presented at the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, A. Crimi and S. Bakas, Eds, Cham: Springer International Publishing, 2021, pp. 327–339.

[258] J. C. Reinhold *et al.*, 'Evaluating the Impact of Intensity Normalization on MR Image Synthesis.', *Image Process.*, 2019.

[259] M. J. Cardoso *et al.*, 'Monai: An open-source framework for deep learning in healthcare', *ArXiv Prepr. ArXiv221102701*, 2022.

[260] D. P. Kingma and J. Ba, 'Adam: A method for stochastic optimization', *ArXiv Prepr. ArXiv14126980*, 2014.

[261] S. Gwynne *et al.*, 'Toward Semi-automated Assessment of Target Volume Delineation in Radiotherapy Trials: The SCOPE 1 Pretrial Test Case', *Int. J. Radiat. Oncol.*, vol. 84, no. 4, pp. 1037–1042, 2012.

[262] S. M. H. Hoque *et al.*, 'Clinical use of a commercial artificial intelligence-based software for autocontouring in radiation therapy: geometric performance and dosimetric impact', *Cancers*, vol. 15, no. 24, pp. 5735–5735, 2023.

[263] H. Baroudi *et al.*, 'Automated contouring and planning in radiation therapy: what is "clinically acceptable"?', *Diagnostics*, vol. 13, no. 4, pp. 667–667, 2023.

[264] A. Ferreira *et al.*, 'How we won brats 2023 adult glioma challenge? just faking it! enhanced synthetic data augmentation and model ensemble for brain tumour segmentation', *ArXiv Prepr. ArXiv240217317*, 2024.

[265] N. Horvat, N. Papanikolaou, and D.-M. Koh, 'Radiomics beyond the hype: a critical evaluation toward oncologic clinical use', *Radiol. Artif. Intell.*, vol. 6, no. 4, pp. e230437–e230437, 2024.

[266] Y. Li *et al.*, 'Impact of Preprocessing and Harmonization Methods on the Removal of Scanner Effects in Brain MRI Radiomic Features.', *Cancers*, 2021.

[267] X. Li, X. T. Li, R. Y. Huang, and R. Y. Huang, 'Standardization of imaging methods for machine learning in neuro-oncology.', 2020.

[268] J. Lao *et al.*, 'A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme', *Sci. Rep.*, vol. 7, no. 1, pp. 10353–10353, Sept. 2017.

[269] N. Braman, J. W. H. Gordon, E. T. Goossens, C. Willis, M. C. Stumpe, and J. Venkataraman, 'Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data', presented at the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer, 2021, pp. 667–677.

[270] W. Kang *et al.*, 'Application of radiomics-based multiomics combinations in the tumor microenvironment and cancer prognosis', *J. Transl. Med.*, vol. 21, no. 1, pp. 598–598, 2023.

[271] D. Veiga-Canuto *et al.*, 'Comparative multicentric evaluation of inter-observer variability in manual and automatic segmentation of neuroblastic tumors in magnetic resonance images', *Cancers*, vol. 14, no. 15, pp. 3648–3648, 2022.

[272] M. Yanzhen, C. Song, L. Wanping, Y. Zufang, and A. Wang, 'Exploring approaches to tackle cross-domain challenges in brain medical image segmentation: a systematic review', *Front. Neurosci.*, vol. 18, pp. 1401329–1401329, 2024.

[273] Q. Yang, R. Jing, and J. Mu, 'Multi-Modal MR Image Segmentation Strategy for Brain Tumors Based on Domain Adaptation', *Computers*, vol. 13, no. 12, pp. 347–347, 2024.

[274] J. Cox *et al.*, 'BrainSegFounder: towards 3D foundation models for neuroimage segmentation', *Med. Image Anal.*, vol. 97, pp. 103301–103301, 2024.

[275] A. K. Jha *et al.*, 'Repeatability and reproducibility study of radiomic features on a phantom and human cohort', 2021.

[276] J. Lee *et al.*, 'Radiomics feature robustness as measured using an MRI phantom', *Sci. Rep.*, vol. 11, no. 1, pp. 3973–3973, 2021.

[277] M. Sun *et al.*, 'Robustness and reproducibility of radiomics in T2 weighted images from magnetic resonance image guided linear accelerator in a phantom study', *Phys. Med.*, vol. 96, pp. 130–139, 2022.

[278] E.-N. Cheong, J. E. Park, S. Y. Park, S. C. Jung, and H. S. Kim, 'Achieving imaging and computational reproducibility on multiparametric MRI radiomics

features in brain tumor diagnosis: Phantom and clinical validation', *Eur. Radiol.*, vol. 34, no. 3, pp. 2008–2023, 2024.

[279] A. Bettinelli, F. Marturano, A. Sarnelli, A. Bertoldo, and M. Paiusco, 'The ImSURE phantoms: a digital dataset for radiomic software benchmarking and investigation', *Sci. Data*, vol. 9, no. 1, pp. 695–695, 2022.

[280] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement', *Circulation*, vol. 131, no. 2, pp. 211–219, 2015.

[281] S. Maleki Varnosfaderani and M. Forouzanfar, 'The role of AI in hospitals and clinics: transforming healthcare in the 21st century', *Bioengineering*, vol. 11, no. 4, pp. 337–337, 2024.

[282] A. Shrestha, A. Watkins, F. Yousefirizi, A. Rahmim, and C. F. Uribe, 'RT-utils: A Minimal Python Library for RT-struct Manipulation', *ArXiv Prepr. ArXiv240506184*, 2024.

[283] T. Phil, T. Albrecht, S. Gay, and M. E. Rasmussen, 'Sikerdebaard/dcmrtstruct2nii: dcmrtstruct2nii v5 (Version v5)', 2023.

[284] N. J. Tustison *et al.*, 'The ANTsX ecosystem for quantitative biological and medical imaging', *Sci. Rep.*, vol. 11, no. 1, pp. 9068–9068, 2021.

[285] A. Fedorov *et al.*, '3D Slicer as an image computing platform for the Quantitative Imaging Network', *Magn. Reson. Imaging*, vol. 30, no. 9, pp. 1323–1341, 2012.

[286] A. Rabasco Meneghetti, 'Radiogenomics machine learning analyses for treatment personalization of locally advanced head and neck squamous cell carcinoma', Jan. 2024, Accessed: Aug. 01, 2025. [Online]. Available: https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-891066

8. Appendix

A

I. Skull stripping was achieved using HD-BET (Version 1.0).

II. Image conversion between NIfTI and DICOM, including DICOM RTSTRUCT, formats was handled by rt-utils (Version 1.2.7) [282] and dicom2nifti (Version 2.4.10), Dcmrtstruct2nii (Version v5) [283].

III. Rigid registration was performed using ants (Version 0.0.7) [284]. Rigid registration was performed using ants (Version 0.0.7) [284].

IV. Furthermore, an example of automated tumour segmentation was generated using a custom-trained U-Net model [92].

V. For clinical reusability, the tumour segmentation (GTV/TC [92]) generated by this pipeline was converted back into DICOM format (RTSTRUCT).

VI. For visual comparison, 3D slicer v5.6.2 [285]. The clinical relevance of the pipeline results was confirmed through clinician approval in a clinical setting. For visual comparison, 3D slicer v5.6.2 [285]. The clinical relevance of the pipeline results was confirmed through clinician approval in a clinical setting.

VII. The complete, Python-based alternative workflow is available at [https://github.com/krmdmn/preprocesing_pipeline].

B

IBSI standardised parameters

```
{
 "interpolation": {
   "new_voxel_spacing": [1, 1, 1],
   "method": "spline",
   "rounding_after_interp": true
 },
 "feature_families": ["morph", "stats", "ih", "ivh", "glcm", "glrlm", "glszm", "ngldm", "ngtdm", "gldzm"],
 "re_segmentation_range": {"min": "", "max": ""},
 "re_segmentation_outlier_filtering": {"apply": false, "sigma": 3},
 "bin_method": "FBN",
 "bin_value": 64,
 "analysis_type": "3D",
 "texture_parameters": {
   "glcm": {"aggregation": "merged", "distance": 1},
   "glrlm": {"aggregation": "merged", "distance": 1},
   "ngtdm": {"distance": 1},
   "ngldm": {"distance": 1, "alpha": 0}
 }
}
```

Figure B- 1 The IBSI standardised preprocessing parameters for radiomic analysis.

LASSO-RANK Feature Selection

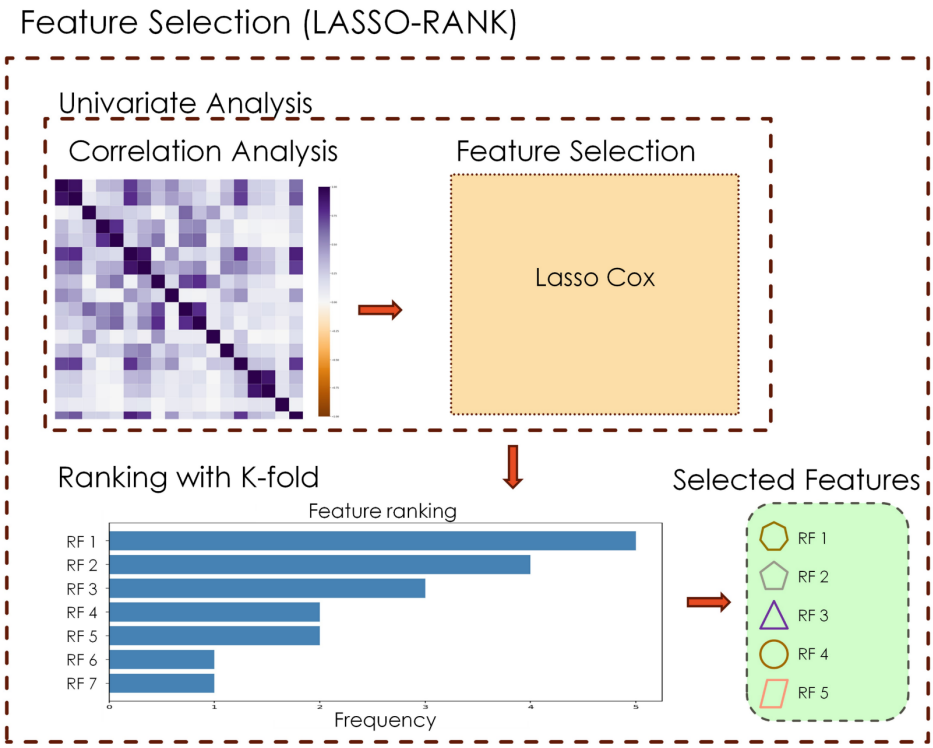


Figure B- 2 The LASSO-RANK feature selection framework.

PSA and GA feature selection Parameters

```
A)For PSO algorithm:
ParticleSwarmFeatureSelectionCV(
        n_particles=30,
        estimator=Lasso,
        cv=3,
        scoring="neg_mean_squared_error",
        max_iter=10,
        n_jobs=-1,
        verbosity=0    )
B)For GA algorithm:
GeneticSelectionCV(
        estimator= Lasso,
        cv=5,
        verbose=1,
        scoring="neg_mean_squared_error",
        max_features=9,
        n_population=50,
        crossover_proba=0.5,
        mutation_proba=0.2,
        n_generations=40,
        crossover_independent_proba=0.5,
        mutation_independent_proba=0.05,
        tournament_size=3,
        n_gen_no_change=10,
        caching=True,
        n_jobs=-1,    )
```

Figure B- 3 Hyperparameters for (A) PSO and (B) GA were set based on the example source codes, with the exception of 'max_features', which was adjusted to 9 (the total feature number for each feature subset) for GA, leading to improved model performance.

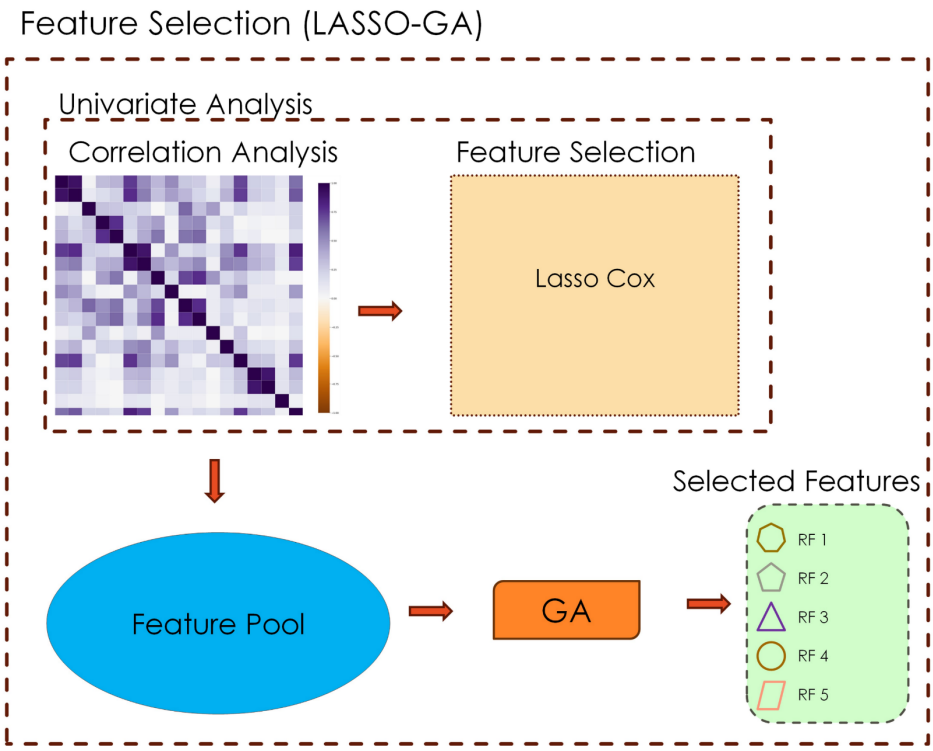


Figure B- 4 The LASSO-GA feature selection framework.

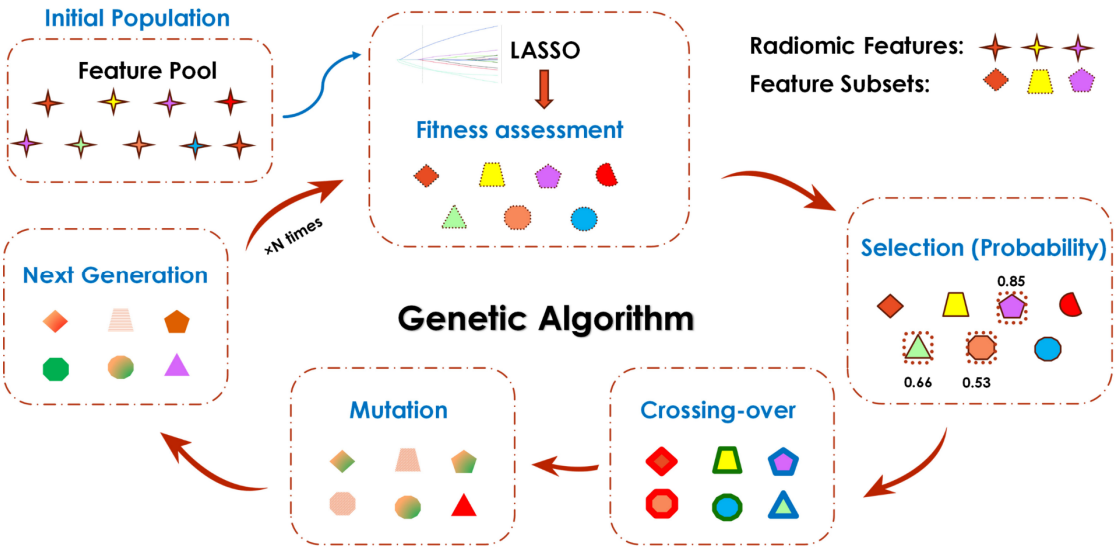


Figure B- 5 GA-based Feature Selection Workflow with the feature pool from LASSO.

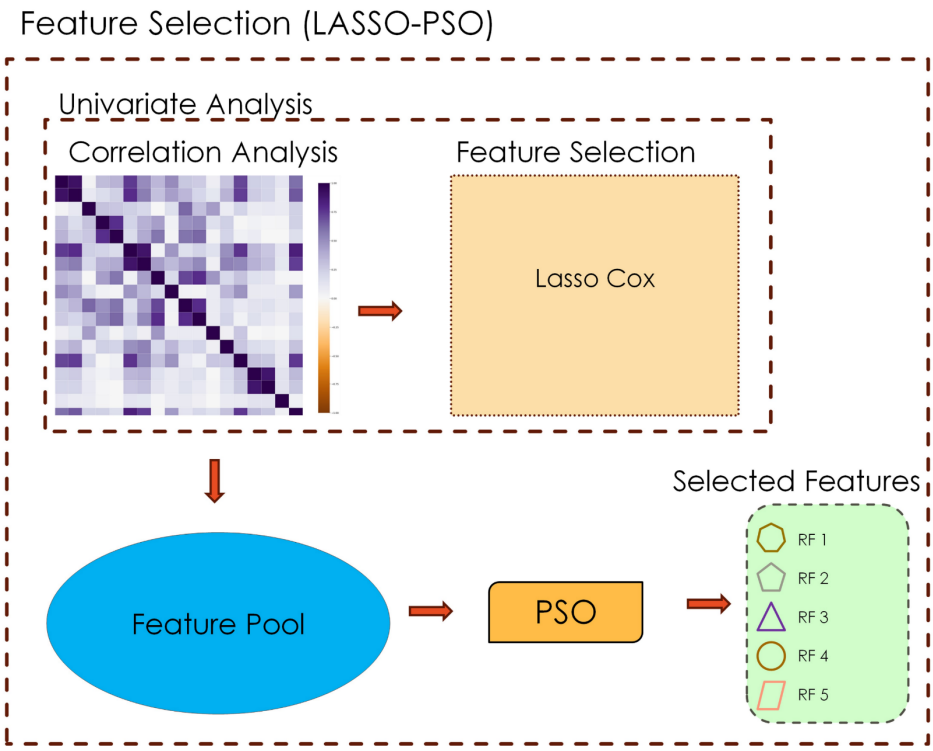


Figure B- 6 The LASSO-PSO feature selection framework.

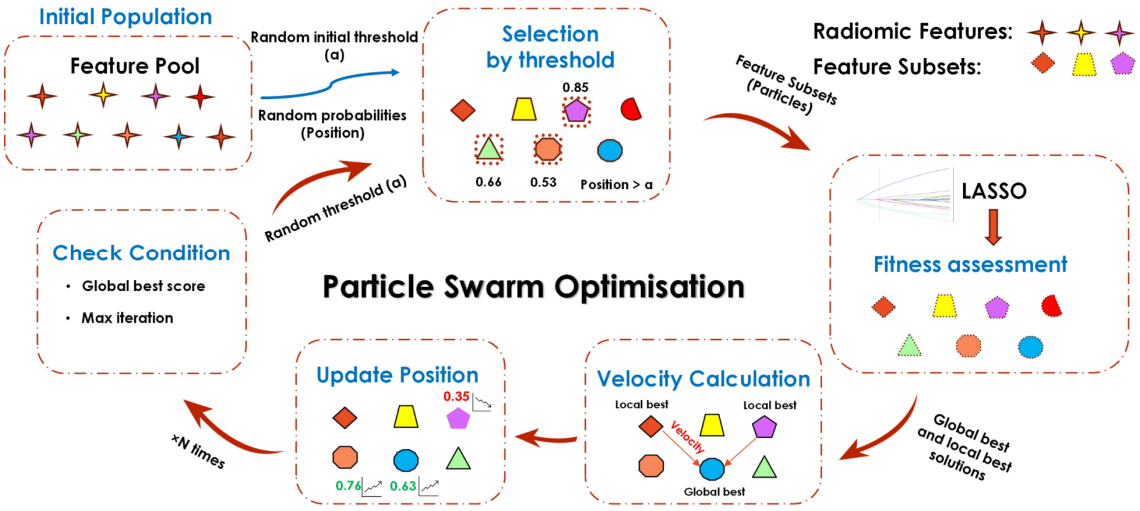


Figure B- 7 PSO-based feature selection workflow with the feature pool from LASSO.

Table B- 1 The hyperparameters for each model and feature selection method from 200 bootstrapped iterations of the training (discovery) dataset.

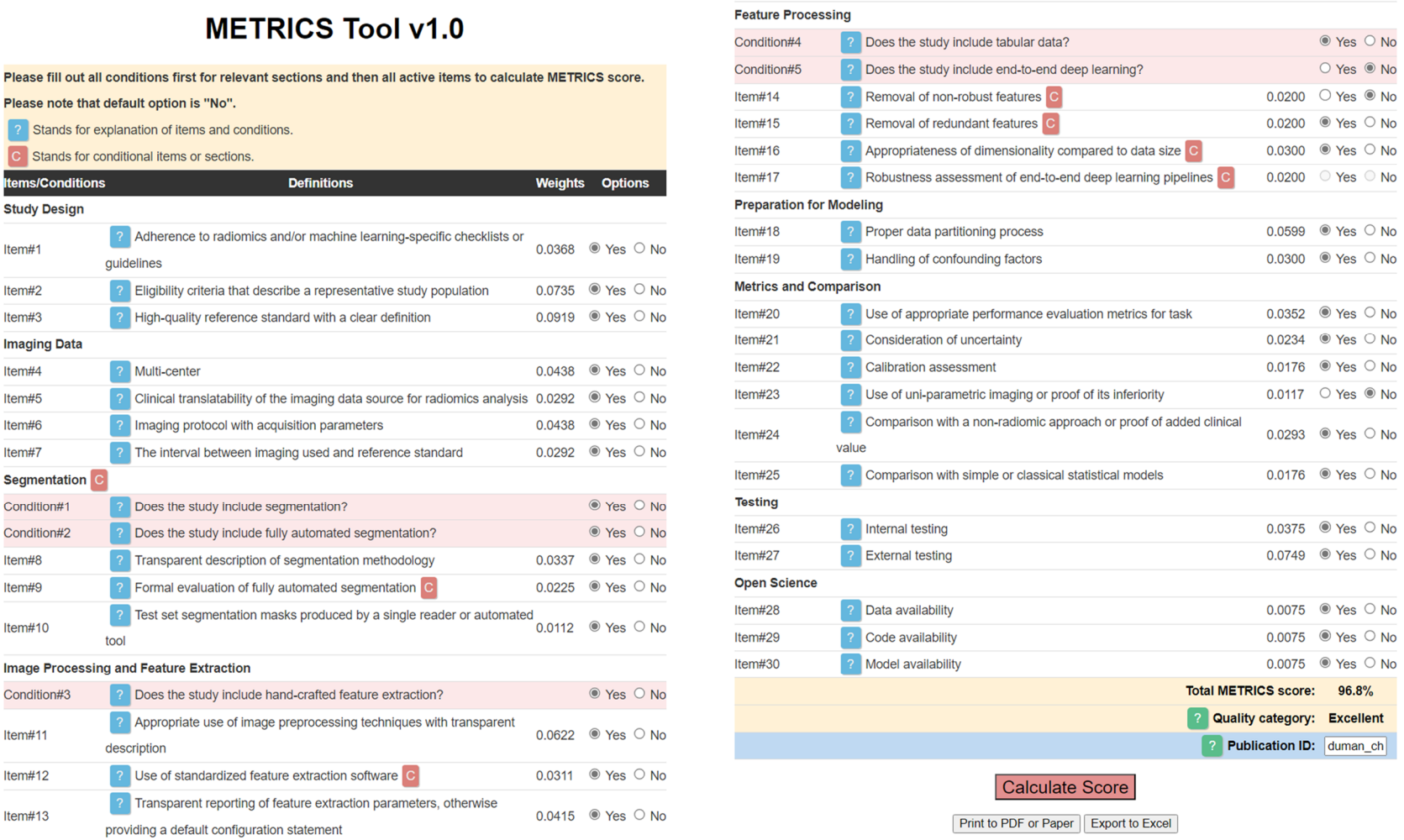
Model	Feature Selection	Hyperparameters (selected [min, max])
Cox-LASSO	LASSO-RANK	alpha = 50 [1,100]
Cox-LASSO	LASSO-GA	alpha = 100 [1,100]
Cox-LASSO	LASSO-PSO	alpha = 5 [1,100]
RSF	LASSO-RANK	n_estimators = 10 [3,10] max_depth = 5 [3,5] min_samples_split =3 [3,10]
RSF	LASSO-GA	n_estimators = 5 [3,10] max_depth = 5 [3,5] min_samples_split =10 [3,10]
RSF	LASSO-PSO	n_estimators = 10 [3,10] max_depth = 4 [3,5] min_samples_split =10 [3,10]

Table B- 2 The RSF model performance on the discovery, the hold-out test and the external validation.

		C-Index		
Model	Feature Selection Method	Discovery Cohort	Hold-out Test Cohort	External Validation Cohort
RSF	LASSO-PSO	0.74	0.63	0.58

Table B- 3 The feature Importance and weights of each feature in the final clinical-radiomic model.

LASSO-PSO, Selected Features	Permutation Importance (Feature Importance)	Feature Weight
morph_pca_maj_axis	0.034	0.21
morph_pca_flatness	0.011	0.16
morph_comp_1	0.012	-0.11
morph_vol_dens_aee	0.004	-0.06
morph_area_dens_aee	0.003	0.04
ngl_dc_entr_3D	0.011	-0.14
dzm_zdnu_3D	0.016	0.13
szm_lgze_3D (ET label)	0.015	-0.16
szm_lgze_3D (TC label)	0.001	-0.01
stat_skew	0.014	-0.13
Age	0.067	0.31
KM curve cut-off value		0.012 (median)



METRICS Tool v1.0

Please fill out all conditions first for relevant sections and then all active items to calculate METRICS score.

Please note that default option is "No".

? Stands for explanation of items and conditions.

C Stands for conditional items or sections.

Items/Conditions	Definitions	Weights	Options
Study Design			
Item#1	Adherence to radiomics and/or machine learning-specific checklists or guidelines	0.0368	Yes / No
Item#2	Eligibility criteria that describe a representative study population	0.0735	Yes / No
Item#3	High-quality reference standard with a clear definition	0.0919	Yes / No
Imaging Data			
Item#4	Multi-center	0.0438	Yes / No
Item#5	Clinical translatability of the imaging data source for radiomics analysis	0.0292	Yes / No
Item#6	Imaging protocol with acquisition parameters	0.0438	Yes / No
Item#7	The interval between imaging used and reference standard	0.0292	Yes / No
Segmentation C			
Condition#1	Does the study include segmentation?		Yes / No
Condition#2	Does the study include fully automated segmentation?		Yes / No
Item#8	Transparent description of segmentation methodology	0.0337	Yes / No
Item#9	Formal evaluation of fully automated segmentation C	0.0225	Yes / No
Item#10	Test set segmentation masks produced by a single reader or automated tool	0.0112	Yes / No
Image Processing and Feature Extraction			
Condition#3	Does the study include hand-crafted feature extraction?		Yes / No
Item#11	Appropriate use of image preprocessing techniques with transparent description	0.0622	Yes / No
Item#12	Use of standardized feature extraction software C	0.0311	Yes / No
Item#13	Transparent reporting of feature extraction parameters, otherwise providing a default configuration statement	0.0415	Yes / No
Feature Processing			
Condition#4	Does the study include tabular data?		Yes / No
Condition#5	Does the study include end-to-end deep learning?		Yes / No
Item#14	Removal of non-robust features C	0.0200	Yes / No
Item#15	Removal of redundant features C	0.0200	Yes / No
Item#16	Appropriateness of dimensionality compared to data size C	0.0300	Yes / No
Item#17	Robustness assessment of end-to-end deep learning pipelines C	0.0200	Yes / No
Preparation for Modeling			
Item#18	Proper data partitioning process	0.0599	Yes / No
Item#19	Handling of confounding factors	0.0300	Yes / No
Metrics and Comparison			
Item#20	Use of appropriate performance evaluation metrics for task	0.0352	Yes / No
Item#21	Consideration of uncertainty	0.0234	Yes / No
Item#22	Calibration assessment	0.0176	Yes / No
Item#23	Use of uni-parametric imaging or proof of its inferiority	0.0117	Yes / No
Item#24	Comparison with a non-radiomic approach or proof of added clinical value	0.0293	Yes / No
Item#25	Comparison with simple or classical statistical models	0.0176	Yes / No
Testing			
Item#26	Internal testing	0.0375	Yes / No
Item#27	External testing	0.0749	Yes / No
Open Science			
Item#28	Data availability	0.0075	Yes / No
Item#29	Code availability	0.0075	Yes / No
Item#30	Model availability	0.0075	Yes / No

Total METRICS score: 96.8%

Quality category: Excellent

Publication ID: duman_ch

Calculate Score

Print to PDF or Paper | Export to Excel

Figure B- 8 METRICS for the radiomic study.

C


```
 "interpolation": {
  "new_voxel_spacing": [1, 1, 1],
  "method": "spline",
  "rounding_after_interp": true
 },
 "feature_families": ["morph", "stats", "ih", "ivh", "glcm", "glrlm", "glszm", "ngldm",
"ngtdm", "gldzm"],
 "re_segmentation_range": {"min": "", "max": ""},
 "re_segmentation_outlier_filtering": {"apply": false, "sigma": 3},
 "bin_method": "FBN",
 "bin_value": 64,
 "analysis_type": "3D",
 "texture_parameters": {
  "glcm": {"aggregation": "merged", "distance": 1},
  "glrlm": {"aggregation": "merged", "distance": 1},
  "ngtdm": {"distance": 1},
  "ngldm": {"distance": 1, "alpha": 0}
 }
}
```


Figure C- 1 Settings in IBSI-compliant terminology for radiomics analysis carried out with the SPAARC code.

Table C- 1 The selected hyperparameters settings from 200 bootstrapped iterations of the training dataset.

Model	Feature Selection	Hyperparameters (selected [min, max])
Cox-LASSO	MutInfo	alpha = 2 [1,5]
Cox-LASSO	mRMR	alpha = 2 [1,5]
Cox-LASSO	Lasso	alpha = 2 [1,5]
GBS	MutInfo	n_estimators = 2 [1,5] max_depth = 2 [1,5] min_samples_split =2 [1,5]
GBS	mRMR	n_estimators = 2 [1,5] max_depth = 2 [1,5] min_samples_split = 2 [1,5]
GBS	Lasso	n_estimators =2 [1,5] max_depth = 2 [1,5] min_samples_split = 2 [1,5]
RSF	MutInfo	n_estimators = 2 [1,5] max_depth = 2 [1,5] min_samples_split = 2 [1,5]
RSF	mRMR	n_estimators = 2 [1,5] max_depth = 2 [1,5] min_samples_split = 2 [1,5]
RSF	Lasso	n_estimators = 2 [1,5] max_depth = 2 [1,5] min_samples_split = 2 [1,5]

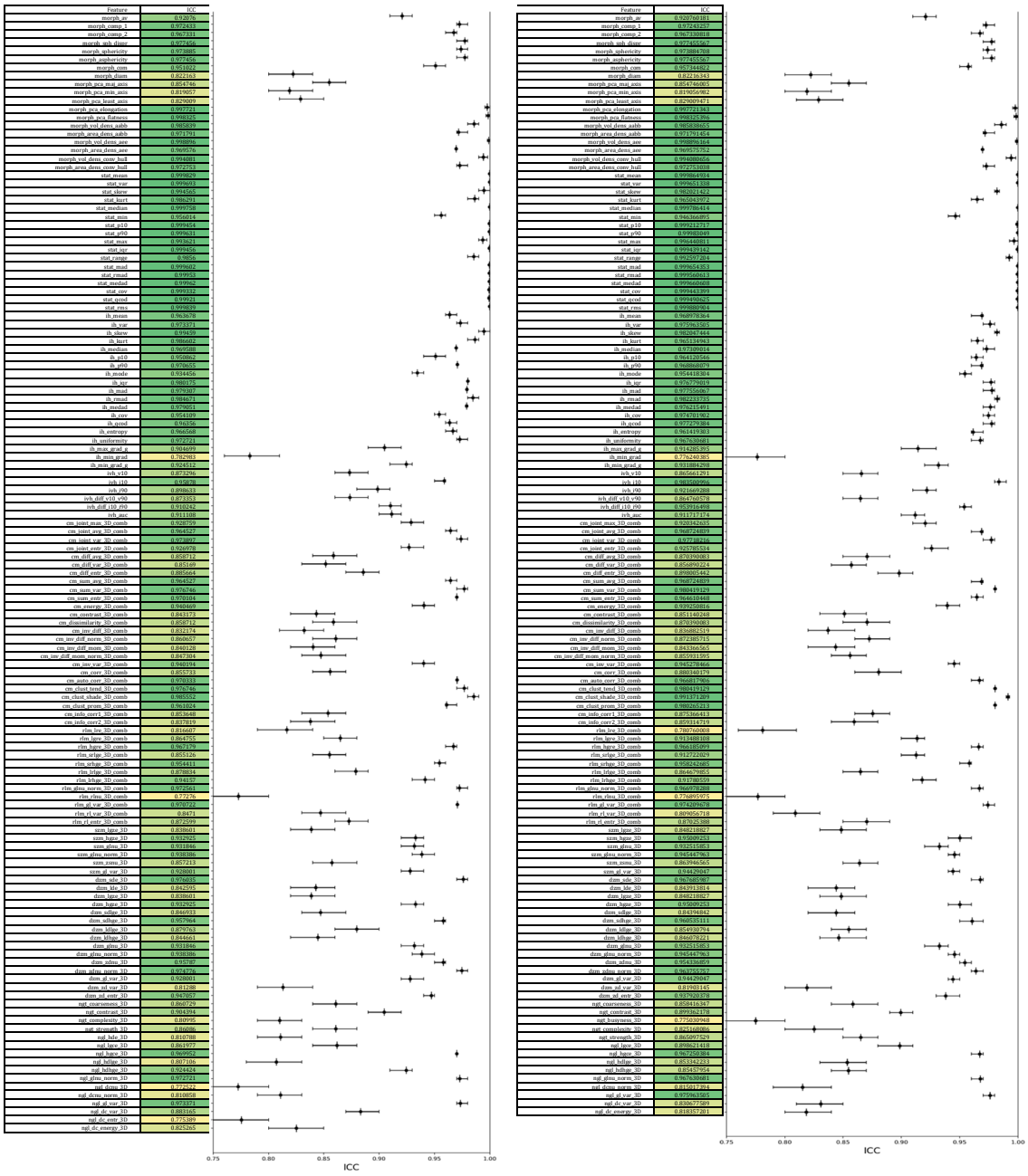


Figure C- 2 The analysis of feature robustness to image perturbation the ICC result with 95% confidence intervals. (a) Robust RFs derived from MRI T2 sequence (b) Robust RFs derived from MRI FLAIR sequence.

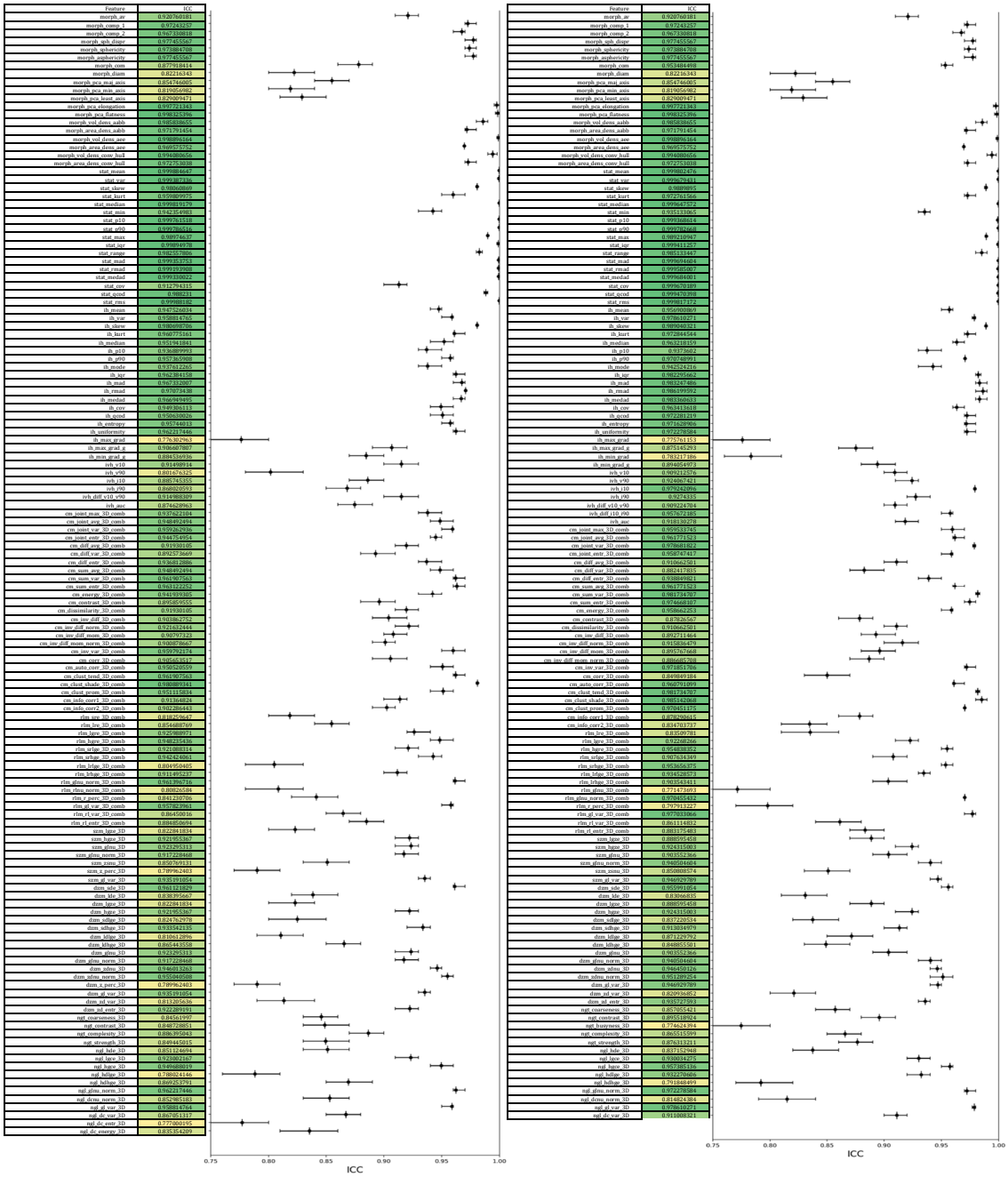


Figure C- 3 The analysis of feature robustness to image perturbation the ICC result with 95% confidence intervals (ICC>0.75). (a) Robust RFs derived from MRI T1 sequence (b) Robust RFs derived from MRI T1ce sequence

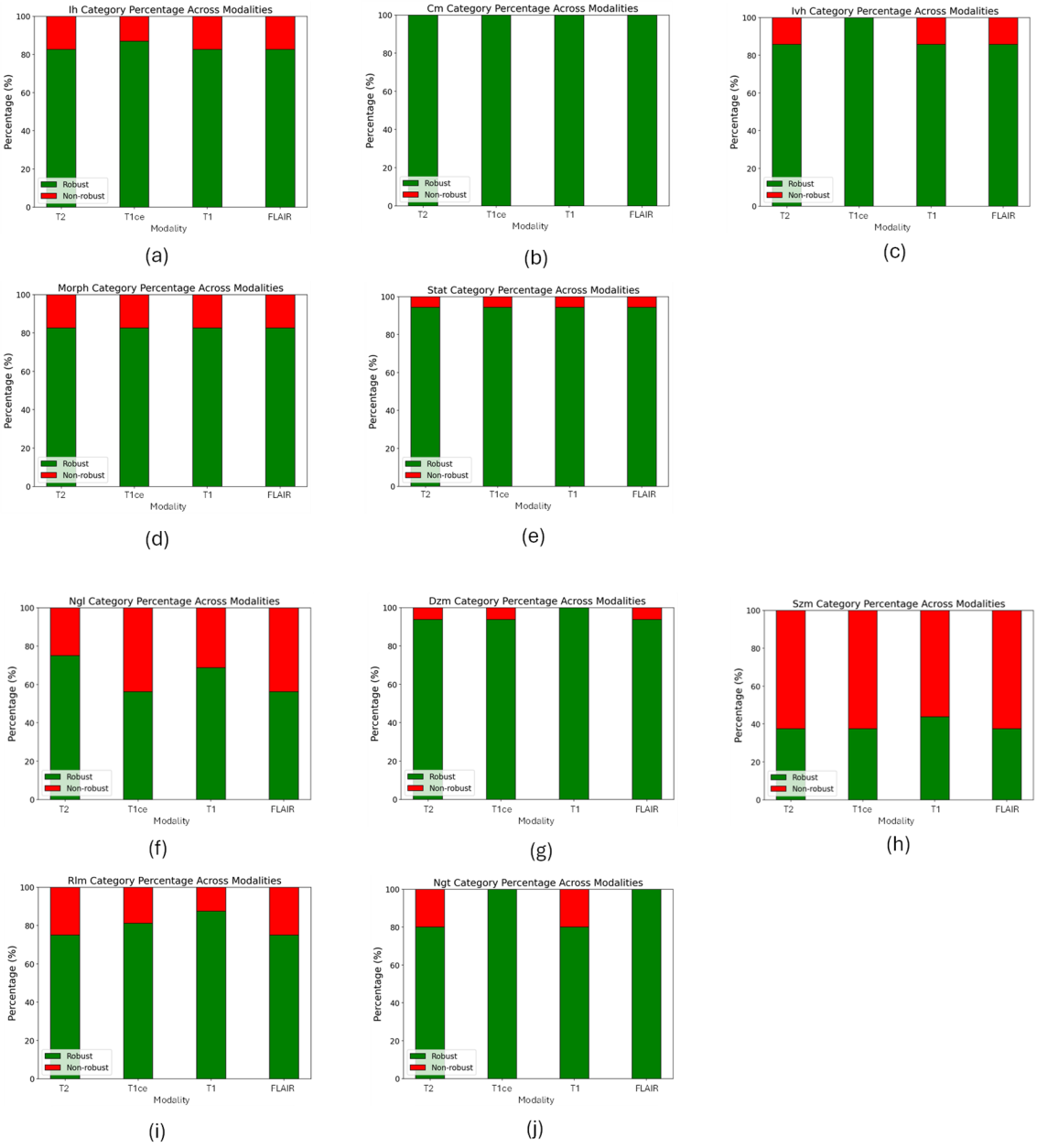


Figure C- 4 Analysis of Feature Robustness for Each Feature Family across MRI sequences (Percentage): (a) IH (Intensity Histograms, IBSI reference: ZVCW), (b)CM (Grey Level Co-occurrence Matrix, IBSI reference: LFYI), (c)IVH (Intensity-Volume Histogram, IBSI reference: P88C), (d) Morphological (MORPH, IBSI reference: HCUG), (e)STAT (Intensity-Based Statistics, IBSI reference: UHIW), (f) NGL (Neighbourhood Grey Level Dependence Matrix, IBSI reference: REK0), (g) DZM (Grey Level Distance Zone Matrix, IBSI reference: VMDZ), (h) SZM (Grey Level Size Zone Matrix, IBSI reference: 9SAK), (i) RLM (Grey Level Run Length Matrix, IBSI reference: TP0I), (j) NGT (Neighbourhood Grey Tone Difference Matrix, IBSI reference: IPET)

Feature robustness analysis:

Radiomic models can be severely affected by differences in positioning, image acquisition, and ROI segmentation, which introduce feature variability and limit the model generalisability [286]. Based on this, rigorous assessment of feature robustness is important. To distinguish robust from non-robust radiomic features in single-image analyses, perturbation-based augmentation techniques were performed [246]. Perturbed images (shown in Figure C- 5) covered rotation and volumetric shrinkage or enlargement (volume adaptation) [215], [286].

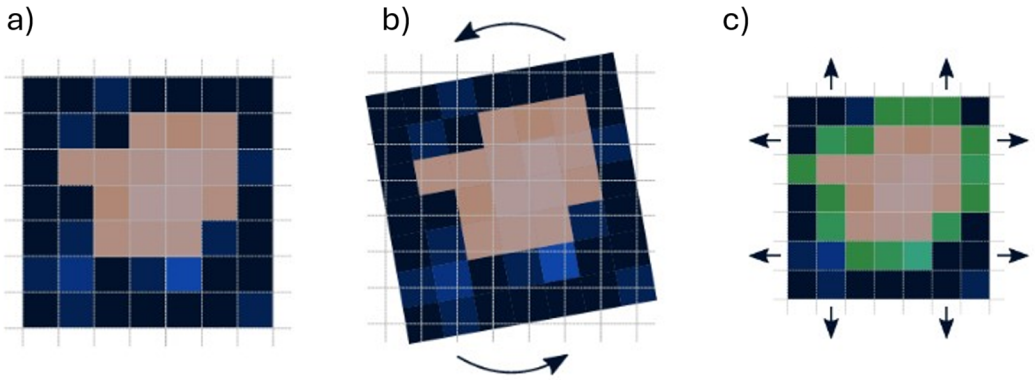


Figure C- 5 The perturbation methods : a) original b) rotation c) volume adaptation [246].

D

A) Normalisation Techniques

For the BraTS dataset, this study implements several normalisation techniques, which can be broadly categorised into two primary types: individual time-point normalisation methods and sample-based normalisation methods. The individual time-point normalisation methods include Z-score normalisation (referred to as zscore-normalise), Fuzzy C-means (FCM)-based tissue mean normalisation (fcm-normalise), Kernel Density Estimate (KDE)-based white matter mode normalisation (kde-normalise), and WhiteStripe normalisation (ws-normalise). On the other hand, the sample-based normalisation methods consist of Least Squares (LSQ)-based tissue mean normalisation (lsq-normalise) and Piecewise Linear Histogram Matching (nyul-normalise), with the exception of RAVEL normalisation (ravel-normalise), which was deemed inapplicable in the current context.

The results of RFS+ on ET, TC, and WT for each segmentation approach, using the various normalisation techniques, are summarised In Table D- 1.

Table D- 1 Results of RFS+ for ET, TC and WT.

Intensity Normalisation Technique	**Segmentation Approach**	**ET**	**TC**	**WT**
Nyul	Multiclass	79.44	79.53	88.98
	Multi-label	83.52	88.78	92.05
	Binary class	84.21	89.42	90.30
Z-score	Multiclass	84.99	89.71	91.65
	Multi-label	82.29	87.27	92.24
	Binary class	85.19	89.48	92.18
Whitestripe	Multiclass	83.61	87.99	90.47
	Multi-label	83.05	88.17	91.77
	Binary class	84.12	88.24	91.83
FCM	Multiclass	78.65	78.23	88.67
	Multi-label	77.56	79.42	87.65
	Binary class	83.65	84.21	88.53
LSQ	Multiclass	78.59	78.04	87.32
	Multi-label	79.34	80.11	86.59
	Binary class	82.34	84.87	83.98
KDE	Multiclass	79.22	77.45	88.67
	Multi-label	81.03	78.66	87.45
	Binary class	84.17	88.22	88.34

B) RFS+ Workflows for each region.

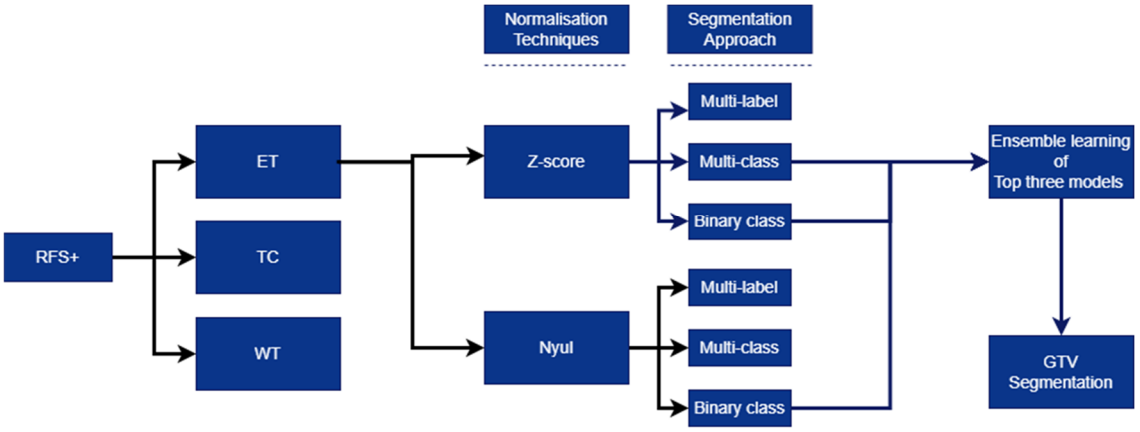


Figure D- 1 RFS+ ET based on Table D- 1.

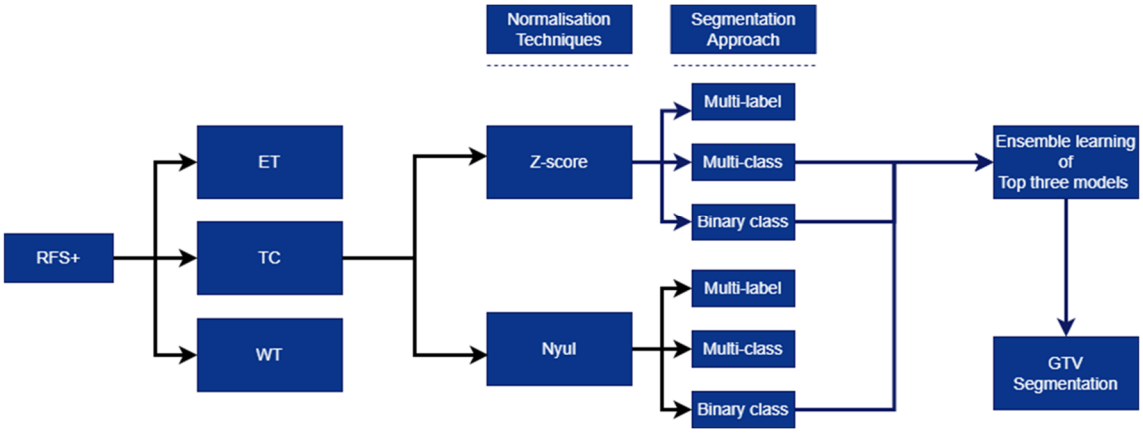


Figure D- 2 RFS+ for TC based on Table D- 1.

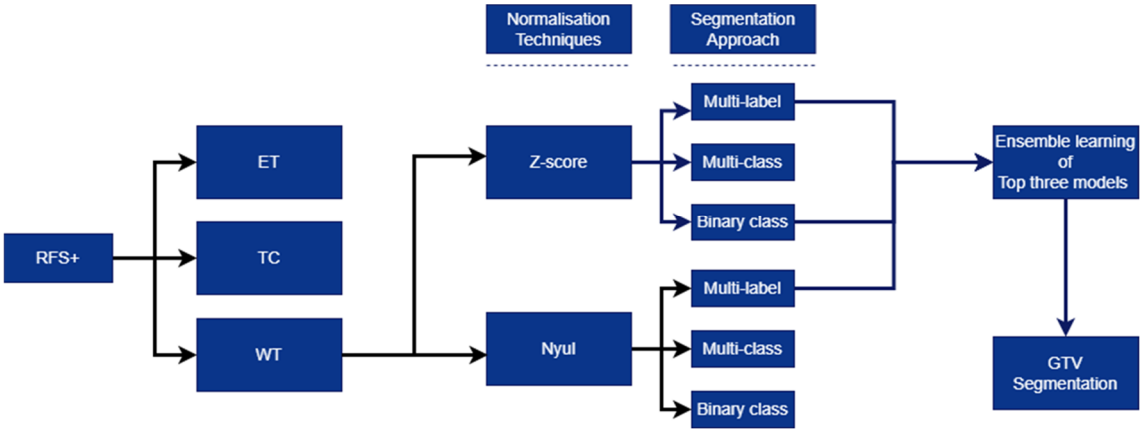


Figure D- 3 RFS+ for WT based on Table D- 1.

C) RFS+ with each segmentation approach for each region

Figure D- 4 shows the segmentation approaches along with their respective inputs and the RFS+ outputs for the ET region.

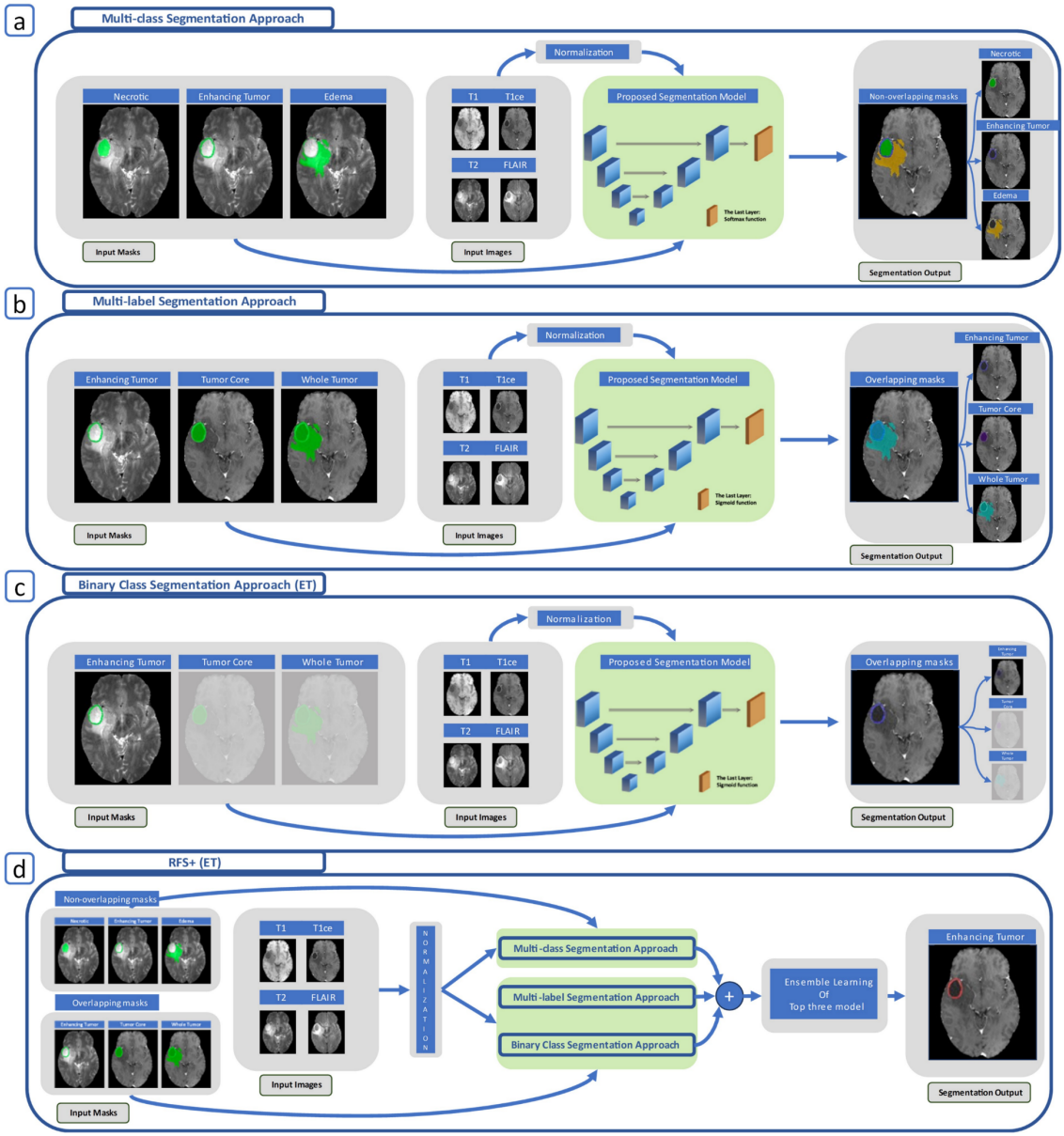


Figure D- 4 a) Multiclass segmentation b) Multi-label segmentation c) Binary class segmentation d) RFS+ for ET.

Figure D- 5 demonstrates the segmentation approaches along with their respective inputs and the RFS+ outputs for the TC region.

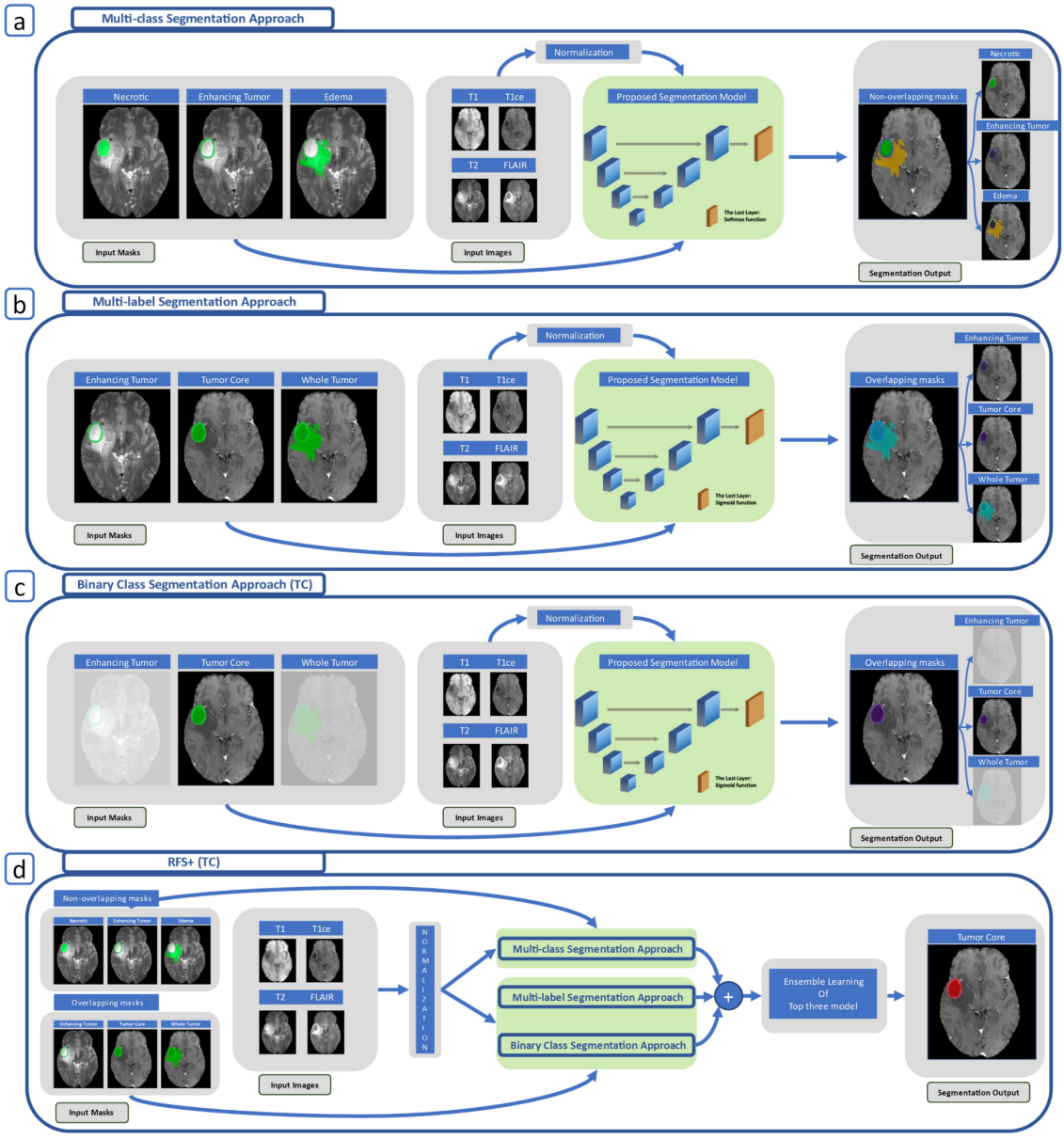


Figure D- 5 a) Multiclass segmentation b) Multi-label segmentation c) Binary class segmentation d) RFS+ for TC.

Figure D- 6 illustrates the segmentation approaches along with their respective inputs and the RFS+ outputs for WT region.

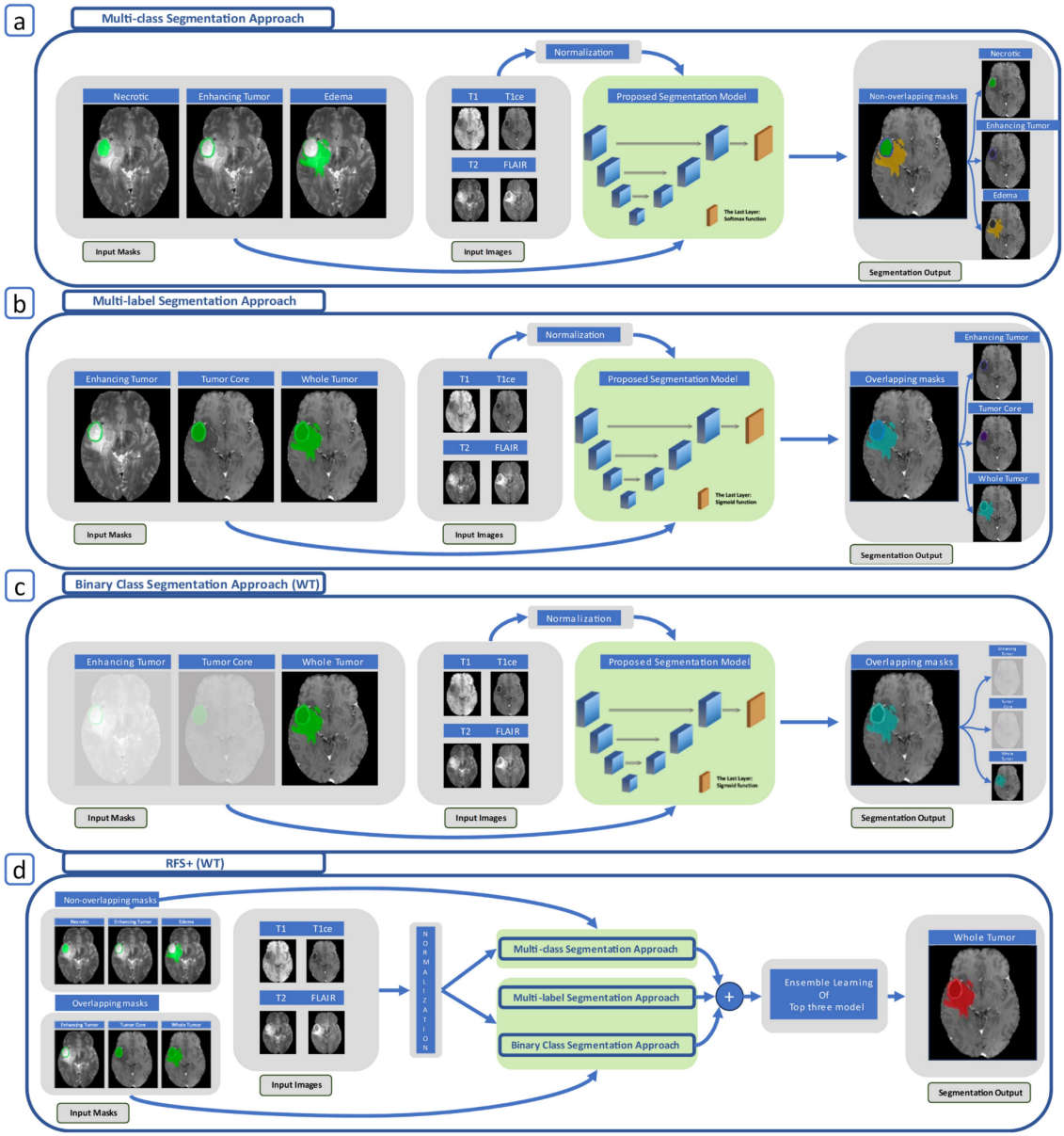


Figure D- 6 a) Multiclass segmentation b) Multi-label segmentation c) Binary class segmentation d) RFS+ for WT.

D) Analysing Training Requirements and Time Efficiency

The comparative analysis highlights RFS+'s exceptional resource efficiency, as it achieves outstanding performance while minimising computational demands, and when combined with 2D U-net, it yields a significant 67% reduction in memory usage and a substantial 92% decrease in training time compared to the extended nnU-net, all while maintaining high-quality segmentation results, making it an attractive solution for resource-constrained environments.

Table D- 2 The extended nnU-net requirements

	Models	RTX 3070 8GB		RTX 3090 24 GB		The model number	Total Time (Days)
		Trainable	Time in Days	Trainable	Time in Days		
Ensemble	BL baseline nnUNet	-	-	X	5	5	25
	BL+L+GN nnUNet with larger Unet	-	-	X	2	5	10
The extended nnU-net			-				35

Table D- 3 The 2D U-Net with RFS+ requirements (Any region).

	Models	RTX 3070 8GB		The model number	Total Time (Days)	RTX 3090 24 GB		The model number	Total Time (Days)
		Trainable	Time in Days			Trainable	Time in Days		
Ensemble	2D U-Net multiclass (Z-score normalisation)	X	3	1	3	X	1	1	1
	2D U-Net binary class (Z-score normalisation)	X	3	1	3	X	1	1	1
	2D U-Net binary class (Nyul normalisation)	X	3	1	3	X	1	1	1
RFS+					9				3

Table D- 4 The comparison of the ensemble methods.

Ensemble	RTX 3070 8GB		RTX 3090 24 GB	
	Trainable	Time in Days	Trainable	Time in Days
The extended nnU-net	-	-	X	35
RFS+	x	9	X	3

E) Acquisition Parameters retrieved from DICOM for STORM_GLIO

Table D- 5 Acquisition Parameters of STORM_GLIO (Average, Standard deviation)

	T1	T1ce	T2	FLAIR
Thickness/mm	4.77 +/-0.47	4.76 +/-0.47	4.74 +/-0.56	4.81 +/-0.39
TR/ms	489 +/-96	494 +/-98	5627 +/-1856	8084 +/-1832
Echo Time/ms	11 +/-2	11 +/-2	97 +/-8	112 +/-27
Inversion Time/ms	0 +/-0	0 +/-0	0 +/-0	2217 +/-259
Field Strength/T	1.54 +/-0.24	1.5 +/-0	1.54 +/-0.24	1.54 +/-0.24
Rows	426 +/-145	424 +/-146	546 +/-185	475 +/-219
Columns	417 +/-147	415 +/-148	527 +/-198	458 +/-232
Pixel spacing/mm	0.62 +/-0.19	0.62 +/-0.19	0.48 +/-0.14	0.59 +/-0.21
Slice Spacing/mm	5.99 +/-0.73	5.98 +/-0.74	6.27 +/-0.96	6.34 +/-0.72
SAR	1.09 +/-0.77	1.07+/-0.76	0.91 +/-0.53	0.69 +/-0.67