FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Machine Learning to Predict IDH Mutation Status through MRI Analysis

Ricardo Jorge Teixeira Tripeça



Master's Degree in Electrical and Computer Engineering

Supervisors: Tânia Pereira and Hélder Oliveira

October 22, 2024

Resumo

Glioblastoma multiforme (GBM) é o tumor cerebral primário mais agressivo e prevalente em adultos, conhecido pela sua rápida progressão e prognóstico desfavorável. Pesquisas recentes destacam o papel crucial da mutação da isocitrato desidrogenase (IDH) no GBM, que se correlaciona com taxas de sobrevivência prolongadas e respostas de tratamento distintas. Esta pesquisa utiliza técnicas de *Machine Learning* para prever o status da mutação IDH através da análise de imagens por ressonância magnética (RM), com o objetivo de avançar métodos de diagnóstico não invasivos.

A pesquisa proposta empregou uma variedade de modelos de aprendizagem supervisionada, incluindo *Random Forest*, *Support Vector Machine*, *eXtreme Gradient Boosting*, *Logistic Regression*, *Naïve Bayes* e *K-Nearest Neighbors* para avaliar características radiomicas extraídas de múltiplas sequências de RM. O estudo utilizou os *datasets* públicos UCSF-PDGM e UPENN-GBM para treinar e testar os modelos.

Seguindo o *pipeline* de *Radiogenomics*, a metodologia englobou o pré-processamento de images de RM, a extração de características radiomicas e a aplicação de técnicas de seleção de características e *data augmentation* para melhorar o desempenho do modelo. Os modelos treinados demonstraram desempenho satisfatório, alcançando um melhor valor de AUC de 0.946 ± 0.048.

Esta dissertação contribui para o desenvolvimento de sistemas de IA para a previsão de mutações IDH, integrando *radiogenomics* e *machine learning*. Também aborda o desequilíbrio de classes através de técnicas de *data augmentation*. Os resultados sugerem que a combinação de RM e *machine learning* pode melhorar os diagnósticos não invasivos para GBM, oferecendo uma abordagem fiável para entender o panorama genético do glioblastoma e auxiliando no planeamento de tratamentos personalizados.

Palavras-chave: Machine Learning, Imagens de Ressonância Magnética, Radiogenomics, Isocitrato Desidrogenase, Glioblastoma multiforme.

Abstract

Glioblastoma multiforme (GBM) stands as the most aggressive and prevalent primary brain tumor in adults, notorious for its rapid progression and poor prognosis. Recent research highlights the crucial role of the isocitrate dehydrogenase (IDH) mutation in GBM, which correlates with extended survival rates and distinct treatment responses. This research leverages machine learning techniques to predict IDH mutation status through magnetic resonance imaging (MRI) analysis, aiming to advance non-invasive diagnostic methods.

The proposed research employed a range of supervised machine learning models, including Random Forest, Support Vector Machine, eXtreme Gradient Boosting, Logistic Regression, Naïve Bayes, and K-Nearest Neighbors to evaluate radiomic features extracted from multiple MRI sequences. The study utilized the UCSF-PDGM and UPENN-GBM public datasets for training and testing the models.

Following the Radiogenomics pipeline, the methodology encompassed MRI image preprocessing, radiomic feature extraction, and the application of feature selection and data augmentation techniques to boost model performance. The trained models demonstrated satisfactory performance, achieving a best AUC value of 0.946 ± 0.048.

This dissertation contributes to the development of AI systems for IDH mutation prediction, integrating radiogenomics and machine learning. It also addresses class imbalance through data augmentation techniques. The findings suggest that combining MRI and machine learning can improve non-invasive diagnostics for GBM, offering a reliable approach for understanding glioblastoma's genetic landscape and aiding personalized treatment planning.

Keywords: Machine Learning, Magnetic Resonance Imaging, Radiogenomics, Isocitrate Dehydrogenase, Glioblastoma multiforme.

United Nations Sustainable Development Goals

The master's dissertation titled "Machine Learning to Predict IDH Mutation Status through MRI Analysis" addresses a critical issue in the treatment of glioblastoma multiforme by utilizing machine learning techniques to predict IDH mutation status based on magnetic resonance imaging analysis. This chapter examines how the project aligns with the United Nations Sustainable Development Goals (SDGs), specifically those aimed at improving health and well-being, and promoting technological innovation.

Identification of SDGs

The dissertation project directly addresses the following SDGs:

- **SDG 3:** Good Health and Well-being
- **SDG 9:** Industry, Innovation, and Infrastructure

Specific Targets

Within the identified SDGs, the specific targets addressed by the project are:

- **SDG 3.4:** *"By 2030, reduce by one third premature mortality from non-communicable diseases through prevention and treatment, and promote mental health and well-being."*
- **SDG 3.8:** *"Achieve universal health coverage, including financial risk protection, access to quality essential health-care services and access to safe, effective, quality and affordable essential medicines and vaccines for all."*
- **SDG 9.5:** *"Enhance scientific research, upgrade the technological capabilities of industrial sectors in all countries, in particular developing countries, including, by 2030, encouraging innovation and substantially increasing the number of research and development workers per 1 million people and public and private research and development spending"*

Contribution

SDG 3: Good Health and Well-being

The project contributes to SDG 3 by improving the diagnosis and treatment of GBM. The use of machine learning to predict IDH mutation status can lead to earlier and more accurate diagnoses, which is critical for effective treatment planning and improving patient outcomes. By facilitating non-invasive diagnostic methods, the project reduces the need for biopsies, minimizing risks and improving the quality of life for patients.

SDG 9: Industry, Innovation, and Infrastructure

The project advances technological innovation in the medical field by integrating advanced machine learning techniques with radiogenomics. This enhances the technological capabilities of healthcare infrastructure and promotes the development of innovative diagnostic tools. The research contributes to scientific knowledge and can stimulate further advancements in medical imaging and cancer treatment technologies.

Performance Indicators

To evaluate the impact of the dissertation on the SDGs, specific performance indicators are established:

1. **For SDG 3:**
   (a) Accuracy of the predictive model in diagnosing IDH mutations.
   (b) Number of early detections of IDH mutations in GBM patients.
   (c) Improvement in patient survival rates and quality of life metrics.
2. **For SDG 9:**
   (a) Number of technological innovations or tools developed as a result of the project.
   (b) Increase in research publications and citations related to the dissertation topic.
   (c) Amount of funding and resources allocated to follow-up research and development.

Acknowledgements

Completing a Master's thesis is a journey that combines both academic rigor and personal growth, made possible through the support and contributions of many individuals and entities. Therefore, I would like to express my deepest gratitude to everyone who played a part in this achievement.

First and foremost, I extend my heartfelt thanks to my advisor from INESC TEC, Tânia Pereira, for her guidance, patience, dedication, and kindness. Her constant availability and prompt responses provided the confidence and efficiency I needed to navigate each stage of my work.

I am also grateful to my co-advisor, Hélder Oliveira, for his insights and practical solutions during our group meetings.

My sincere thanks go to the INESC TEC institute for fostering an environment of innovation and rigorous research. I am honored to have contributed, even in a modest way, to its mission. I hope my work will, in some small measure, benefit the broader scientific community.

I extend my appreciation to FEUP and all its faculty members, who have provided me with five years of rich knowledge, personal development, and memorable moments.

To my parents, Judite and Jorge, words cannot express my gratitude for your 22 years of exceptional education, patience, and unconditional support. Your commitment to providing me with the best educational opportunities has been the foundation of my personal and professional growth.

To my close friends and my girlfriend, Raquel, thank you for the joyful moments and unwavering support that have rejuvenated my spirit and sustained my productivity. Without your companionship, enduring the intense months of hard work would have been immeasurably more challenging.

Once again, my sincere thanks to everyone who has supported me throughout this journey. Your contributions have been invaluable, and I am profoundly grateful for each and every one of you.

Ricardo Jorge Teixeira Tripeça

This research received support from the European Union's Horizon 2020 research and innovation program, within the OpenInnoTrain project under the Marie Skłowdowska-Curie grant agreement no. 823971. The content of this publication does not reflect the official opinion of the European Union. Responsibility for the information and views expressed in the publication lies entirely with the authors.

"You should be glad that bridge fell down.
I was planning to build thirteen more to that same design"

Isambard Kingdom Brunel

Contents

1 Introduction **1**
- 1.1 Motivation and Goals . . . 2
- 1.2 Contributions . . . 3
- 1.3 Document Structure . . . 3

2 Background **5**
- 2.1 Brain Anatomy . . . 5
- 2.2 Glioma . . . 6
  - 2.2.1 Classification of Adult-Type Diffuse Gliomas . . . 7
- 2.3 Isocitrate dehydrogenase . . . 8
  - 2.3.1 Normal Function of IDH . . . 8
  - 2.3.2 Abnormal Function of IDH . . . 8
- 2.4 Techniques for Brain Tumor Diagnosis . . . 9
  - 2.4.1 Computed Tomography . . . 9
  - 2.4.2 Magnetic Resonance Imaging . . . 10
  - 2.4.3 Positron Emission Tomography . . . 11
  - 2.4.4 Electroencephalography . . . 12
  - 2.4.5 Summary . . . 12
- 2.5 Radiogenomics . . . 13
  - 2.5.1 Conventional and Deep Radiomics . . . 13
  - 2.5.2 Genomics . . . 13
  - 2.5.3 Integration into Radiogenomics . . . 14
  - 2.5.4 Data acquisition . . . 15
  - 2.5.5 Preprocessing . . . 16
  - 2.5.6 Feature Extraction . . . 17
  - 2.5.7 Feature Selection / Dimensionality Reduction . . . 17
  - 2.5.8 Association of Radiomics and Genomics Features . . . 22
  - 2.5.9 Performance Assessment: . . . 22
- 2.6 Radiomic Features . . . 24
  - 2.6.1 Statistical / Histogram Features . . . 24
  - 2.6.2 Texture Features . . . 25
  - 2.6.3 Morphological features . . . 26
  - 2.6.4 Software Tools for Radiomic Features Extraction . . . 26
- 2.7 Machine Learning Algorithms . . . 28
  - 2.7.1 Support Vector Machine . . . 28
  - 2.7.2 Random Forest Classifier . . . 30
  - 2.7.3 XGBoost (eXtreme Gradient Boosting) . . . 33
  - 2.7.4 Logistic Regression . . . 33

2.7.5 K-Nearest Neighbors . . . 35
2.7.6 Naïve Bayes . . . 36
2.8 Interactive Tools . . . 37
2.8.1 CaPTk (Cancer Imaging Phenomics Toolkit) . . . 38
2.8.2 Qt Designer . . . 38
2.8.3 Importance of Interactive Tools in Radiogenomics . . . 39
2.9 Summary . . . 39

3 Literature Review **41**
3.1 State of the Art in GBM Radiogenomics . . . 41
3.1.1 Deep Learning vs Machine Learning . . . 42
3.1.2 Predicting IDH Mutation Status Using Deep Learning Approaches . . . 43
3.1.3 Predicting IDH Mutation Status Using Machine Learning Techniques . . . 45
3.2 Summary . . . 46

4 Data Description **48**
4.1 UCSF-PDGM Dataset . . . 48
4.1.1 Patient Population . . . 48
4.1.2 Clinical Data . . . 50
4.1.3 Image Acquisition . . . 51
4.1.4 Tumor Segmentation . . . 52
4.2 UPENN-GBM Dataset . . . 53
4.2.1 Patient Population . . . 53
4.2.2 Clinical Data . . . 53
4.2.3 Image Acquisition . . . 54
4.2.4 Tumor Segmentation . . . 55
4.3 Summary . . . 55

5 Methodology **56**
5.1 Dataset Collection . . . 57
5.2 Pre-processing . . . 60
5.3 Feature Extraction . . . 60
5.4 Feature Selection / Dimensionality Reduction . . . 61
5.5 Classification Models and Hyperparameter Tuning . . . 62
5.5.1 Hyperparameters Search . . . 62
5.5.2 Support Vector Machine . . . 63
5.5.3 Random Forest . . . 63
5.5.4 XGBoost . . . 64
5.5.5 Logistic Regression . . . 65
5.5.6 k-Nearest Neighbours . . . 66
5.5.7 Näive Bayes . . . 66
5.6 Training and Evaluation . . . 67
5.7 Interactive Tool . . . 68
5.7.1 MRI Image and ROI Visualization . . . 68
5.7.2 Clinical Data Visualization . . . 69
5.7.3 Feature Extraction Command Page . . . 69
5.7.4 Training Station . . . 71
5.7.5 Importance and Impact . . . 71
5.8 Summary . . . 72

6 Results and Discussions **74**
6.1 Results for UCSF-PDGM dataset . . . 74
6.1.1 Best Hyperparameters . . . 79
6.1.2 Limitations . . . 80
6.2 Results for UPENN-GBM dataset . . . 80
6.2.1 Best Hyperparameters . . . 85
6.2.2 Limitations . . . 85
6.3 Results for Combined Dataset . . . 86
6.3.1 Best Hyperparameters . . . 90
6.3.2 Limitations . . . 91
6.4 Summary . . . 91

7 Conclusion **93**

References **95**

A Detailed Experimental Results **102**
A.1 Dataset 1: UCSF-PDGM . . . 102
A.2 Dataset 2: UPENN-GBM . . . 109
A.3 Dataset 3: Combined . . . 116

List of Figures

2.1 Illustration highlighting the primary regions of the brain. 6
2.2 The 2021 WHO classification of adult-type diffuse gliomas, highlighting the importance of IDH mutation status. 7
2.3 Normal CT scan of the brain, illustrating standard anatomical structures without abnormalities. 9
2.4 The three prominent views of MRI scanning: axial, sagittal, and coronal (top). The bottom section compares brain MRI slices of T1-weighted, T2-weighted, and FLAIR sequences, highlighting the differences in tissue contrast. 11
2.5 MRI brain scan (left), PET brain scan from a "healthy" patient (middle). 12
2.6 Comparison between traditional and deep radiomics 14
2.7 Radiogenomics pipeline 15
2.8 Comparison of PCA and t-SNE. 20
2.9 Multiple hyperplanes separate the data from two classes. 28
2.10 Nonlinear SVM. Left: Original 1D dataset for classification. Right: Mapping 1D data to 2D to become able to separate the two classes. 29
2.11 Ensemble types of methods. 30
2.12 Random Forest Algorithm in Machine Learning. 31
2.13 Logistic Regression. 34
2.14 KNN simple example. 35
2.15 CaPTk Screenshot showing example annotated ROIs and initialized tumor. 38

3.1 AI and its models used in radiogenomics studies (a) AI; (b) AI models. 42
3.2 Fully automated hybrid model for IDH status prediction. 44
3.3 Architecture of the deep learning model consisting of two selection steps. 45

4.1 Gender and Age distributions in UCSF-PDGM. Left: Gender. Right: Age. 49
4.2 WHO Grade distribution in UCSF-PDGM 49
4.3 Final Pathologic Diagnosis (WHO 2021) distribution in UCSF-PDGM 50
4.4 Distribution of IDH Mutations in the UCSF-PDGM Dataset. 51
4.5 MRI Sequences included in the UCSF-PDGM dataset. 52
4.6 Gender and age distributions in UPENN-GBM. Left: Gender. Right: Age. 53
4.7 Distribution of IDH1 Mutations in the UCSF-PDGM Dataset. 54
4.8 Examples of the different sequences available in UPENN-GBM. 55

5.1 Schematic representation of the classification methodology for predicting IDH mutation status. 57
5.2 T1, T1GD, T2, and FLAIR sequences, respectively, Slice 84 of Patient 0025's Brain, from UCSF-PDGM Dataset. 58

5.3 Flowchart of patient inclusion and exclusion of UCSF-PDGM and UPENN-GBM dataset. 59
5.4 Screenshot of the interactive tool's image visualization feature. 69
5.5 Screenshot of the interactive tool's clinical data visualization feature. 70
5.6 Screenshot of the interactive tool's feature extraction command page. 70
5.7 Screenshot of the interactive tool's Training Station. 71

List of Tables

2.1 Comparison of PCA and t-SNE - techniques for dimensionality reduction. 21
2.2 Difference between Feature Selection and Dimensionality Reduction. 22
2.3 Most widely used and publicly available software tools for radiomic features extraction. 27
2.4 Distribution of IBSI-standardized features across various radiomic tools, including Pyradiomics, MITK, LIFEx, SERA and CaPTk. The table details the number of features in each category, illustrating the coverage and focus areas of different tools. 27
2.5 Advantages and Disadvantages of Support Vector Machines 30
2.6 Difference between Decision Tree and Random Forest 32
2.7 Advantages and Disadvantages of Random Forest Classifier 32
2.8 Differences between XGBoost and Random Forest. 33
2.9 Advantages and Disadvantages of Logistic Regression 35
2.10 Advantages and Disadvantages of K-Nearest Neighbors 36
2.11 Advantages and Disadvantages of Näive Bayes 37

3.1 Summary of studies on predicting IDH1 mutation status in gliomas using MRI radiomics and machine learning approaches. 46

5.1 Support Vector Machine Hyperparameter Values 63
5.2 Random Forest Hyperparameter Values 64
5.3 XGBoost Hyperparameter Values 64
5.4 Logistic Regression Hyperparameter Values 65
5.5 k-Nearest Neighbours Hyperparameter Values 66
5.6 Näive Bayes Hyperparameter Values 66
5.7 IDH Mutation Status of Patients in Training and Test Sets Across the Three Datasets. 68

6.1 Best Performances of SVM model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. 75
6.2 Best Performances of KNN model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. 75
6.3 Best Performances of XGBoost model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. 76
6.4 Best Performances of RF model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. 76
6.5 Best Performances of LogR model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. 77
6.6 Best Performances of NB model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. 77
6.7 Comparison of best-performing classifiers for UCSF-PDGM. 79

6.8 Best Hyperparameters found for each of the Best-Performing Classifiers for UCSF-PDGM . . . 79
6.9 Best Performances of SVM model Across different MRI Sequences and Feature Selection Methods for UPENN-GBM. . . . 81
6.10 Best Performances of KNN model Across different MRI Sequences and Feature Selection Methods for UPENN-GBM. . . . 81
6.11 Best Performances of XGBoost model Across different MRI Sequences and Feature Selection Methods for UPENN-GBM. . . . 82
6.12 Best Performances of RF model Across different MRI Sequences and Feature Selection Methods for UPENN-GBM. . . . 82
6.13 Best Performances of LogR model Across different MRI Sequences and Feature Selection Methods for UPENN-GBM. . . . 83
6.14 Best Performances of NB model Across different MRI Sequences and Feature Selection Methods for UPENN-GBM. . . . 83
6.15 Comparison of best-performing classifiers for UPENN-GBM. . . . 84
6.16 Best Hyperparameters found for each of the Best-Performing Classifiers for UPENN-GBM . . . 85
6.17 Best Performances of SVM model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. . . . 86
6.18 Best Performances of KNN model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. . . . 87
6.19 Best Performances of XGBoost model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. . . . 87
6.20 Best Performances of RF model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. . . . 88
6.21 Best Performances of LogR model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. . . . 88
6.22 Best Performances of NB model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. . . . 89
6.23 Comparison of best-performing classifiers for Combined Dataset. . . . 90
6.24 Best Hyperparameters found for each of the Best-Performing Classifiers for Combined Dataset . . . 90

A.1 Performance of SVM Model for UCSD-PDGM. . . . 102
A.2 SVM SMOTE ratio and K features for UCSD-PDGM . . . 103
A.3 Performance of XGBoost Model for UCSD-PDGM. . . . 103
A.4 XGBoost SMOTE ratio and K features for UCSD-PDGM . . . 104
A.5 Performance of RF Model for UCSD-PDGM. . . . 104
A.6 RF SMOTE ratio and K features for UCSD-PDGM . . . 105
A.7 Performance of LogR Model for UCSD-PDGM. . . . 105
A.8 LogR SMOTE ratio and K features for UCSD-PDGM . . . 106
A.9 Performance of NB Model for UCSD-PDGM. . . . 106
A.10 NB SMOTE ratio and K features for UCSD-PDGM . . . 107
A.11 Performance of KNN Model for UCSD-PDGM. . . . 107
A.12 KNN SMOTE ratio and K features for UCSD-PDGM . . . 108
A.13 Performance of SVM Model for UPENN-GBM. . . . 109
A.14 SVM SMOTE ratio and K features for UPENN-PDGM . . . 110
A.15 Performance of XGBoost Model for UPENN-GBM. . . . 110
A.16 XGBoost SMOTE ratio and K features for UPENN-PDGM . . . 111

A.17 Performance of RF Model for UPENN-GBM. 111
A.18 RF SMOTE ratio and K features for UPENN-GBM 112
A.19 Performance of LogR Model for UPENN-GBM. 112
A.20 LogR SMOTE ratio and K features for UPENN-PDGM 113
A.21 Performance of NB Model for UPENN-GBM. 113
A.22 NB SMOTE ratio and K features for UPENN-PDGM 114
A.23 Performance of KNN Model for UPENN-GBM. 114
A.24 KNN SMOTE ratio and K features for UPENN-PDGM 115
A.25 Performance of SVM Model for Combined Dataset. 116
A.26 SVM SMOTE ratio and K features for Combined Dataset 117
A.27 Performance of XGBoost Model for Combined Dataset. 117
A.28 XGBoost SMOTE ratio and K features for Combined Dataset 118
A.29 Performance of RF Model for Combined Dataset. 118
A.30 RF SMOTE ratio and K features for Combined Dataset. 119
A.31 Performance of LogR Model for Combined Dataset. 119
A.32 LogR SMOTE ratio and K features for Combined Dataset 120
A.33 Performance of NB Model for Combined Dataset. 120
A.34 NB SMOTE ratio and K features for Combined Dataset. 121
A.35 Performance of KNN Model for Combined Dataset. 121
A.36 KNN SMOTE ratio and K features for Combined Dataset. 122

Abbreviations

2D	Two Dimensional
3D	Three Dimensional
AI	Artificial Intelligence
ASL	3D Arterial Spin Labeling
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BraTS	Brain Tumor Segmentation
CaPTk	Cancer Imaging Phenomics Toolkit
CNN	Convolutional Neural Network
CNS	Central Nervous System
CT	Computed Tomography
CSV	Comma-Separated Values
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DLR	Learning based Radiomics
DNA	Deoxyribonucleic acid
DSC	Dynamic Susceptibility Contrast
DTI	Diffusion-Tensor Imaging
DWI	Diffusion-weighted imaging
EEG	Electroencephalography
EGFR	Epidermal Growth Factor Receptor
FDG	fluorodeoxyglucose
FLAIR	Fluid-Attenuated Inversion Recovery
FN	False Negative
FP	False Positive
GBM	Glioblastoma Multiforme
Gd	Gadolinium
GLCM	Gray-Level Co-occurrence Matrix
GLDZM	Gray-Level Distance Zone Matrix
GLRLM	Gray-Level Run Length Matrix
GLSZM	Gray-Level Size Zone Matrix
GUI	Graphical User Interface
HARDI	2D 55-direction High Angular Resolution Diffusion Imaging
IBSI	Image Biomarker Standardization Initiative
IDH	Isocitrate Dehydrogenase
kNN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis
LoG	Laplacian of Gaussian
LogR	Logistic Regression

MAD	Mean Absolute Difference
MGMT	O^6-Methylguanine-DNA methyltransferase
MITK	Medical Imaging Interaction Toolkit
ML	Machine Learning
mpMRI	Multi-parametric Magnetic Resonance Imaging
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
mRNA	messenger ribonucleic acid
NADH	nicotinamide adenine dinucleotide (NAD) + hydrogen (H)
NADPH	nicotinamide adenine dinucleotide phosphate
NB	Naïve Bayes
NEC	Not Elsewhere Classified
NGLDM	Neighboring Gray-Level Dependence Matrix
NGTDM	Neighborhood Gray Tone Difference Matrix
NIfTI	Neuroimaging Informatics Technology Initiative
NMDA	N-methyl-D-aspartate
NOS	Not Otherwise Specified
NRRD	Nearly Raw Raster Data
OOB	Out-of-bag
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PWI	Perfusion-weighted imaging
RBF	Radial Basis Function
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
SDG	Sustainable Development Goals
SERA	Standardized Environment for Radiomics Analysis
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
T1GD	T1 with gadolinium contrast enhancement
TCIA	The Cancer Imaging Archive
TN	True Negative
TP	True Positive
t-SNE	t-Distributed Stochastic Neighbor Embedding
UCSF-PDGM	University of California San Francisco Preoperative Diffuse Glioma MRI
UPENN-GBM	University of Pennsylvania Health System Glioblastoma
VOIs	Volumes Of Interest
WHO	World Health Organization
XGBoost	eXtreme Gradient Boosting

Chapter 1

Introduction

Glioblastoma multiforme (GBM) is the most aggressive and prevalent primary brain tumor in adults. Characterized by rapid progression and a poor prognosis, GBM presents a significant clinical challenge. According to Orphanet data, it affects approximately 3 per 100,000 adults annually in North America and Europe, with a higher incidence observed in males [31]. Globally, the incidence rises to 6 per 100,000, with significantly lower rates observed in non-Caucasian populations [79].

Without treatment, GBM can be fatal within six months. Patients often present with severe symptoms including headaches, seizures, and cognitive or motor impairments, underscoring the aggressive nature of this disease [43]. Despite advancements in medical interventions such as surgery, radiation, and chemotherapy, the average survival time for GBM patients remains around one year. This poor prognosis is largely attributed to the molecular and genomic heterogeneity of GBM, resulting in diverse responses to treatment.

Recent research has demonstrated that mutations in the isocitrate dehydrogenase (IDH) gene are associated with longer overall survival in GBM patients compared to those with wild-type IDH [33]. Therefore, understanding the IDH mutation status is crucial for predictive assessments and offers valuable insights into the disease's progression.

The IDH mutation status also has significant implications for treatment strategies. Patients with IDH-mutant glioblastomas often respond differently to therapies compared to those with wild-type IDH, influencing clinical decision-making and personalized treatment plans [33]. Consequently, exploring the molecular landscape influenced by IDH mutations is important for developing tailored treatment strategies and enhancing the management of glioblastoma patients.

This dissertation aims to deepen our understanding of GBM, and gliomas in general, by leveraging machine learning techniques to predict IDH mutation status based on magnetic resonance imaging (MRI) analysis. By integrating radiogenomic data, this study seeks to bridge the gap between radiological imaging and genomic information, ultimately contributing to more personalized and effective treatment approaches for GBM patients.

1.1 Motivation and Goals

Traditional methods for delineating glioblastoma characteristics, specifically brain biopsy and mRNA profiling, present significant challenges that require the exploration of alternative, non-invasive, approaches. Brain biopsy, a standard diagnostic procedure, is inherently invasive, involving the extraction of tissue samples directly from the tumor site [74]. This invasiveness not only poses risks to patients, including potential complications and side effects, but also limits the frequency with which samples can be obtained for monitoring disease progression. Additionally, the heterogeneity of glioblastomas raises concerns about the representativeness of a single biopsy, potentially providing an incomplete and biased view of the tumor's molecular landscape [74]. Furthermore, these methods are also resource-intensive and expensive. It requires sophisticated laboratory techniques and specialized equipment, contributing to the overall cost of diagnosis and impeding widespread accessibility.

The motivation for this research stems from the imperative to address these limitations by leveraging non-invasive methods. Magnetic resonance imaging has evolved as a promising alternative in neuro-oncology, enabling high-resolution visualization of brain structures and abnormalities, and allowing for the non-invasive assessment of tumor characteristics [5]. Its ability to capture dynamic changes over time makes it an ideal tool for monitoring disease progression. However, the full potential of MRI in glioblastoma diagnosis and prognosis can be further realized through the integration of machine learning.

Machine learning (ML) techniques offer a means to extract intricate patterns and relationships from the vast and complex datasets generated by MRI scans. By automating the analysis of radiological images, ML can provide quantitative insights into tumor characteristics, allowing for more accurate and reproducible assessments. This integration of MRI and ML not only overcomes the limitations of invasive procedures but also paves the way for a more comprehensive, timely, and cost-effective approach to understanding the molecular foundations of gliomas.

In recent years, the application of ML methods for characterizing gliomas from medical imaging has gained significant attention [55]. Leveraging radiogenomic features alongside ML methodologies has shown considerable success in predicting diagnosis, prognosis, and treatment responses for glioma patients [17]. Particularly, the correlation between radiomic features and IDH genotype has been investigated, revealing promising results in non-invasively identifying genetic mutations [53].

We hypothesized that supervised ML approaches employing various classifiers, including Random Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), Logistic Regression (LogR), Naïve Bayes (NB), and K-Nearest Neighbors (KNN), could effectively predict IDH mutation status based solely on MRI radiomic features. The primary objective of this study is to compare the performance of these classifiers using multiple MRI sequences — namely T1-weighted, T2-weighted, T1-weighted with gadolinium contrast enhancement (T1GD), and Fluid-Attenuated Inversion Recovery (FLAIR) — to accurately predict IDH mutation status.

By exploring these various classifiers and MRI sequences, we aim to identify the most effective

combination for predicting IDH mutations. This approach could enhance the non-invasive diagnostic process, offering a reliable and accessible method for understanding the genetic landscape of glioblastomas. The findings from this study are expected to contribute to the advancement of neuro-oncology, improving patient outcomes through better diagnostic accuracy and personalized treatment strategies.

1.2 Contributions

This dissertation introduces contributions aimed at advancing the understanding and predictive capabilities in the context of IDH mutations in gliomas. The study's contributions are as follows:

- An in-depth study on the use of Artificial Intelligence (AI) systems to predict IDH mutation status through MRI analysis.
- Application of Machine Learning methods to explore the correlation between radiomics features and IDH mutation status.
- Implementation of data augmentation techniques to address data imbalance issues and evaluation of their impact on predictive performance.
- Comprehensive assessment to determine the best combinations of features, MRI sequences, feature selection methods, and classifiers for predicting IDH mutation status.

1.3 Document Structure

This document is structured into seven chapters, each contributing to a comprehensive exploration of the research. The present chapter introduces the research topic, outlining the significance of GBM and the importance of predicting IDH mutation status. It also presents the motivation, goals, and contributions of the study.

Chapter 2 provides a detailed overview of the clinical and technical foundations relevant to the research. It covers brain anatomy, the characteristics and classification of gliomas, the role of IDH mutations, and various imaging techniques used in brain tumor diagnosis. Additionally, it introduces the field of radiogenomics and discusses methods for feature extraction and selection, as well as data analysis using machine learning. Finally, it discusses the importance of interactive tools.

Chapter 3 reviews existing literature on the application of machine learning and radiogenomics in the context of GBM. It compares state-of-the-art techniques for predicting IDH mutation status and evaluates the strengths and limitations of both traditional machine learning and deep learning approaches.

Chapter 4 describes the datasets used in this research, detailing their characteristics, the patient populations, clinical data, image acquisition procedures, and tumor segmentation methods.

Chapter 5 outlines the research design and procedures used to achieve the study's objectives. It details the processes of dataset collection, image preprocessing, feature extraction, and feature selection. It also describes the machine learning models implemented, the hyperparameter tuning strategies employed, and the evaluation metrics used to assess model performance. Finally, it discusses the development and functionality of the interactive tool created for this project.

Chapter 6 presents the results of the machine learning experiments conducted on the datasets. It includes a performance analysis of the different classifiers across various MRI sequences. The chapter also discusses the implications of the findings, the limitations of the study, and potential areas for future research.

Lastly, Chapter 7 summarizes the key findings of the research, highlights the contributions made to the field of neuro-oncology, and suggests directions for future work. It reflects on the significance of integrating machine learning with radiogenomic data for non-invasive diagnostics and personalized treatment planning in GBM.

Each chapter builds upon the previous one, providing a comprehensive and logical progression from the introduction of the problem to the presentation of the research findings and conclusions.

Chapter 2

Background

This chapter provides a detailed overview of the clinical and technical foundations relevant to understanding the research's scope. It begins with a succinct overview of brain anatomy (Section 2.1). Following this, it delves into the complexities and clinical significance of gliomas (Section 2.2), detailing their classification and characteristics. The chapter then focuses on isocitrate dehydrogenase mutations (Section 2.3), emphasizing their normal and abnormal functions and their impact on gliomas. Next, it provides an overview of some of the existing imaging modalities (Section 2.4) used in this context. Subsequently, the chapter explores the field of radiogenomics (Section 2.5), discussing conventional and deep radiomics, genomics, and their integration. It also covers background on data acquisition, preprocessing, feature extraction and selection, the association of radiomics and genomics features, data analysis using machine learning, and performance assessment. Section 2.6 elaborates on Radiomic Features. The chapter concludes with a discussion on the importance of interactive tools in radiogenomics (Section 2.8).

2.1 Brain Anatomy

The brain is a complex organ serving as the primary control center for a multitude of bodily functions and processes. It orchestrates the nervous system, processes information, coordinates actions, and maintains the body's internal balance, known as homeostasis.

Structurally, the brain is divided into three main parts: the cerebrum, cerebellum, and brainstem, as depicted in Figure 2.1. The cerebrum, the brain's most significant region, consists of the right and left hemispheres. It is crucial for complex activities, including sensory perception (touch, vision, and hearing), speech, reasoning, emotions, learning, and the fine control of movement. Below the cerebrum, the cerebellum plays a vital role in coordinating muscle movements, maintaining posture, and ensuring balance. The brainstem, which connects the cerebrum and cerebellum to the spinal cord, autonomously regulates essential functions such as breathing, heart rate, body temperature, sleep-wake cycles, and digestion. It also controls reflexive actions like sneezing, coughing, vomiting, and swallowing [35].

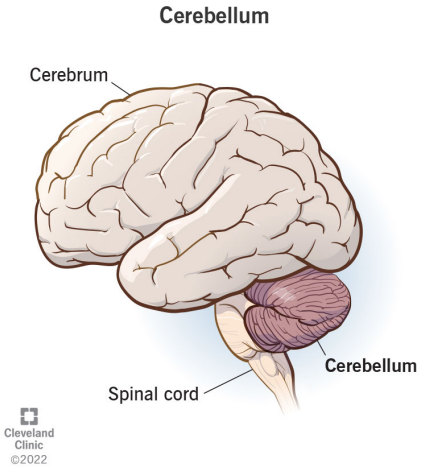


Figure 2.1: Illustration highlighting the primary regions of the brain. From [58].

2.2 Glioma

Brain tumors are commonly seen in clinical settings, highlighting the importance for general medical providers to possess a basic understanding of their diagnosis and treatment. Among the most common types of brain tumors are intracranial metastases from systemic cancers, meningiomas, and gliomas, with a particular focus on glioblastoma [57].

Gliomas, a varied group of primary central nervous system (CNS) tumors, arise in the brain or spinal cord and are identified by cells that resemble glial cells—normal brain cells that support and protect nerve cells [19]. These tumors are the most prevalent type of malignant primary CNS tumors in adults, with an annual incidence of about 6 cases per 100,000 globally. Notably, the incidence is significantly lower in non-Caucasian populations, particularly for glioblastoma [79]. Traditionally, gliomas have been classified based on their histopathological characteristics. However, advances in cancer genomics have transformed this approach by incorporating molecular profiles into glioma diagnoses [76]. This shift acknowledges that different genomic signatures correlate with different clinical outcomes, necessitating tailored treatment strategies and expectations [12].

The new 2021 World Health Organization (WHO) classification categorizes gliomas into four groups [12, 56]:

1. Adult-type diffuse gliomas
2. Pediatric-type diffuse low-grade gliomas
3. Pediatric-type diffuse high-grade gliomas
4. Circumscribed astrocytic gliomas

This updated classification underscores the importance of molecular profiling in accurately diagnosing and tailoring treatments for gliomas, reflecting the evolving understanding of these complex tumors.

2.2.1 Classification of Adult-Type Diffuse Gliomas

Adult-type diffuse gliomas, which represent the majority of gliomas, are classified into the following categories:

1. Astrocytoma, IDH-mutant (grade 2/3/4)
2. Oligodendroglioma, IDH-mutant (grade 2/3) and 1p/19q-codeleted
3. Glioblastoma, IDH-wildtype (grade 4)

Figure 2.2 provides an overview of this classification system. As depicted, the IDH mutation status is a critical factor in categorizing adult-type diffuse gliomas in the WHO classification of brain tumors, emphasizing its growing importance in therapeutic decision-making.

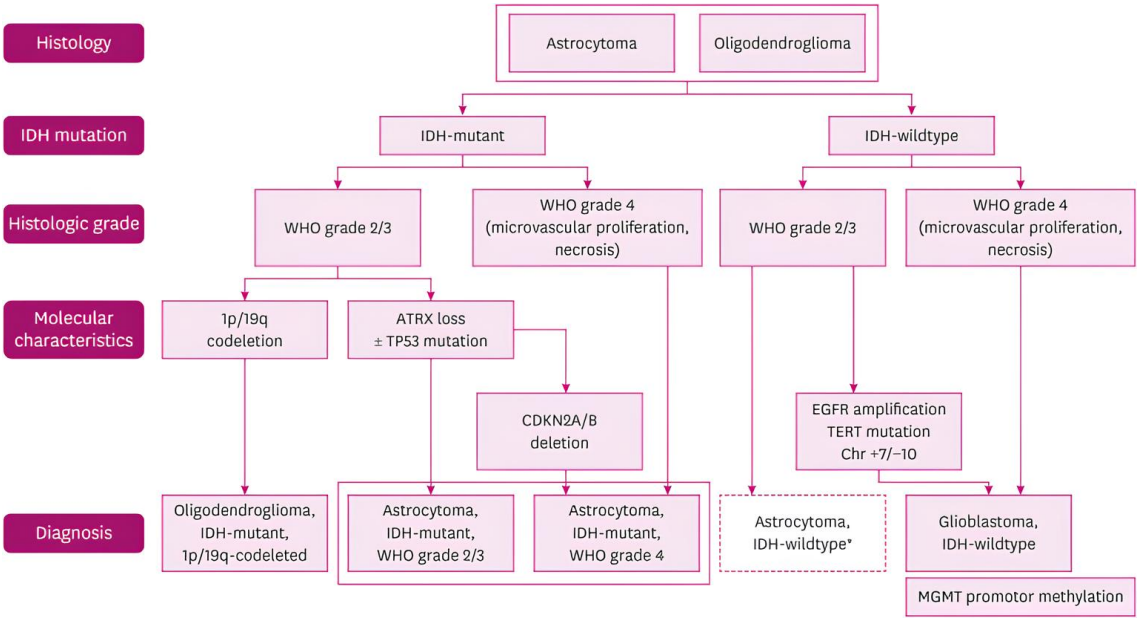


Figure 2.2: The 2021 WHO classification of adult-type diffuse gliomas, highlighting the importance of IDH mutation status. *IDH-wildtype glioma without high-grade features (WHO grade 2/3) should be investigated further and classified into other categories. From [12].

GBM is classified as a grade 4 IDH-wildtype glioma and is recognized as the most common and aggressive form of malignant primary brain tumor. It accounts for approximately 57% of all gliomas and 48% [75] of all primary malignant tumors in the CNS tumors. Despite the recent progress in employing a combination of treatments such as surgery, radiotherapy, systemic therapy, and supportive care for glioblastoma, the overall prognosis remains unfavorable, with rare cases of long-term survival, with a median survival of just one year [75]. This poor prognosis lies in

the molecular and genomic heterogeneity of GBM, leading to varied treatment responses. Studies have shown that gliomas with IDH mutations are associated with longer overall survival compared to their IDH-wildtype counterparts [34, 81, 48].

2.3 Isocitrate dehydrogenase

As discussed in Section 2.2, IDH status holds significant clinical implications, influencing prognostic assessments and dictating treatment strategies. The current section delves into the normal function of IDH, its abnormal functions when mutated, and the clinical ramifications of these mutations, particularly in the context of gliomas.

2.3.1 Normal Function of IDH

Isocitrate dehydrogenases are enzymes that play a crucial role in cellular metabolism. , particularly within the citric acid cycle (Krebs cycle). There are three isoforms of IDH: IDH1, IDH2, and IDH3 [30].

- **IDH1 and IDH2:** These isoforms are located in the cytoplasm and mitochondria, respectively. They are involved in producing a molecule vital for cellular defense against oxidative stress.
- **IDH3:** This isoform is a mitochondrial enzyme that directly participates in the citric acid cycle, playing a critical role in energy production.

2.3.2 Abnormal Function of IDH

When IDH1 and IDH2 are mutated, their enzymatic function changes. Instead of producing 2-oxoglutarate, these mutated enzymes gain the ability to produce 2-hydroxyglutarate, which accumulates in cells. This metabolite has oncogenic properties, contributing to tumor development and progression. Additionally, 2-hydroxyglutarate can activate certain receptors in the brain, increasing the likelihood of seizures. This explains the higher incidence of seizures in diffuse gliomas with IDH mutations, such as astrocytomas and oligodendrogliomas [30].

The frequency of IDH mutations in gliomas is significantly lower than that of wild-type IDH, with approximately 12% of gliomas harboring IDH mutations. Among these, IDH1 mutations are more common than IDH2 and IDH3 mutations, accounting for about 90% of cases [30]. This imbalance in the distribution of IDH mutation status in gliomas highlights its importance, impacting both diagnostic evaluations and therapeutic approaches.

2.4 Techniques for Brain Tumor Diagnosis

In 1924, the German psychiatrist Hans Berger performed the first human electroencephalography (EEG) [10]. This pioneering EEG could detect the brain's electrical waves, which varied as different brain cells interacted. Since then, neuroimaging techniques have advanced significantly and are now vital tools for neurology and mental health professionals.

Imaging plays a crucial role in the diagnosis, characterization, and treatment planning of GBMs. Various imaging modalities provide complementary information about the tumor's anatomical, functional, and molecular characteristics, which are essential for predicting the IDH mutation status.

The most common brain tumor diagnosis techniques are:

- Computed tomography;
- Magnetic Resonance Imaging;
- Positron Emission Tomography;
- Electroencephalography.

2.4.1 Computed Tomography

Computed tomography (CT) scans can reveal the presence of a tumor and related findings, through a combination of multiple X-ray images captured from different angles to generate detailed images of body tissues. However, they may fail to detect small tumors when diagnosing glioblastoma. Overlooking a small, low-grade glioma in a screening study might lead to its progression into glioblastoma. Additionally, CT scans may not capture all multifocal lesions, and diagnosing cerebrospinal fluid spread, especially in its early stages, can be challenging with CT imaging. Despite these limitations, CT scans offer valuable information on factors such as calcification or hemorrhage and prove beneficial for patients unable to undergo MRI [54].



Figure 2.3: Normal CT sc an of the brain, illustrating standard anatomical structures without abnormalities. From [15].

2.4.2 Magnetic Resonance Imaging

MRI, compared to CT, is significantly more sensitive in detecting tumors and associated findings, including peritumoral edema. It is the preferred modality for evaluating patients suspected or confirmed to have glioma, being the imaging modality with the highest degree of confidence [54]. MRI uses a magnetic field and computer-generated radio waves to produce detailed images of the organs and tissues, avoiding the use of ionizing radiation. Instead, it may require the administration of a contrast substance known as gadolinium (Gd). Gadolinium-based contrast agents enhance the visibility of abnormal tissues and blood-brain barrier disruptions. They pose a lower risk of allergic reactions compared to iodinated contrast materials used in CT scans.

MRI can utilize a variety of sequences to obtain detailed images, each providing different information about the tissues. Common sequences include [36]:

- **T1-weighted imaging (T1) -** These sequences produce high-contrast images of tissues with different water and fat content. In T1-weighted images, fatty tissues are bright, while areas that contain a lot of water are dark [20].
- **T1GD (T1 with Gadolinium Contrast Enhancement) -** These sequences uses a gadolinium-based contrast agent to enhance the visibility of blood-brain barrier disruptions, highlighting areas with abnormal vascularity.
- **T2-Weighted Imaging (T2) -** These sequences produce images in which areas with a lot of water are bright and fatty tissues are dark. They are particularly useful in detecting pathologies such as edema or tumours [20].
- **Fluid Attenuated Inversion Recovery (FLAIR) -** This sequence is a special type of T2-weighted sequence that suppresses the cerebrospinal fluid signals, making it easier to identify lesions in the brain by providing high contrast between normal and abnormal tissues [20].
- **Diffusion-weighted imaging (DWI) -** This technique measures the diffusion of water molecules within tissues, making it particularly effective for detecting early signs of stroke [20].
- **Perfusion-weighted imaging (PWI) -** Measures the passage of blood through the brain's vasculature. It provides information about blood volume, flow, and permeability within brain tissues, which is crucial for assessing tumor angiogenesis and distinguishing between high-grade and low-grade gliomas [60].

The characteristics of MRI images are influenced by the magnetic field produced by the scanner and various timing parameters. T1-weighted and T2-weighted scans are the most frequently used MRI sequences, with FLAIR being the third most common. Figure 2.4 illustrates a comparison of these primary imaging sequences.

Each MRI sequence offers unique advantages, allowing radiologists to comprehensively assess brain pathology and accurately diagnose conditions such as gliomas. The ability to combine these sequences provides a detailed and multifaceted view of the brain, facilitating precise treatment planning and monitoring.

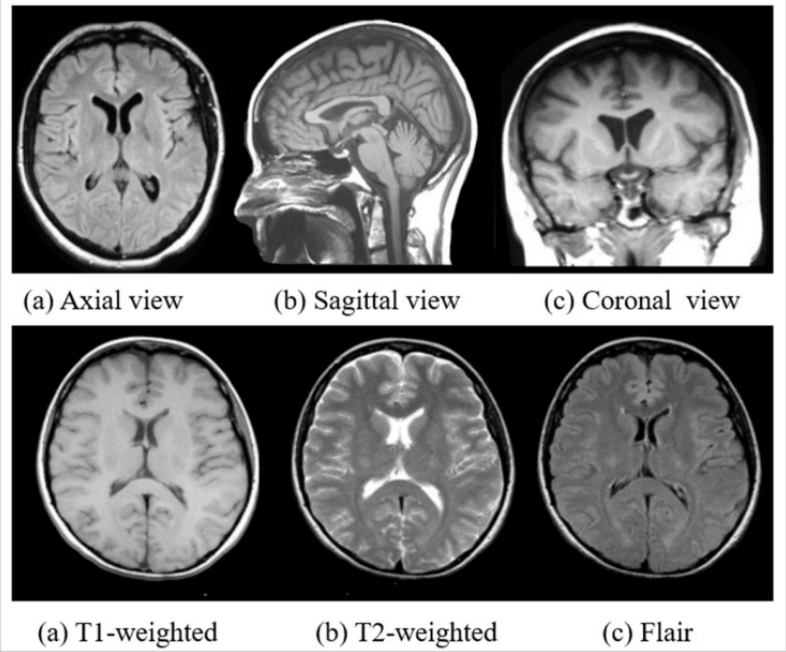


Figure 2.4: The three prominent views of MRI scanning: axial, sagittal, and coronal (top). The bottom section compares brain MRI slices of T1-weighted, T2-weighted, and FLAIR sequences, highlighting the differences in tissue contrast. From [68]

2.4.3 Positron Emission Tomography

A Positron Emission Tomography (PET) scan is an imaging technique that provides insights into the metabolic and functional activities of tissues and organs. PET scans use radioactive tracers, most commonly fluorodeoxyglucose (FDG), which mimic glucose and accumulate in areas of high metabolic activity. This is particularly useful in oncology because cancer cells typically exhibit higher glucose metabolism compared to normal cells [9].

Upon administration, the tracer is taken up by active cells, and the PET scanner detects the emitted positrons, creating images that reflect metabolic activity. PET scans are particularly valuable in glioblastoma for distinguishing between tumor recurrence and radiation necrosis, assessing treatment response, and guiding biopsy to the most metabolically active tumor region.

Compared to MRI, which excels in providing detailed anatomical images, PET is superior for evaluating functional and metabolic aspects. This distinction makes PET an excellent complementary tool to MRI, especially in the comprehensive assessment of glioblastomas.

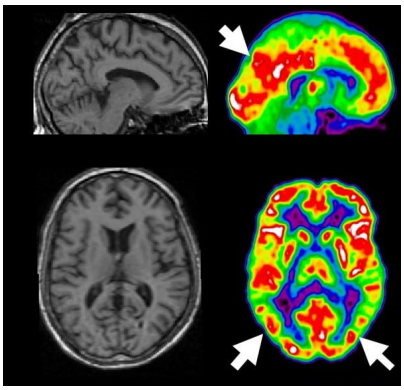


Figure 2.5: MRI brain scan (left), PET brain scan from a "healthy" patient (middle). Addapted from [9].

2.4.4 Electroencephalography

EEG is a non-invasive technique that records electrical activity of the brain through electrodes placed on the scalp [9]. The signals recorded reflect the synchronous activity of large groups of neurons, primarily in the cerebral cortex. These signals are captured by electrodes and amplified to produce waveforms that represent various states of brain activity, such as wakefulness, sleep, or pathological conditions. While it is traditionally used to diagnose and monitor neurological conditions such as epilepsy, EEG also provides valuable information in the context of glioblastomas. For example, EEG is essential in seizure detection, which is a common symptom in GBM patients.

2.4.5 Summary

Neuroimaging techniques have evolved significantly since Hans Berger's pioneering work with EEG in 1924. These advancements have made imaging an indispensable tool in the diagnosis, characterization, and treatment planning of gliomas. Each technique offers unique and complementary insights into the anatomical, functional, and molecular characteristics of brain tumors.

- **Computed Tomography** Provides valuable information on calcification and hemorrhage, and serves as a useful alternative for patients unable to undergo MRI.
- **Magnetic Resonance Imaging** Offers superior sensitivity and specificity in detecting tumors and associated abnormalities, making it the preferred modality for GBM evaluation.
- **Positron Emission Tomography** Excels in evaluating the metabolic and functional aspects of tissues, particularly valuable for distinguishing tumor recurrence from radiation necrosis and guiding biopsies.
- **Electroencephalography** Non-invasively monitors the brain's electrical activity, essential for seizure detection and functional assessment in GBM patients.

By integrating these diverse techniques, clinicians can achieve a comprehensive understanding of glioma characteristics, leading to more accurate diagnoses, better treatment planning, and

improved patient outcomes. The ongoing development of advanced imaging techniques continues to enhance our ability to non-invasively assess molecular and functional properties of gliomas, paving the way for personalized and effective therapeutic strategies.

2.5 Radiogenomics

Radiogenomics, an interdisciplinary field combining "Radiomics" and "Genomics" with the power of Artificial Intelligence, has emerged as a state-of-the-art science in precision medicine, particularly in oncology care. This innovative methodology establishes statistical correlations between the radiomic features of tumors and their underlying genetic composition. This correlation of an imaging phenotype with a specific genotype through radiogenomics presents a potential solution to the challenges associated with traditional biopsy sampling.

2.5.1 Conventional and Deep Radiomics

Radiomics methods are vital for the objective and reproducible identification and quantification of complex data patterns that are imperceptible to the human eye. These methods complement and support clinical decision-making, aiding in determining the most effective treatment options for glioma patients [7].

Radiomics technology is increasingly recognized in cancer research for its contribution to personalized treatment strategies, although it can be applied to medical imaging for various diseases. It has been effectively used across different anatomical regions in cancer imaging using multiple modalities, including MRI, CT, and PET scans. Traditional radiomics models primarily depend on manually crafted features extracted from radiological images [3]. These features encompass a broad spectrum, including texture, geometric, intensity, shape, histogram, morphological, among others [68].

In contrast, Deep Radiomics employs deep learning techniques to automatically learn and extract relevant features from images. This approach can identify more complex patterns and interactions within the data, making it less reliant on manual feature selection. Consequently, deep radiomics can adapt more effectively to various imaging modalities and clinical scenarios [3].

Figure 2.6 illustrates the differing methodologies of traditional radiomics and deep learning approaches in extracting phenotypic information from medical images.

2.5.2 Genomics

Genomics focuses on the analysis of an organism's complete set of DNA, including all of its genes. In the context of oncology, genomics aims to identify genetic mutations and alterations that drive cancer progression and influence response to treatments. Techniques such as whole-genome sequencing and targeted gene panels are used to uncover the genetic landscape of tumors, providing insights into their behavior and potential therapeutic targets [3].

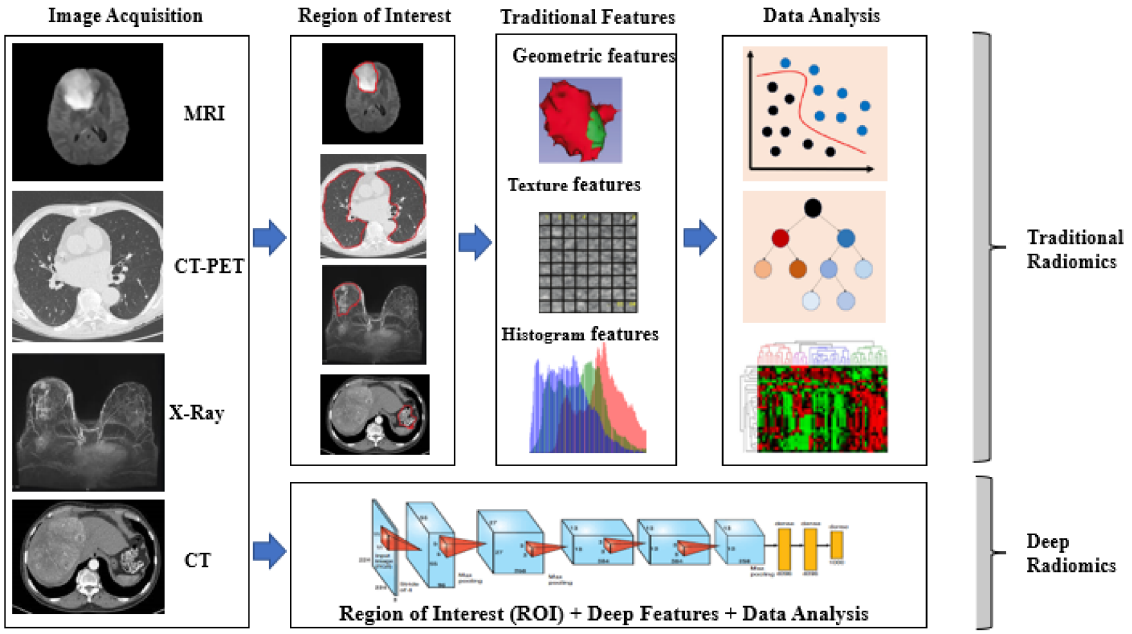


Figure 2.6: Comparison between traditional and deep radiomics: Traditional radiomics involves multiple steps such as ROI detection, feature extraction, selection, and analysis, whereas deep radiomics consolidates all these steps into a single process. From [3]

2.5.3 Integration into Radiogenomics

The integration of radiomics and genomics into radiogenomics creates a powerful tool for precision medicine. Radiogenomics enables the non-invasive prediction of specific genetic mutations and chromosomal alterations through imaging features, effectively creating "virtual biopsy" maps [56]. These maps offer in vivo markers that reflect the spatial and molecular heterogeneity of tumors, allowing for more precise patient stratification and personalized treatment planning. Furthermore, radiogenomics facilitates "real-time" monitoring of treatment efficacy, which is vital for optimizing therapeutic outcomes in personalized medicine [22].

The radiogenomics process typically involves workflow encompassing six distinct stages, as outlined in Figure 2.7 [68]:

1. **Acquisition of Radiological Images:** The initial step involves obtaining high-quality radiological images.

2. **Preprocessing Steps:** The images are then preprocessed to enhance their quality and suitability for analysis.

3. **Feature Extraction:** Relevant radiomic features are extracted from the preprocessed images.

4. **Feature Selection / Dimensionality Reduction:** Techniques are applied to reduce the dimensionality of the extracted features, making the data more manageable and interpretable.

5. **Association of Radiomics and Genomics Features:** Radiomic features are correlated with genomic data to derive meaningful insights into the genetic characteristics of tumors.

6. **Data Analysis:** The final step involves analyzing the data to identify significant correlations and derive actionable insights.

This analysis offers valuable insights into the genetic composition of tumors, without the need for invasive biopsy procedures, paving the way for more precise and personalized oncology care.

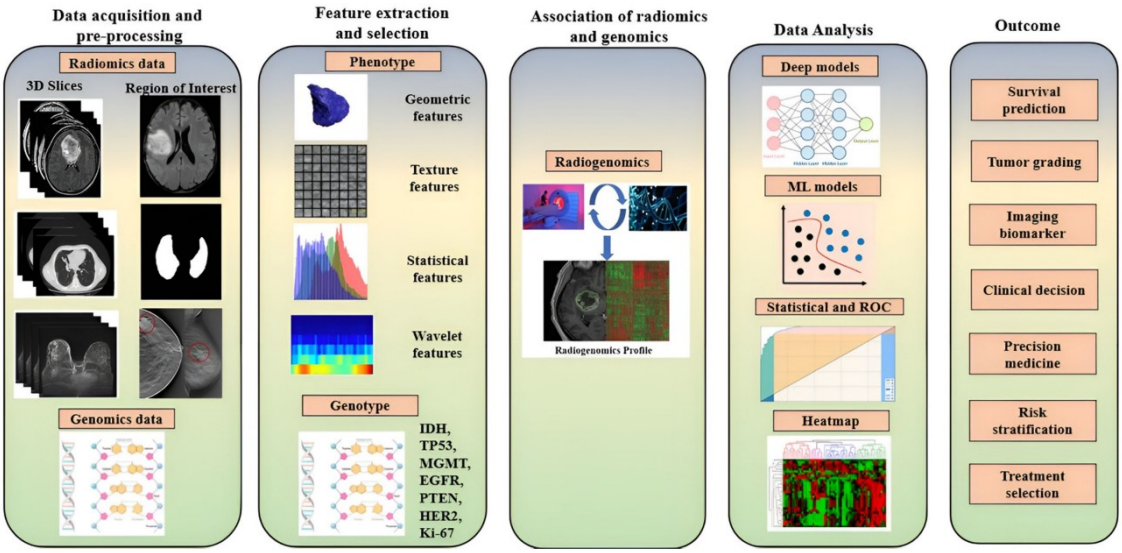


Figure 2.7: Radiogenomics pipeline, illustrating the steps from data acquisition and preprocessing, through feature extraction and selection, to the association of radiomics and genomics, culminating in data analysis to identify different outcomes. From [68].

2.5.4 Data acquisition

Data acquisition is the foundational stage in the radiogenomics workflow, involving the collection of high-quality radiological images. This stage is important to ensure the reliability of subsequent analyses. It can be conducted through diverse modalities like CT scans, MRI, PET, among others, that must be selected based on the clinical objectives and the specific characteristics of the tumor under investigation. Each modality provides distinct insights into the tumor's morphology and functional attributes, as discussed in Section 2.4.

2.5.5 Preprocessing

Preprocessing is the second critical stage in the radiogenomics workflow. This stage involves preparing the acquired radiological images for subsequent analysis by enhancing their quality and standardizing their format. Key components of this stage include noise reduction, normalization, and segmentation, all of which are essential for accurate feature extraction and analysis. Specifically, for radiomics data derived from cancerous images, addressing issues like class imbalance, data augmentation, randomization, and standardization is crucial [47].

- **Noise Reduction:** Reducing noise in the images improves clarity and ensure that subtle features are not obscured. Techniques such as filtering and smoothing can be applied to remove random variations in the image data that are not related to the tumor or surrounding tissue.
- **Normalization:** Normalizing the images involves adjusting them to a common scale or reference, which helps in comparing images across different patients or imaging sessions. This step is essential for ensuring that variations in image intensity are due to biological differences rather than technical inconsistencies.
- **Segmentation:** Segmentation is the process of delineating regions of interest (ROI) within the images. This step is crucial for isolating the tumor and relevant anatomical structures from the surrounding tissues. This can be achieved using various procedures, including manual, semi-automated, or fully automated methods [5].
  - **Manual Segmentation:** Involves a radiologist or trained expert manually outlining the ROI. This method is highly accurate but can be time-consuming and subject to inter-observer variability.
  - **Fully Automated Segmentation:** Utilizes algorithms and AI techniques to automatically delineate the ROI. This method is faster and can reduce variability, though it requires validation against manual segmentation to ensure accuracy.
  - **Semi-Automated Segmentation:** Combine manual and automated methods to leverage the strengths of both.

  Currently, manual segmentations are acknowledged as the benchmark for radiomic analysis due to their ability to ensure precision [5]. However, the field has seen the emergence of various automated segmentation approaches, particularly those leveraging deep learning (DL). Moreover, semi-automatic algorithms are gaining attention as viable alternatives in the quest for automated segmentation methodologies [5].

2.5.6 Feature Extraction

Feature extraction in radiomics involves obtaining quantitative data from medical images that can be utilized for various analytical purposes, including lesion detection and classification. This process can involve hand-crafted features, leading to a conventional radiomics approach, or features extracted through advanced deep learning techniques.

Radiomic features can be categorized into several types, such as statistical, geometric, and textural attributes. These features provide detailed information about the tumor's shape, intensity, and texture, which can be crucial for accurate diagnosis and treatment planning.

In radiogenomics, feature extraction extends to include genomic data. This involves identifying various genomic markers or genotypic features, such as MGMT, IDH, and EGFR [5], which correspond to the radiomic features of the patient. Integrating radiomic and genomic data allows for the mapping of phenotype and genotype information, enabling comprehensive analysis and deeper insights into disease mechanisms.

2.5.7 Feature Selection / Dimensionality Reduction

After extracting all relevant features, it is essential to perform feature selection to pinpoint the most pertinent ones for subsequent analysis. This step is vital for reducing the dataset's dimensionality and enhancing the efficiency and performance of machine learning models. The following is a comparison between Feature Selection and Dimensionality Reduction, as well as a comparison between different techniques.

Both feature selection and dimensionality reduction are key techniques in data processing and machine learning, particularly with high-dimensional datasets such as radiogenomic features. However, they have different objectives and methods. Feature selection focuses on identifying a subset of the most relevant features from the original dataset, thereby removing those that are redundant or irrelevant. On the other hand, dimensionality reduction involves transforming the original features into a new set of variables, usually of lower dimensions, while retaining as much essential information as possible. While feature selection retains the original features, dimensionality reduction creates new features, making it a powerful choice when the goal is to reduce data complexity [28].

The role of feature selection and dimensionality reduction in machine learning is:

1. **To improve model performance:** Reducing the number of features can lead to more efficient and effective learning processes. It simplifies the model, reduces overfitting, and often results in better generalization on unseen data.

2. **To improve Computational Efficiency:** Fewer features mean less computational resources are required for processing. This is particularly important in applications involving large datasets.

3. **To improve interpretability:** A model with fewer, more meaningful features is easier to interpret and understand, which is crucial in domains like medical imaging and genomics.

2.5.7.1 Feature Selection

Feature selection methods can be categorized into three types: filter methods, wrapper methods, and embedded methods.

1. **Filter Methods:**

   Filter methods evaluate the relevance of features by examining their intrinsic properties, independent of the learning algorithm. Some examples include:

   - **Pairwise Correlation:** Measures the linear relationship between each feature and the target variable or between pairs of features. Highly correlated features can be redundant, so one of them can be removed [28].

   - **Chi-Square Test (X^2):** Assesses the independence between categorical features and the target variable. It compares the observed values from different attributes of the dataset to its expected value [28].

   $$X^2 = \sum \frac{(Observed\ value - Expected\ value)^2}{Expected\ value} \tag{2.1}$$

   - **Mean Absolute Difference (MAD):** Measures the average absolute difference between all pairs of feature values. This metric helps identify features with high variability, which are likely to be more informative [28].

   - **Mutual Information:** Measures the amount of information that a feature provides about the target variable. The higher the mutual information between a feature and the target variable, the more informative the feature is, and the more relevant it is to the predictive model [72].

2. **Wrapper Methods:**

   Wrapper methods evaluate subsets of features based on their performance with a specific learning algorithm. Based on the results obtained from training before the model, the addition and removal of features occur. Unlike filter methods, these consider feature interactions and are often more accurate. Some examples include:

   - **Recursive Feature Elimination:** Iteratively removes the least important features based on model performance. Initially, the estimator is trained on the complete set of features, and the importance of each feature is assessed using a specific attribute or function. The least important features are then removed from the current set. This process is recursively repeated on the reduced set of features until the desired number of features is reached [71].

- **Forward Selection:** Starts with an empty set of features and adds features one by one based on their contribution to the model's performance. The stopping criterion is until adding a new variable does not improve the model's performance [28].

- **Boruta:** Boruta is a feature selection algorithm that does not require a subjective threshold. It first creates *shadow features*, which are randomly shuffled versions of each original feature. Then, it computes the importance of each *shadow feature*, and the highest score among these becomes the threshold. Finally, if the importance of the original feature is higher than this threshold, the feature is retained; otherwise, it is discarded [2].

3. **Embedded Methods:**

   Embedded methods perform feature selection as part of the model training process. Embedded methods encounter the drawbacks of filter and wrapper methods and merge their advantages. Some examples include:

   - **Lasso Regression (L1 Regularization):** Adds a penalty term to the regression model that shrinks the coefficients of less important features to zero. The features having zero coefficient can be removed from the dataset [28].

   - **Tree-Based Methods:** Decision trees, Random Forests and Gradient Boosting naturally perform feature selection by selecting the most informative features for splitting nodes [28].

2.5.7.2 Dimensionality Reduction

Dimensionality reduction can be accomplished using a variety of techniques, including [66]:

- **Principal Component Analysis (PCA):** PCA is a statistical technique used for dimensionality reduction that transforms original features into a new set of linearly uncorrelated components called principal components. These components are chosen based on their ability to explain the variance in the data. The first principal component explains the most variance in the data, and each successive component explains less.

  PCA begins by computing the covariance matrix of the data to understand how variables relate. Eigenvalues and eigenvectors are then calculated from this matrix, where eigenvectors indicate directions of maximum variance and eigenvalues measure the magnitude of variance in these directions. The data is projected onto these eigenvectors, transforming it into a new coordinate system defined by the principal components [49].

  By selecting the top principal components, PCA reduces dimensionality while retaining most of the data's variance. This orthogonal transformation ensures no redundancy in the

feature space.

- **t-Distributed Stochastic Neighbor Embedding (t-SNE) [67]:** t-SNE is a non-linear dimensionality reduction technique primarily utilized for data visualization. It begins by evaluating the similarity between each pair of data points in the high-dimensional space. The technique then aims to map these similarities to a lower-dimensional space, striving to preserve the relationships between the points as accurately as possible. Through an iterative process, t-SNE adjusts the positions of the points in the reduced space to closely match the similarities found in the original high-dimensional space. Finally, the points are plotted in the lower-dimensional space, often resulting in a graph that makes patterns and structures within the data easier to recognize and analyze. Table 2.1 and Figure 2.8 highlight the distinctions between PCA and t-SNE.

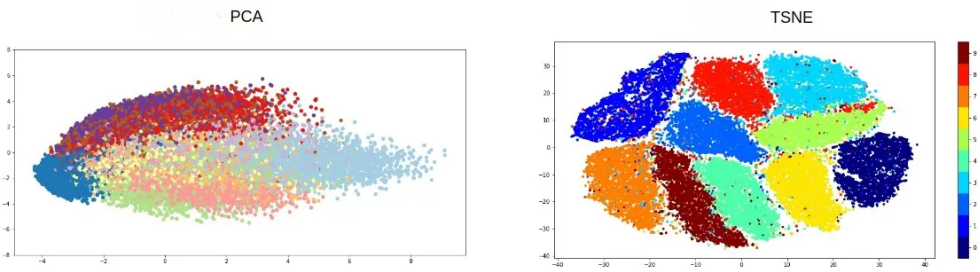


Figure 2.8: Comparison of PCA and t-SNE. Adapted from [64]

Table 2.1: Comparison of PCA and t-SNE - techniques for dimensionality reduction. Adapted from [27].

PCA	t-SNE
Linear Dimensionality reduction technique.	Non-linear Dimensionality reduction technique.
Preserve the global structure of the data.	Preserve the local structure(cluster) of data.
It does not involve Hyperparameters.	It involves Hyperparameters such as perplexity, learning rate and number of steps.
Highly affected by outliers.	Can handle outliers.
Deterministic algorithm.	Non-deterministic or randomised algorithm.
It works by rotating the vectors for preserving variance.	It works by minimising the distance between the point in a Gaussian.
Computationally less expensive.	Can be computationally expensive, especially for high-dimensional datasets.
It can be used for visualization of high-dimensional data in a low-dimensional space.	It is specifically designed for visualization and is known to perform better in this regard.
Suitable for linearly separable datasets.	More suitable for non-linearly separable datasets.
It can be used for feature extraction	It is mainly used for visualization and exploratory data analysis.

- **Linear Discriminant Analysis (LDA):** LDA is a supervised dimensionality reduction technique that aims to maximize the separability between different classes. It projects the data onto a lower-dimensional space by maximizing the ratio of between-class variance to within-class variance.

2.5.7.3 Summary:

Table 2.2 sums up the differences between Feature Selection and Dimensionality Reduction.

Table 2.2: Difference between Feature Selection and Dimensionality Reduction. Adapted from [66].

	Feature Selection	**Dimensionality Reduction**
Definition	Selecting a subset of the most relevant features from the dataset	Reducing the number of dimensions in a dataset while retaining as much information as possible
Techniques	Filter methods, wrapper methods, embedded methods	PCA, t-SNE, LDA, etc.
Goal	Identify the most relevant and important features for predicting the target variable	Improve performance and interpretability of a machine learning model
Impact on data	Keeps a subset of the original features	Transforms the data into a lower-dimensional space
Impact on model	Can improve performance and interpretability by reducing the number of features and risk of overfitting	Can improve performance and interpretability by reducing complexity and risk of overfitting

2.5.8 Association of Radiomics and Genomics Features

During this phase, radiomics and genomics features from cancer patients are integrated to gain insights into the tissue-level characterization of regions. This combination enables the differentiation of cancerous and non-cancerous regions using radiomics features and facilitates the identification of specific gene mutations.

2.5.9 Performance Assessment:

Performance assessment is a required step to evaluate the effectiveness of machine learning models. In radiogenomics, employing robust evaluation metrics is crucial to ensure that the models deliver reliable predictions.

This stage serves as a crucial decision support system, impacting a wide range of clinical outcomes. These outcomes include tumor grading, predicting patient survival, imaging biomarker generation, aiding in clinical decision-making, facilitating precision medicine, conducting risk stratification, and formulating personalized treatment plans for cancer patients [68].

Key perfomance metrics include:

1. **Accuracy:** The ratio of correctly predicted instances to the total instances.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{2.2}$$

where TP denotes True Positives, TN denotes True Negatives, FP denotes False Positives, and FN denotes False Negatives.

2. **Precision:** The ratio of correctly predicted positive instances to the total predicted positive instances. Precision is useful when the cost of false positives is high.

$$Precision = \frac{TP}{TP+FP} \tag{2.3}$$

3. **Recall (Sensitivity):** The ratio of correctly predicted positive instances to the total actual positive instances. Recall is important when the cost of false negatives is high.

$$Recall = \frac{TP}{TP+FN} \tag{2.4}$$

4. **Specificity:** The ratio of correctly predicted negative instances to the total actual negative instances. Specificity is important when the cost of false positives is high.

$$Specificity = \frac{TN}{TN+FP} \tag{2.5}$$

5. **Balanced Accuracy:** Adjusts for imbalanced datasets by averaging the true positive rate (sensitivity) and the true negative rate (specificity).

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{2.6}$$

6. **F1 Score:** The harmonic mean of precision and recall. F1 Score is a balanced metric that is useful when both precision and recall are important.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.7}$$

7. **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** AUC-ROC represents the likelihood that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. AUC-ROC is useful for evaluating the performance of classification models, particularly when dealing with imbalanced datasets.

2.6 Radiomic Features

In radiogenomics, feature extraction involves the use of algorithms to capture tumor heterogeneity from local pixel or voxel neighborhoods. However, the widespread adoption of radiomics in clinical practice has been hindered by several factors. These include the lack of standardized methods for defining and calculating radiomic features, inconsistent terminology, and the limited reproducibility of radiomic studies [6].

To address these issues, the Image Biomarker Standardization Initiative (IBSI) has published a detailed reference manual [83]. This manual provides standardized definitions for 174 radiomic features, an image processing framework for radiomics, and guidelines for reporting. The increased awareness of the need for standardization has led several radiomic tools to conform to the IBSI guidelines, improving the consistency and reliability of radiomic studies.

IBSI grouped the radiomic features into 11 classes, which can be categorized into three major categories: [6, 42]:

1. **Statistical / Histogram features**
   (a) Local Intensity (2 features)
   (b) Intensity-based Statistic (18 features)
   (c) Intensity Histogram (23 features)
   (d) Intensity-volume Histogram (7 features)
2. **Texture features**
   (a) Gray-Level Co-occurrence Matrix (GLCM) (25 features)
   (b) Gray-Level Run Length Matrix (GLRLM) (16 features)
   (c) Gray-Level Size Zone Matrix (GLSZM) (16 features)
   (d) Gray-Level Distance Zone Matrix (GLDZM) (16 features)
   (e) Gray-Level Dependence Matrix (GLDM) (17 features)
   (f) Neighborhood Gray Tone Difference Matrix (NGTDM) (5 features)
3. **Morphology (shape) features**
   (a) Morphology features (29 features)

2.6.1 Statistical / Histogram Features

These features provide a quantitative measure on the distribution of voxel intensities within the tumor region, offering information about the tumor's overall intensity characteristics. They can be classified into four different components [42]:

1. **Local Intensity:** Measures the intensity of pixels or voxels in a localized neighborhood.
2. **Intensity-based Statistic:** Describes statistical properties of the intensity values, such as mean, variance, skewness, and kurtosis.
3. **Intensity Histogram:** Captures the frequency distribution of intensity values, providing a summary of the intensity range and its distribution within the tumor.
4. **Intensity-volume Histogram:** Combines intensity and spatial information to describe the relationship between intensity values and the volume of the tumor.

2.6.2 Texture Features

Textural features capture the spatial arrangement and intensity variations within the tumor, capturing patterns that can reflect the complexity and heterogeneity of the tumor microenvironment. They can be classified into six different components [42]:

1. **Gray-Level Co-occurrence Matrix:** GLCM is a statistical technique used to analyze the texture of an image by evaluating the spatial relationships between voxels. It measures the frequency of specific voxel pairs with particular gray levels occurring at a defined distance and angle. The commonly extracted features from GLCM, such as contrast, correlation, energy, and homogeneity, offer valuable information about the texture patterns and structure of the image.
2. **Gray Level Run Length Matrix:** GLRLM is another texture analysis method that measures the length of consecutive runs of voxels having the same gray level in a specified direction. It captures the texture in terms of the distribution of runs, offering features such as short-run emphasis, long-run emphasis, gray level non-uniformity, and run length nonuniformity. These features help in identifying the texture's granularity and pattern continuity.
3. **Gray Level Size Zone Matrix:** GLSZM extends texture analysis by examining the size of homogeneous zones of gray levels within an image. A zone is defined as a group of connected voxels that share the same gray level. The GLSZM quantifies the number and size of these zones, providing features like zone size non-uniformity, large zone emphasis, and gray level non-uniformity. This method is useful for characterizing the spatial distribution and the texture's uniformity.
4. **Gray-Level Distance Zone Matrix:** GLDZM analyzes the distance between homogeneous zones of the same intensity, offering insights into the spatial distribution of texture patterns.
5. **Gray Level Dependence Matrix:** GLDM measures the degree of dependence of a voxel on its neighbors, considering different levels of dependencies. It defines the number of voxels within a certain neighborhood that have a specific gray level. Features derived from the GLDM include dependence non-uniformity, dependence entropy, and dependence variance, which are useful for characterizing the texture's complexity and homogeneity.

6. **Neighbouring Gray Tone Difference Matrix:** NGTDM assesses texture by evaluating the difference between a voxel's gray level and the average gray level of its neighbors within a defined distance. It captures local texture information, such as coarseness, contrast, busyness, complexity, and strength. NGTDM features are effective in differentiating textures based on their local intensity variations and edge sharpness.

2.6.3 Morphological features

Morphological features refer to characteristics related to the irregularities and aggressive infiltration of the tumor, leading to changes in the surface and shape within the tumor and peritumoral regions. These features can be broadly classified into local and global components [5].

- **Local Features** focus on specific aspects of the tumor's geometry:
  - **Curvature:** Differentiates flat areas from highly curved ones.
  - **Sharpness:** Measures the sharpness of curvature, with highly curved masses exhibiting sharper curvatures.
  - **Shape Index:** Characterizes the shape topology of the tumor, distinguishing various geometric shapes.
- **Global Features** provide an overview of the tumor's overall structure:
  - **Primary Axes:** Represents the ratio between the major and minor axes of the region of interest, providing a broad view of the tumor's shape.

In essence, morphological features aim to capture the nuanced geometric attributes of a tumor, offering valuable insights into its morphological characteristics.

2.6.4 Software Tools for Radiomic Features Extraction

Various software tools and platforms have been developed to facilitate automated feature extraction in the field of radiogenomics. These tools streamline the process of deriving meaningful features from medical images, ensuring consistency, accuracy, and efficiency. Some of the most widely used and publicly available software tools include Pyradiomics [77, 25], the Medical Imaging Interaction Toolkit (MITK) [32], LIFEx [61], the Standardized Environment for Radiomics Analysis (SERA) [26], and Cancer Imaging Phenomics Toolkit (CaPTk) [23], all available in Windows, Linux, and MacOS. Table 2.3 summarizes these radiomics software programs, highlighting their programming languages, supported data formats, and significant dependent libraries.

Table 2.3: Most widely used and publicly available software tools for radiomic features extraction.

	Language	Data Format	Significant Dependent Library
Pyradiomics	Python	DICOM, NIfTI, NRRD	SimpleITK
MITK	C++	DICOM	Insight Toolkit (ITK), Visualization Toolkit (VTK)
LIFEx	Java	DICOM, NIfTI, png, jpg	None
SERA	MATLAB	DICOM	None
CaPTk	C++	DICOM, NIfTI	OpenCV, ITK, VTK

DICOM = Digital Imaging and Communications in Medicine, NIfTI = Neuroimaging Informatics Technology Initiative, NRRD = Nearly Raw Raster Data.

Table 2.4 provides a detailed breakdown of the number of features within each category as defined by the IBSI standards, illustrating the coverage and focus areas of different tools.

Table 2.4: Distribution of IBSI-standardized features across various radiomic tools, including Pyradiomics, MITK, LIFEx, SERA and CaPTk. The table details the number of features in each category, illustrating the coverage and focus areas of different tools. Adapted from [50].

	Pyradiomics	MITK	LIFEx	SERA	CaPTk
Morphological features	14	20	2	25	3
Local intensity features	0	2	0	2	0
Intensity-based statistical features	15	18	5	18	15
Intensity histogram features	2	21	2	19	17
Intensity-volume histogram features	0	7	0	7	0
GLCM	24	25	6	25	6
GLRLM	16	16	11	16	4
GLSZM	16	16	11	16	16
GLDZM	0	0	0	16	0
GLDM	14	16	0	16	16
NGTDM	5	5	3	5	5

Each of these tools has its strengths and is suited to different aspects of radiomics research. The choice of tool often depends on the specific requirements of the study, including the type of imaging data, the desired features, and the level of customization needed.

2.7 Machine Learning Algorithms

In the field of machine learning, classification algorithms are essential tools for predicting outcomes based on data. This section explores several popular classification algorithms. The algorithms covered include Support Vector Machine, Random Forest, XGBoost, Logistic Regression, K-Nearest Neighbors, and Naïve Bayes. Each algorithm has unique characteristics and is suited to different types of data and classification tasks, making them valuable for medical image analysis and the prediction of disease-related mutations.

2.7.1 Support Vector Machine

SVM is a supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the data into different classes. The main idea is to maximize the margin between the hyperplane and the nearest data points from each class, known as support vectors. SVM is effective in high-dimensional spaces and is versatile with the use of different kernel functions to handle non-linear data [41].

Linear SVMs are suitable for linearly separable data, meaning no transformations are needed to distinguish between classes. The decision boundary is represented as $wx+b=0$, where w is the weight vector, x is the input vector, and b is the bias term. There are two approaches for calculating the margin:

- **Hard-Margin Classification:** This approach is used when the data is perfectly linearly separable. It aims to find the hyperplane that separates the classes without any misclassification. The margin is maximized as $\gamma = \frac{a}{||w||}$, where $||w||$ is the Euclidean norm of the weight vector and a is the margin projected onto w. Figure 2.9 illustrates this method, with hyperplane L2 being selected because it maximizes the distance to the nearest data point on either side (hard margin). However, this method is very sensitive to outliers and noise.

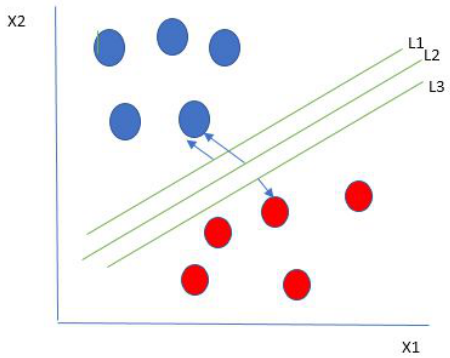


Figure 2.9: Multiple hyperplanes separate the data from two classes. From [63]

- **Soft-Margin Classification:** This approach allows for some misclassifications to handle cases where the data is not perfectly separable. The optimization problem becomes a trade-off between maximizing the margin and minimizing the classification error. The hyper-parameter C controls this trade-off, where a larger C puts more emphasis on minimizing misclassification errors, and a smaller C allows for a wider margin with more tolerance for misclassifications.

Nonlinear SVMs handle data that is not linearly separable by transforming it into a higher-dimensional feature space using preprocessing methods. This can increase complexity and risk of overfitting, but the "kernel trick" mitigates this by replacing dot product calculations with kernel functions. For example, if our data resembles the data shown in Figure 2.10 (Left), the solution involves creating a new variable using a kernel. This consists in selecting a point x_i on the line and generating a new variable y_i as a function of its distance from the origin, resulting in the representation shown on the right of the figure [41].

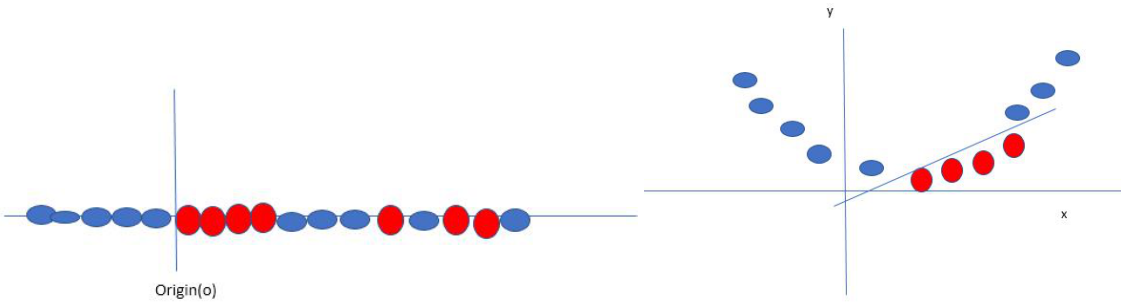


Figure 2.10: Nonlinear SVM. Left: Original 1D dataset for classification. Right: Mapping 1D data to 2D to become able to separate the two classes. From [63]

Popular kernels include:

- **Polynomial Kernel -** represents the similarity of vectors in a feature space over polynomials of the original variables, allowing the classifier to fit a wide range of complex decision boundaries.
- **Radial Basis Function (RBF) Kernel -** It can map the input space into a higher-dimensional space, effectively handling non-linear relationships.
- **Sigmoid Kernel -** This kernel is similar to the activation function of a two-layer perceptron neural network. It can map the input space into an infinite-dimensional space, allowing it to handle non-linear relationships. It is particularly useful in situations where the relationship between features can be captured by a function resembling the sigmoid activation in neural networks. However, it is not as widely used as the RBF kernel due to its sensitivity to the choice of parameters and potential issues with convergence.

SVM offer several advantages and disadvantages. Table 2.5 summarizes these key points.

Table 2.5: Advantages and Disadvantages of Support Vector Machines

Advantages	Disadvantages
Powerful in cases where the number of dimensions exceeds the number of samples	Computationally expensive, especially with large datasets
With appropriate kernel and parameters, SVM can be quite robust to overfitting	Tuning hyperparameters can be challenging and time-consuming
Only support vectors are used to define the decision boundary, making SVM memory efficient	The performance of SVMs can degrade with very large datasets, making them less scalable
The concept of margin maximization provides a clear and interpretable measure of the confidence of classification	SVMs can be sensitive to noise and overlapping classes, especially in the case of soft-margin

2.7.2 Random Forest Classifier

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs to enhance accuracy and mitigate overfitting. "Ensemble" refers to the practice of integrating multiple models, so instead of relying on a single model, a collection of models is utilized to make predictions [40]. There are two primary types of ensemble methods: Bagging and Boosting, as illustrated in Figure 2.11. Random Forest Classifier works on the Bagging principle, where multiple subsets of the training data are created with replacement, and individual models are trained on these subsets.

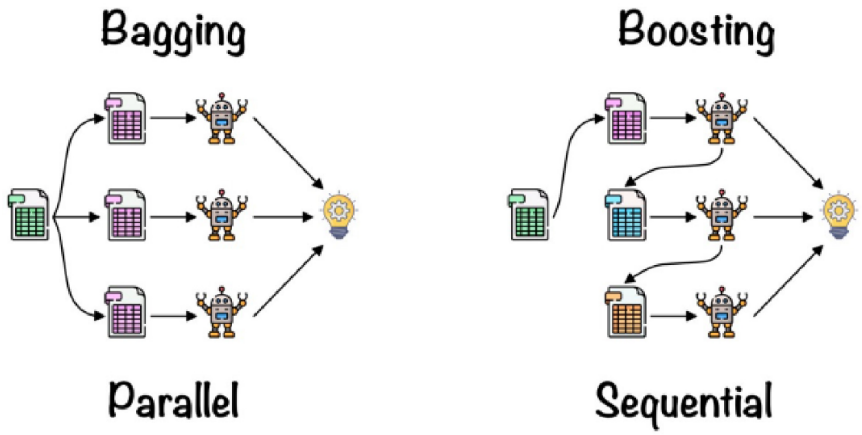


Figure 2.11: Ensemble types of methods. From [63].

Each tree in the Random Forest is constructed from a bootstrap sample of the data, which means a random sample of the training data with replacement. This approach ensures that some data points may be used multiple times, while others may be left out. The randomness in data

sampling helps create trees that are less correlated with each other, enhancing the overall model performance and robustness. Figure 2.12 illustrates the pipeline of a random forest classifier.

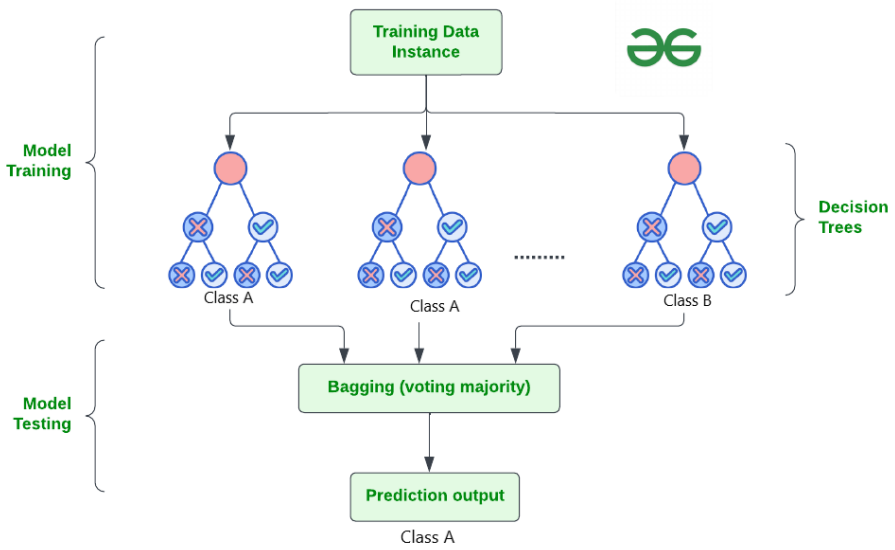


Figure 2.12: Random Forest Algorithm in Machine Learning. From [29]

At each split in a tree, only a random subset of features is considered for splitting. This method introduces diversity among the trees and reduces the likelihood of overfitting. The final output is determined by aggregating the results from all the trees through a majority voting process in classification tasks. The outcome that is most frequently predicted by the individual trees is selected as the final prediction in classification.

Random Forest also employs an out-of-bag (OOB) error estimation, where approximately one-third of the training sample, known as the OOB sample, is not used in the training of a particular tree. This OOB sample is used for cross-validation, providing an unbiased estimate of the model's performance without the need for a separate validation set.

Difference between Decision Tree and Random Forest

Although a Random Forest is a collection of decision trees, there are significant differences in their behavior. Table 2.6 provides a comparison between the two methods.

Table 2.6: Difference between Decision Tree and Random Forest

Aspect	Decision Trees	Random Forest
Basic Structure	Single tree structure	Ensemble of multiple trees
Overfitting Tendency	Prone to overfitting, especially with deep trees	Less prone to overfitting due to averaging
Variance	High variance in predictions	Lower variance due to averaging
Bias	Low bias, can capture complex details	Slightly higher bias due to simpler aggregated models
Interpretability	Easy to visualize and understand	Less interpretable, complex to visualize
Computational Cost	Lower; faster to train and predict	Higher; requires more resources
Robustness to Noise	Sensitive to noise	More robust to noise due to ensemble approach
Feature Importance	Clear indication of feature importance	Aggregated feature importance, more reliable
Handling of Imbalance	Can struggle with imbalanced datasets	More effective with imbalanced datasets
Decision Boundaries	Highly irregular decision boundaries	Smoother decision boundaries

Random Forest offer several advantages and disadvantages. Table 2.7 summarizes these key points.

Table 2.7: Advantages and Disadvantages of Random Forest Classifier

Advantages	Disadvantages
Handles missing data naturally by using surrogate splits	Training can be computationally intensive and memory-demanding
Provides estimates of feature importance, aiding in model understanding	Requires more resources, especially with a large number of trees and features
Training multiple trees can be parallelized, scalable to large datasets	Making predictions can be slower compared to simpler models

2.7.3 XGBoost (eXtreme Gradient Boosting)

XGBoost is a scalable, high-performance gradient boosting algorithm used for classification and regression tasks. It builds an ensemble of decision trees in a sequential manner, where each tree attempts to correct the errors of the previous ones. XGBoost incorporates regularization techniques, such as L1 and L2, to control overfitting and optimize performance [24].

XGBoost is then based on the concept of boosting, in contrast to Random Forest, which relies on bagging. While Random Forest combines the predictions of multiple base learners to produce a strong learner, XGBoost iteratively adds new trees that address the errors made by the existing ensemble, refining the overall predictive accuracy [78].

Comparison with Random Forest

Although both XGBoost and Random Forest are ensemble methods, they have distinct differences in their approach and application. Table 2.8 highlights key differences between these two algorithms:

Table 2.8: Differences between XGBoost and Random Forest. Adapted from [78]

Feature	XGBoost	Random Forest
Description	Improves mistakes from previous trees	Builds trees independently
Algorithm Type	Boosting	Bagging
Handling of Weak Learners	Corrects errors sequentially	Combines predictions of independently built trees
Regularization	Uses L1 and L2 regularization to prevent overfitting	Usually doesn't employ regularization techniques
Performance	Often performs better on structured data but needs more tuning	Simpler and less prone to overfitting

2.7.4 Logistic Regression

Logistic Regression is a statistical model used for binary classification tasks. It estimates the probability that a given input belongs to a particular class using the logistic function, also known as the sigmoid function. This function maps any real-valued number to a value between 0 and 1, producing an S-shaped curve [38]. To make a binary decision, such as determining whether an IDH mutation is positive or negative, the model compares the estimated probability to a predefined threshold. If the probability is higher than the threshold, the model predicts that the instance belongs to the class. If the probability is lower, the model predicts that the instance does not belong to the class.

Logistic Regression estimates the parameters of the logistic function to best fit the data. The logistic function, is defined as [70]:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{2.8}$$

with β_1 representing the weight vector, x the input feature vector, and β_0 the bias term. Figure 2.13 represents the logistic function.

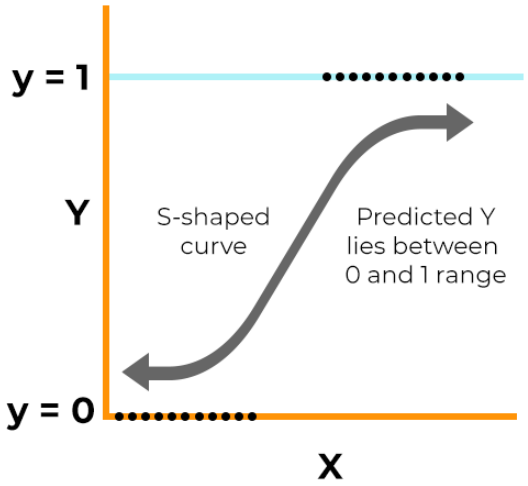


Figure 2.13: Logistic Regression. From [44]

The goal of logistic regression is to find the optimal weights β_1 and bias β_0 that maximize the likelihood of the observed data. This is typically done by minimizing the negative log-likelihood, also known as the binary cross-entropy loss [70]:

$$L(\beta_1, \beta_0) = -\frac{1}{N} \sum_{i=1}^{N} [y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \tag{2.9}$$

where N is the number of training samples, y_i is the true label, and $\hat{y}_i$ is the predicted probability given by the logistic function.

Logistic Regression offers several advantages and disadvantages. Table 2.9 summarizes these key points.

Table 2.9: Advantages and Disadvantages of Logistic Regression

Advantages	Disadvantages
Simple to implement and interpret; coefficients indicate feature importance	Assumes a linear relationship between features and log-odds of the target. It may not perform well if the true relationship is highly non-linear
Provides probability estimates for each class	Tuning hyperparameters can be challenging and time-consuming
Computationally efficient and can handle large datasets	Requires proper scaling of features to ensure convergence during optimization
Memory efficient	Can be sensitive to outliers

2.7.5 K-Nearest Neighbors

K-Nearest Neighbors is a simple, non-parametric algorithm used for classification and regression. It works by finding the k closest training examples in the feature space to a given test point and assigns the most common class among these neighbors to the test point [37].

KNN is then based on the assumption that similar points in the feature space tend to belong to the same class. The algorithm follows these steps [73]:

1. **Compute Distance:** Calculate the distance between the test point and all training points using a distance metric such as Euclidean distance.
2. **Find Neighbors:** Identify the k training points that are closest to the test point. Lower values of k can have high variance, but low bias, and larger values of k may lead to high bias and lower variance.
3. **Assign Class:** For classification tasks, the algorithm predicts the class label of the input data point based on the most frequent class among the k nearest neighbors.

Figure 2.14 illustrates a simple example of how the KNN algorithm works.

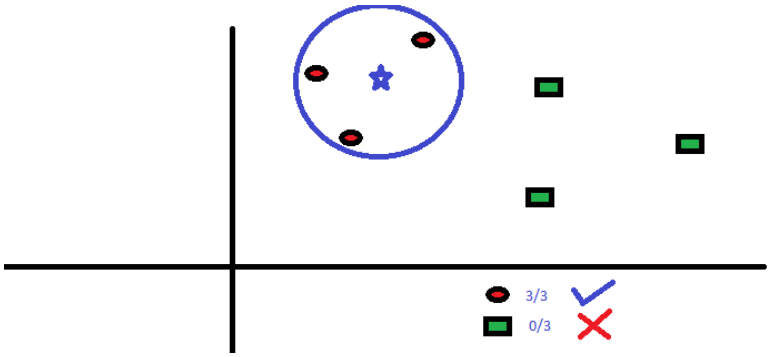


Figure 2.14: KNN simple example. From [73]

In the figure above, to determine the class of the blue star, a k value of 3 was chosen. A circle is drawn with the star at the center, large enough to enclose only three data points on the plane. The three closest points to the star are all red circles. Therefore, with a high level of confidence, it can be inferred that the star belongs to the class of the red circles [73].

KNN offers several advantages and disadvantages. Table 2.10 summarizes these key points.

Table 2.10: Advantages and Disadvantages of K-Nearest Neighbors

Advantages	**Disadvantages**
Straightforward to implement and understand	Performance degrades with many irrelevant or redundant features
Lazy learner, no training phase required; stores all training data and performs classification only when a query is received.	Storing all training data can require significant memory
Few hyperparameters: Only requires a k value and a distance metric	Performance depends on choice of k and distance metric; inappropriate choices can impact accuracy

2.7.6 Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of strong (naive) independence between the features [39]. In other words, a Naïve Bayes classifier assumes that the presence of a particular feature in a class is independent of the presence of any other feature, which is often not true in real-world scenarios. This simplification allows for the efficient computation of probabilities and makes the model particularly useful for large datasets and high-dimensional feature spaces.

Bayes' theorem is given by:

$$P(y|X) = \frac{P(X|y).P(y)}{P(X)} \tag{2.10}$$

where:

- $P(y|X)$ is the posterior probability of class y given the feature vector X.
- $P(X|y)$ is the likelihood of observing the feature vector X given class y.
- $P(y)$ is the prior probability of class y.
- $P(X)$ is the evidence, or the total probability of observing the feature vector X.

Since $P(X)$ is constant for all classes, it can be ignored in the classification process, leading to the simplified form:

$$P(y|X) \propto P(X|y).P(y) \tag{2.11}$$

The "naive" assumption implies that the features are conditionally independent given the class, allowing us to factorize the likelihood:

$$P(X|y) = \prod_{i=1}^{n} P(x_i|y) \tag{2.12}$$

where x_i is the i-th feature.

Types of Naïve Bayes Classifiers:

1. **Gaussian Naive Bayes:** Assumes that the continuous features follow a Gaussian (normal) distribution.
2. **Multinomial Naive Bayes:** Typically used for discrete data, such as word counts in text classification.
3. **Bernoulli Naive Bayes:** Used for binary/Boolean features, where each feature is assumed to follow a Bernoulli distribution.

Näive Bayes offers several advantages and disadvantages. Table 2.11 summarizes these key points.

Table 2.11: Advantages and Disadvantages of Näive Bayes

Advantages	Disadvantages
Easy to implement and understand	Assigns zero probability to unseen categorical features in training data; can be mitigated with Laplace smoothing.
When the assumption of independence holds, the classifier often outperforms other machine learning models and requires less training data	The assumption of feature independence is often unrealistic in real-world applications, potentially affecting the model's performance

2.8 Interactive Tools

In the field of radiogenomics, interactive tools play an important role in facilitating data analysis, visualization, and user interaction. These tools provide researchers and clinicians with intuitive interfaces to explore complex datasets, perform analyses, and derive meaningful insights. This section introduces examples of interactive tools and highlights the importance of interface design using frameworks like Qt Designer.

2.8.1 CaPTk (Cancer Imaging Phenomics Toolkit)

CaPTk, previously referenced in this chapter (Section 2.6.4), is a comprehensive software platform designed to support the analysis of cancer imaging data. It integrates advanced image processing techniques and machine learning algorithms, providing a user-friendly interface for both researchers and clinicians [16].

One of the key strengths of CaPTk is its advanced visualization tools, which facilitate user interaction with imaging data, allow for the viewing of segmentation results, and aid in the interpretation of model outputs. The intuitive graphical user interface (GUI) of CaPTk is accessible to users with varying levels of technical expertise, promoting its adoption in both clinical and research settings. Figure 2.15 illustrates the GUI of CaPTk, showcasing annotated ROIs and initialized tumor segmentation.

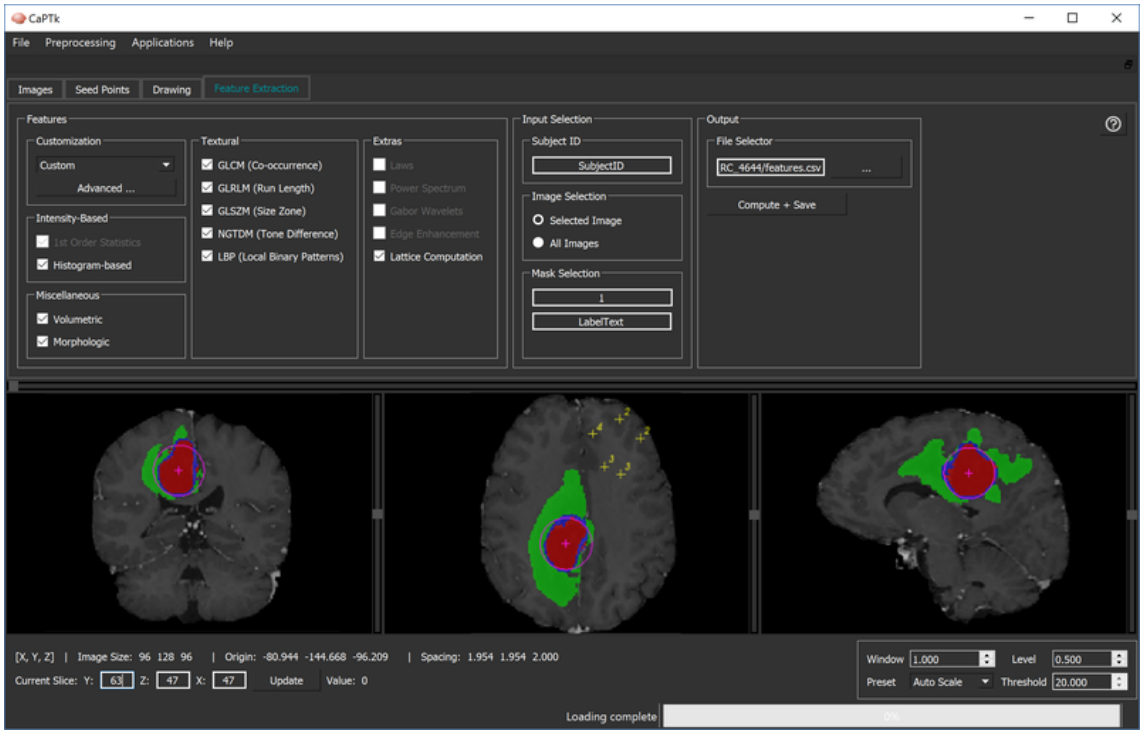


Figure 2.15: CaPTk Screenshot showing example annotated ROIs and initialized tumor. From [16].

2.8.2 Qt Designer

Qt Designer is a tool for designing and building GUIs for applications. It is part of the Qt framework, which is widely used for developing cross-platform software with rich GUIs [62]. Some notable aspects of Qt Designer include:

- **Drag-and-Drop Interface:** Qt Designer offers a drag-and-drop interface for designing GUIs, allowing users to create complex interfaces without writing extensive code. This feature accelerates the development process and enhances productivity.

- **Customizable Widgets:** The tool provides a wide range of customizable widgets, such as buttons, sliders, and text boxes, which can be tailored to meet the specific needs of an application.
- **Cross-Platform Compatibility:** Applications developed with Qt Designer can be deployed across multiple platforms, including Windows, macOS, and Linux, ensuring broad accessibility and usability.

2.8.3 Importance of Interactive Tools in Radiogenomics

Interactive GUIs significantly enhance the workflow in radiogenomics by providing:

- **Enhanced Usability:** Intuitive interfaces reduce the learning curve and enable users to interact with complex datasets and analysis tools more effectively.
- **Data Visualization:** Advanced visualization capabilities help in better interpretation of imaging and genomic data, leading to more informed decision-making.

2.9 Summary

This chapter is structured to cover several key areas, each contributing to a comprehensive understanding of gliomas and its molecular characterization, imaging modalities, and integration of radiomics and genomics.

- The chapter begins with an overview of brain anatomy, detailing the cerebrum, cerebellum, and brainstem.
- Next, it delves into the complexities and clinical significance of GBM, detailing its classification, characteristics, and the impact of molecular profiles on diagnosis and treatment strategies. The new WHO classification of gliomas is also discussed, emphasizing the role of IDH mutations in prognosis and treatment.
- The chapter then focuses on IDH mutations, explaining their normal and abnormal functions. It highlights the clinical importance of these mutations in gliomas and their association with better prognosis compared to IDH wild-type gliomas.
- An overview of various imaging techniques, including CT, MRI, and PET, is provided.
- Radiogenomics, which integrates radiomics and genomics, is explored in depth. The chapter covers conventional and deep radiomics, genomic analysis, and their integration to create powerful tools for precision medicine. Key steps such as data acquisition, preprocessing, feature extraction, feature selection, and data analysis using machine learning are detailed.

- Various techniques for selecting relevant features and reducing dimensionality are discussed, including filter methods, wrapper methods, and embedded methods. Methods such as PCA and LDA are explained to highlight their role in improving model performance and interpretability.
- The chapter describes different machine learning models used for classification tasks, including SVM, Random Forest, XGBoost, Logistic Regression, K-Nearest Neighbors, and Naive Bayes. Each model's advantages and disadvantages are outlined, emphasizing their suitability for different types of data and classification tasks.
- The importance of robust evaluation metrics is discussed, covering accuracy, precision, recall, F1 score, AUC-ROC, balanced accuracy, and specificity.
- The chapter concludes with an overview of interactive tools like CaPTk and Qt Designer, emphasizing their role in enhancing data analysis and visualization in radiogenomics.

Chapter 3

Literature Review

In recent years, the rapid evolution of Artificial Intelligence has captured increasing attention from both the public and researchers. The prevailing discourse on AI is characterized by a duality of perspectives. On one hand, there exists a concern that AI poses a potential threat, with fears of it replacing human roles in various industries, thereby giving rise to apprehension about its unchecked progression. On the other hand, there is a more optimistic view that perceives AI as a valuable tool designed to augment and enhance human capabilities rather than displace them.

Within the realm of healthcare, AI holds immense promise and utility. It serves as a powerful ally to healthcare professionals, aiding them in several tasks such as image analysis and decision support systems. The integration of AI in medical practices offers innovative solutions to complex challenges. By leveraging machine learning algorithms, AI can analyse medical images with unprecedented precision, assisting doctors in making more accurate diagnoses and treatment plans. Moreover, AI's role as a decision-support system in healthcare cannot be overstated. It can process vast amounts of medical data, identifying patterns and correlations that may elude the human eye.

This literature review delves into the intersection of AI, MRI analysis, and radiogenomics, with a focused lens on understanding IDH mutations in GBM. The purpose of this exploration is to critically examine the current state of AI-based methodologies, particularly in the context of radiogenomics, and their role in deciphering the complex genomic landscape of GBM. By unraveling the nuances of existing studies, we aim to identify gaps, trends, and methodological strengths.

3.1 State of the Art in GBM Radiogenomics

Building upon the discussion in the preceding chapter, Radiogenomics has gained relevance in cancer medicine, largely driven by the inherent limitations of molecular pathology and conventional radiology. These traditional methods are often hampered by temporal and spatial constraints of tissue sampling, making them less comprehensive. In contrast, advancements in ML algorithms applied to vast imaging databases, coupled with the growing recognition of the potential insights from imaging data, have significantly propelled the field forwards [68].

The advent of genomic sequencing and medical radiography has substantially increased the volume of patient data available to clinicians. This data influx is particularly beneficial in radiogenomics, where AI excels across a spectrum of applications. These include image acquisition and reconstruction, feature extraction, data analysis, and the development of predictive models for cancer analysis. These models are instrumental in prognostication, treatment planning, follow-up strategies, and more [68, 45].

For instance, a notable study in this field is "*Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from MRI*" by Le et al. [48]. This research aimed to distinguish distinct molecular subtypes of GBM based on genetic characteristics. Utilizing advanced radiomics techniques, the study focused on categorizing GBM into classical, mesenchymal, neural, and proneural subtypes. The findings highlighted critical molecular features that influence tumor behavior, achieving predictive accuracies of 70.9%, 73.3%, 88.4%, and 88.4% for the classical, mesenchymal, neural, and proneural subtypes, respectively. This study employed XGBoost and a two-level feature selection analysis (Spearman correlation and F-score), demonstrating this combination as a powerful approach for high-performance classification of transcriptome subtypes.

3.1.1 Deep Learning vs Machine Learning

Both machine learning and deep learning approaches have been extensively utilized to predict genetic information from radiomic features, each offering unique advantages and limitations. Although ML-based signatures are more frequently applied, as illustrated in Figure 3.1, the choice between ML and DL models depends significantly on the specific application and the size of the dataset.

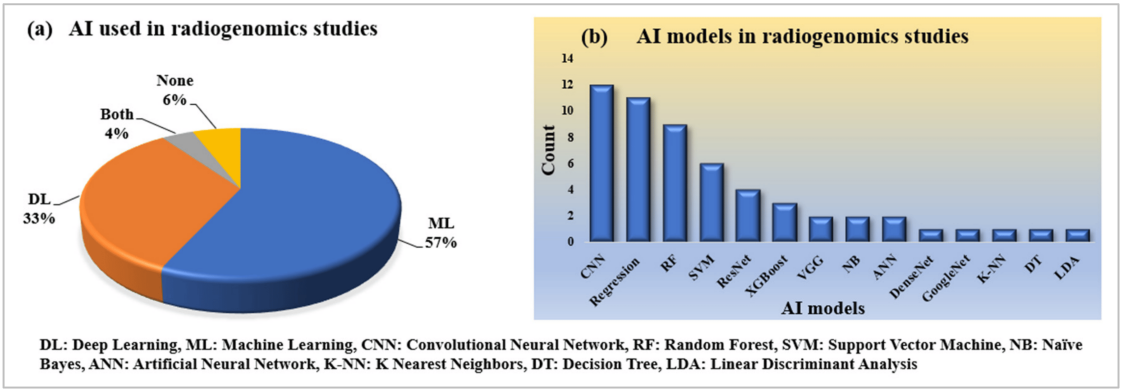


Figure 3.1: AI and its models used in radiogenomics studies (a) AI; (b) AI models. Year 2022. From [69]

ML-based imaging signatures rely on hand-crafted radiomic features derived from specific attributes such as shape, size, grade, tissue composition, texture, and histograms. These features are manually selected based on domain knowledge. In contrast, DL-based imaging signatures automatically generate deep radiomic features through sophisticated pattern recognition, predicting

genetic information without manual intervention. This automation reduces the need for manual feature extraction, providing a more streamlined and potentially more powerful approach than traditional ML models.

Researchers often prefer ML-based imaging signatures when aiming to minimize training time and when faced with limited data availability. ML models can be effective with smaller datasets and predefined features. On the other hand, DL models, which require significantly larger datasets and longer training periods, are more data-intensive but can yield more precise results due to their ability to learn complex patterns directly from raw images. The primary challenge in current research is the limited access to comprehensive radiogenomics datasets, which hampers the effective training and application of deep learning models.

In summary:

- ML models can be effective with smaller datasets and predefined features, while DL models generally require larger datasets to learn meaningful patterns directly from raw images.
- ML approaches rely on manual feature extraction, which can be both a strength and a limitation. It allows for interpretability but may miss complex patterns that DL models can uncover through automated feature learning.
- ML models are typically more interpretable, enabling clinicians to understand the factors driving predictions. Conversely, while DL models are powerful, they often function as "black boxes", which can hinder their adoption in clinical settings due to a lack of transparency.

3.1.2 Predicting IDH Mutation Status Using Deep Learning Approaches

While genetic biomarkers are traditionally determined through histopathological testing, the noninvasive prediction of biomarker status presents significant clinical advantages. Histopathological diagnosis typically necessitates large tissue specimens and can be prone to inaccuracies due to tumor heterogeneity. Furthermore, identifying these biomarkers pre-surgically can aid in surgical planning and determining the extent of resection required. Noninvasive prediction methods thus offer valuable information, enhancing surgical outcomes and overall patient management.

In the study titled "*Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics*" by Choi et al. [18] and in a similar study titled "*Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma*", by Li et al. [51], the researchers developed noninvasive methods to assess IDH mutation status in gliomas, specifically targeting GBM and low-grade gliomas, respectively.

Choi et al. employed a combination of deep learning techniques and radiomics to create a fully automated hybrid model, illustrated in Figure 3.2. This model integrates a Convolutional Neural Network (CNN) for analyzing 2-dimensional tumor signal intensity with radiomic features derived from the 3-dimensional shape and location of the tumor. The CNN component captures

detailed texture and intensity patterns from MRI scans, while the radiomics component provides complementary information about the tumor's morphological and spatial characteristics.

The study reported an accuracy ranging from 78.8% to 93.8% and areas under the receiver operating characteristic (ROC) curves between 0.86 and 0.96. The model's high accuracy and robust ROC values underscore its potential as a reliable, noninvasive tool for characterizing gliomas.

Similarly, Li et al. [51] proposed a DL approach for the prediction of the IDH1 mutation status in glioma patients through a private MRI imaging dataset. The used pipeline is illustrated in Figure 3.3.

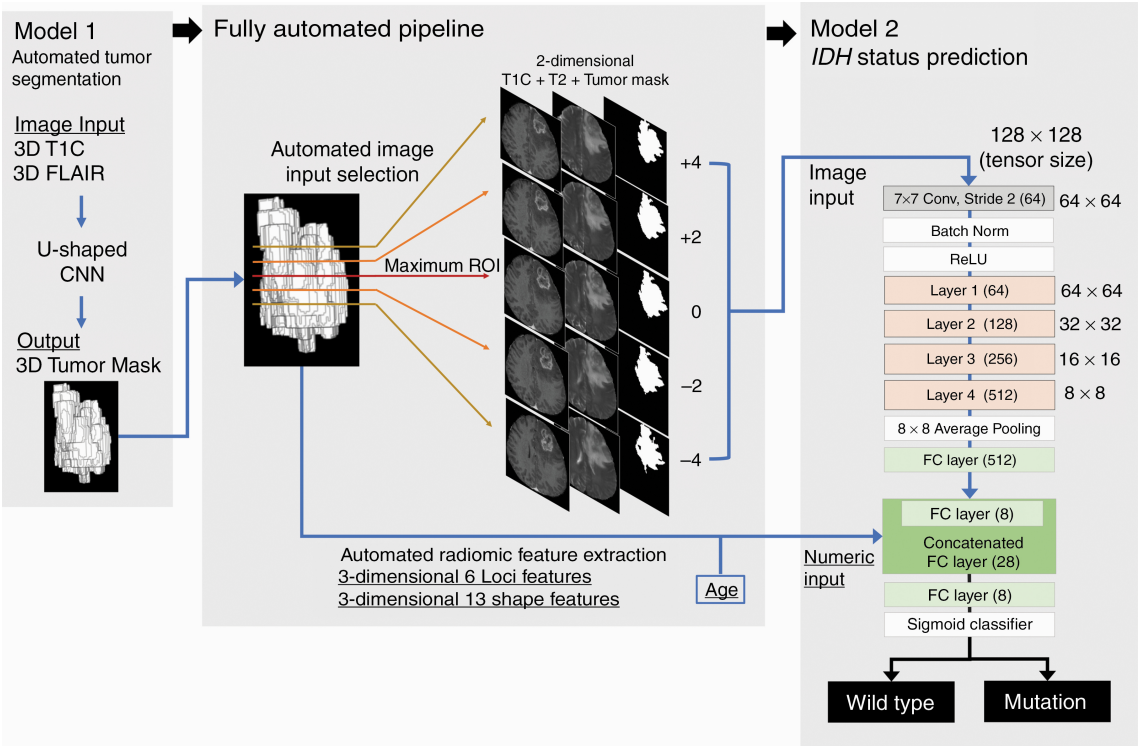


Figure 3.2: Fully automated hybrid model for IDH status prediction. In Model 2, layers 1–4 consisted of 3, 4, 6, and 3 residual blocks, with each block containing the 3 × 3 convolutions twice. From [18]

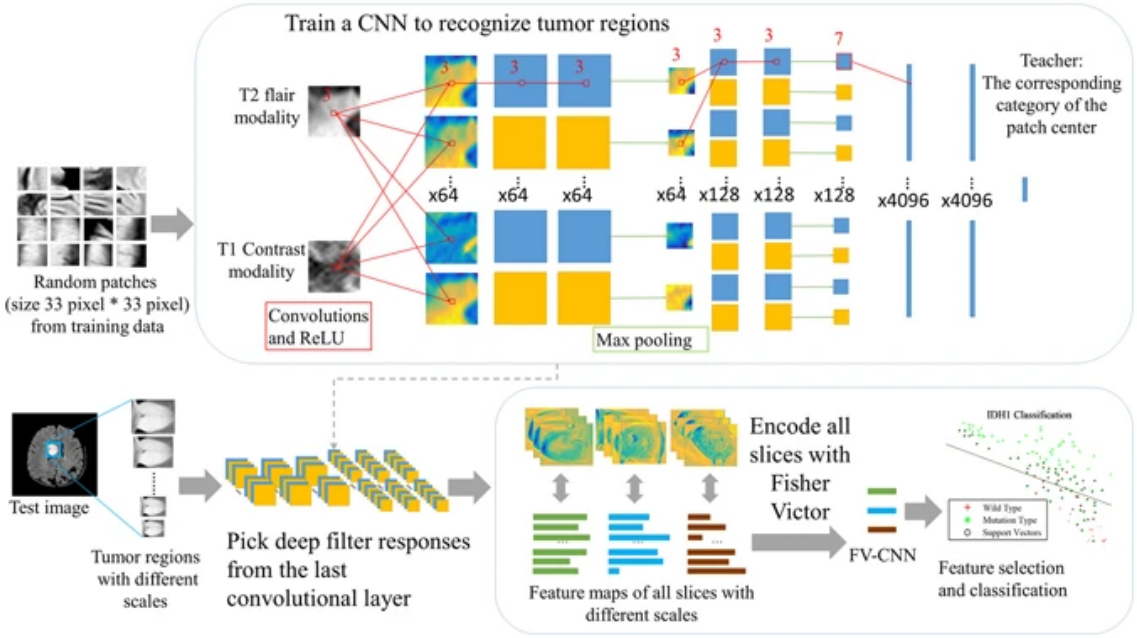


Figure 3.3: Architecture of the deep learning model consisting of two selection steps. First, the model identifies tumor regions in MR images using a state-of-the-art CNN structure. Second, deep filter responses from the last convolutional layer are extracted and encoded using Fisher vector encoding. The model's predictions are then evaluated using leave-one-out cross-validation with an SVM. From [51]

3.1.3 Predicting IDH Mutation Status Using Machine Learning Techniques

Recent advancements in ML algorithms and computation have created promising opportunities for exploring MRI radiomic features. These features are particularly suitable for ML analysis due to the large data size and multimodal nature of MRI. Consequently, ML methods have been increasingly investigated for predicting glioma genetic biomarkers from MRI radiomic features.

Research into the use of ML and MRI-radiomics for predicting the IDH1 genotype has primarily focused on SVM [8, 82] and RF [1, 52, 80] models. Some studies have utilized only conventional MRI sequences, such as T1-weighted and T2-weighted [52, 80, 82], while others have examined the added value of advanced MRI techniques [1, 8], such as MR diffusion or perfusion imaging.

For instance, the study "*Radiomics Strategy for Molecular Subtype Stratification of Lower-Grade Glioma: Detecting IDH and TP53 Mutations Based on Multimodal MRI*" [82] by Zang et al., and "*Multiregional Radiomics Profiling from Multiparametric MRI: Identifying an Imaging Predictor of IDH1 Mutation Status in Glioblastoma*" [52] by Zhi-Cheng Li et al., highlight the effectiveness of these approaches. Zang et al. utilized conventional MRI sequences and achieved an AUC ranging from 75.6% to 83% with an SVM-RFE model. On the other hand, Zhi-Cheng Li et al. focused on glioblastoma patients and achieved an Area Under the Curve (AUC) between 86% and 96% using a Random Forest model.

Expanding on the previously mentioned studies, the study titled "*MRI Radiomic Features to*

Predict IDH1 Mutation Status in Gliomas: A Machine Learning Approach using Gradient Tree Boosting" [65], Sakai et al. selected XGBoost as their classifier. They trained models using FLAIR and DWI radiomic features, achieving their best AUC value of 97% with the DWI MRI sequence. This study further highlights the potential of advanced machine learning techniques in accurately predicting IDH1 mutation status from MRI radiomic features.

3.2 Summary

This chapter presents relevant studies on both DL and ML approaches in MRI analysis and radiogenomics to predict IDH1 mutations. The studies reviewed employed various robust methodologies to investigate the role of AI in understanding and predicting IDH mutations. Table 3.1 summarizes the works that have been reviewed.

Table 3.1: Summary of studies on predicting IDH1 mutation status in gliomas using MRI radiomics and machine learning approaches.

Study	Technique	MRI Sequences Used	Dataset	Best AUC
Choi et al. [18] (2020)	Deep Radiomics (CNN)	T1GD, T2, FLAIR	856 Low/High-Grade Glioma patients	86% - 96%
Li et al. [51] (2017)	Deep Radiomics (CNN)	T1GD, FLAIR	151 Low-Grade Glioma patients	95%
Zang et al. [82] (2018)	Support Vector Machine	T1, T1GD, T2, FLAIR	199 Low-Grade Glioma patients	83%
Zhi-Cheng Li et al. [52] (2018)	Random Forest	T1, T1GD, T2, FLAIR	225 GBM patients	86% - 96%
Sakai et al. [65] (2020)	eXtreme Gradient tree Boosting	FLAIR, DWI	100 Low/High-Grade Glioma patients	97%

Note: T1 = T1-weighted, T1GD = postcontrast T1-weighted, T2 = T2-weighted, FLAIR = Fluid-Attenuated Inversion Recovery, DWI = Diffusion-Weighted Imaging

The reviewed studies underscore the potential of AI in bridging the gap between radiological imaging and genomic information, offering a more comprehensive understanding of the underlying molecular landscape of diseases. However, certain limitations are noteworthy. Many studies rely on retrospective analyses and relatively small sample sizes, due to the limited size of the publicly available datasets, potentially limiting the generalizability of findings. Additionally, the lack of standardized methodologies and diverse datasets poses challenges in comparing results across studies. The dynamic nature of AI technologies also suggests that the field is continuously evolving, requiring ongoing validation and refinement of models. Furthermore, the interpretability of AI models remains a challenge, i.e. it is important to develop models that provide transparent insights into the features influencing predictions, fostering trust and acceptance among healthcare practitioners.

As we position our work within the state-of-the-art, it becomes evident that there is a need for further research that not only validates existing methodologies but also explores new approaches to enhance the accuracy and clinical utility of AI models in predicting IDH mutations. Our research aims to contribute to this evolving landscape by evaluating various machine learning models for predicting IDH mutation status in gliomas, as well as investigating different MRI sequences and feature selection methods, thereby adding to the collective knowledge in the field of AI-based medical image analysis and radiogenomics.

Chapter 4

Data Description

This chapter provides a detailed description of the datasets employed in our research, which are critical for the analysis and conclusions drawn in subsequent sections. The study leverages two prominent datasets, the University of California San Francisco Preoperative Diffuse Glioma MRI (Section 4.1) and the University of Pennsylvania Health System Glioblastoma (Section 4.2). These datasets contain comprehensive magnetic resonance imaging and clinical information relevant to gliomas, including data on IDH mutation status.

The public accessibility of glioma MRI datasets has been instrumental in the development of advanced AI techniques, including automated tumor segmentation, radiogenomics, and MRI-based survival prediction. These datasets have facilitated significant progress in the field by providing a rich source of data for various machine learning and imaging studies. However, the current publicly available glioma MRI datasets are generally confined to only four MRI contrasts (T2, FLAIR, and T1 pre- and post-contrast). Furthermore, there is considerable variation in imaging protocols, particularly regarding magnetic field strength and acquisition parameters, which poses challenges for standardization and reproducibility in research.

4.1 UCSF-PDGM Dataset

The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM) dataset [13] is accessible through The Cancer Imaging Archive (TCIA). Developed by the University of California San Francisco, this dataset contains preoperative MRI scans from patients with histopathologically confirmed diffuse gliomas. The data collection adhered to strict guidelines and regulations, receiving approval from the University of California San Francisco institutional review board, which exempted the requirement for patient consent.

4.1.1 Patient Population

The UCSF-PDGM dataset comprises imaging data for 501 cases, representing 495 unique adult patients, with six cases being follow-up images. These patients were diagnosed with grade 2-4 diffuse gliomas and underwent imaging between 2015 and 2021.

The patient cohort is comprised of 56 (≈11%) grade 2, 42 (≈9%) grade 3, and 403 (≈80%) grade 4 tumors. There is a male predominance across all tumor grades, with males representing 56% for grade 2 and 60% of both grade 3 and grade 4. Figure 4.1 provides a detailed summary of the demographic characteristics of patients included in the UCSF-PDGM dataset, and Figure 4.2 and 4.3 the clinical characteristics.

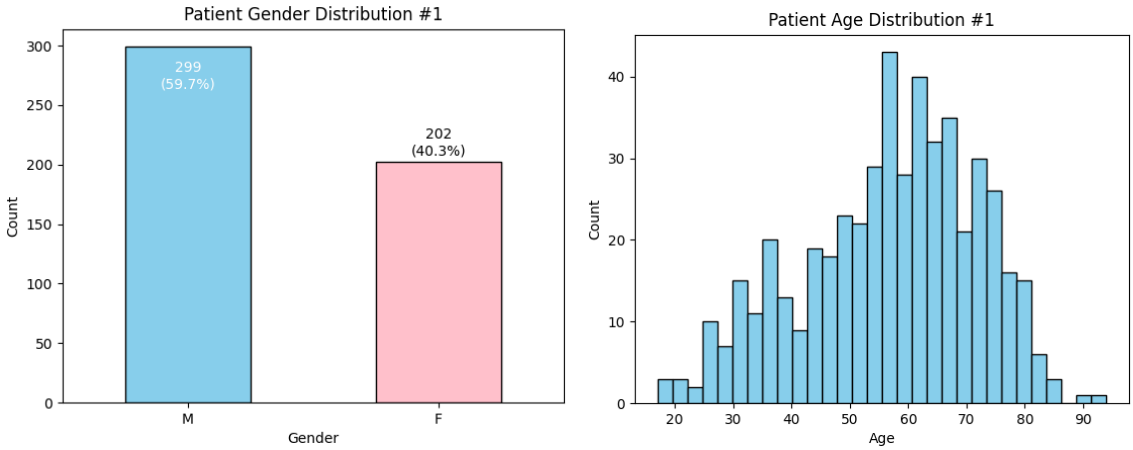


Figure 4.1: Gender and Age distributions in UCSF-PDGM. Left: Gender. Right: Age.

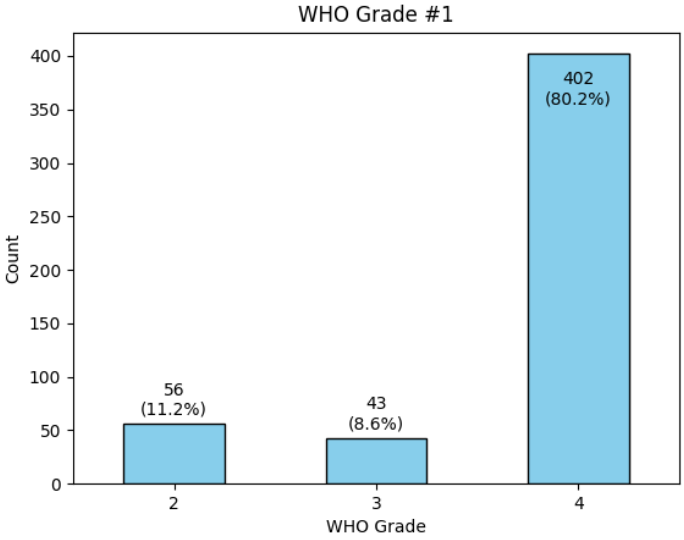


Figure 4.2: WHO Grade distribution in UCSF-PDGM

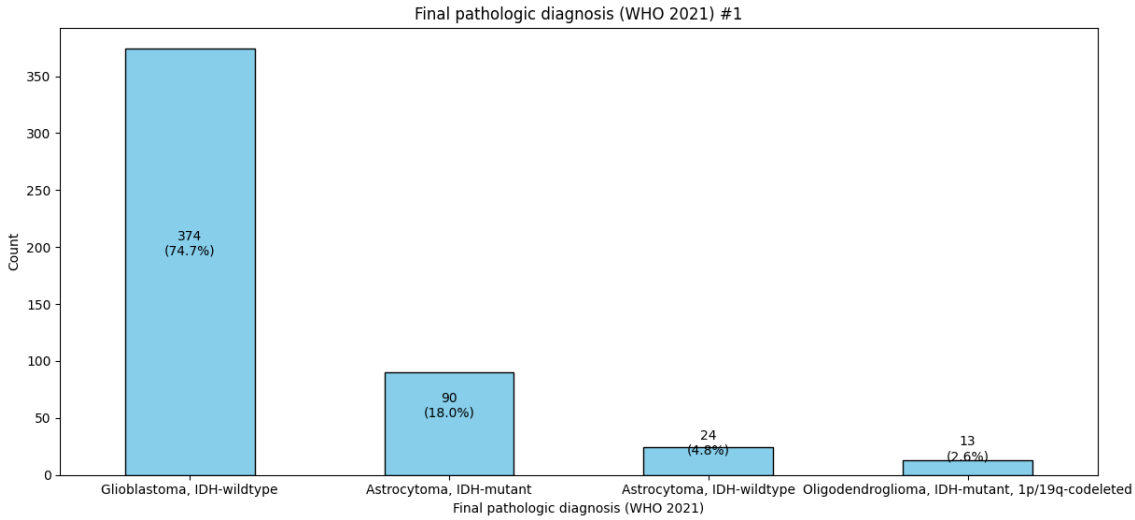


Figure 4.3: Final Pathologic Diagnosis (WHO 2021) distribution in UCSF-PDGM

4.1.2 Clinical Data

All tumors in the UCSF-PDGM dataset were tested for IDH mutations using genetic sequencing of tissue samples obtained during biopsy or resection. The results revealed that IDH mutations were present in a majority of grade 2 (83%) and grade 3 (67%) tumors, but only in a small minority of grade 4 tumors (8%).

Figure 4.4 illustrates the distribution of IDH mutations within the patient cohort, which will serve as the target label for training the designed frameworks. Additionally, the mutations are categorized as follows:

- IDH1 p.R132H (10.8% - 54 cases)
- IDH1 p.R132C (1.6% - 8 cases)
- IDH1 p.R132G (1% - 5 cases)
- IDH1 p.R132S (0.4% - 2 cases)
- IDH1 p.Arg132His (0.2% - 1 case)
- IDH2 p.Arg172Trp (0.2% - 1 case)
- IDH2 p.R172K (0.2% - 1 case)
- Mutated (Not otherwise specified) (6.2% - 31 cases)

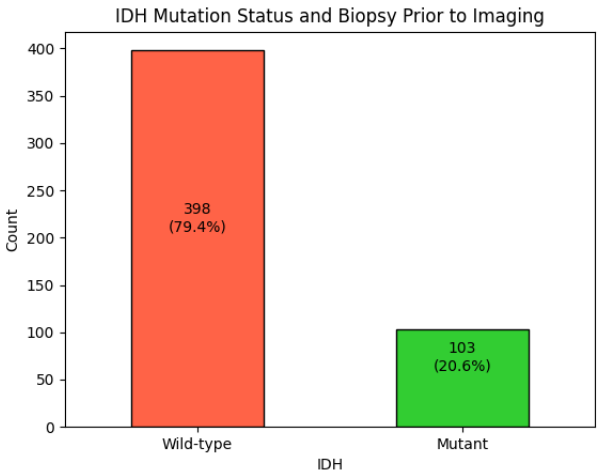


Figure 4.4: Distribution of IDH Mutations in the UCSF-PDGM Dataset.

Furthermore, this dataset includes the following additional information:

- **MGMT Status:** Classified as negative, positive, or indeterminate.
- **Presence of Codeletion of 1p and 19q genes:** Categorized as intact, co-deletion, relative co-deletion, or unknown.
- **Survival Status:** The survival status of the patient at the last clinical follow-up.
- **Overall survival:** Measured in days from initial diagnosis to the last clinical follow-up.
- **Extent of resection:** Classified as biopsy, subtotal resection, or gross total resection.
- **Biopsy prior to imaging:** Indicates whether a biopsy was performed before imaging (yes or no).

4.1.3 Image Acquisition

All preoperative MRI scans were conducted using a 3.0 tesla scanner (Discovery 750, GE Healthcare, Waukesha, Wisconsin, USA) and a dedicated 8-channel head coil (Invivo, Gainesville, Florida, USA). The imaging protocol included the following sequences (Figure 4.5):

- T2-weighted
- FLAIR-weighted
- Susceptibility-weighted
- Diffusion-weighted
- Pre- and post-contrast T1-weighted images
- 3D arterial spin labeling (ASL) perfusion images

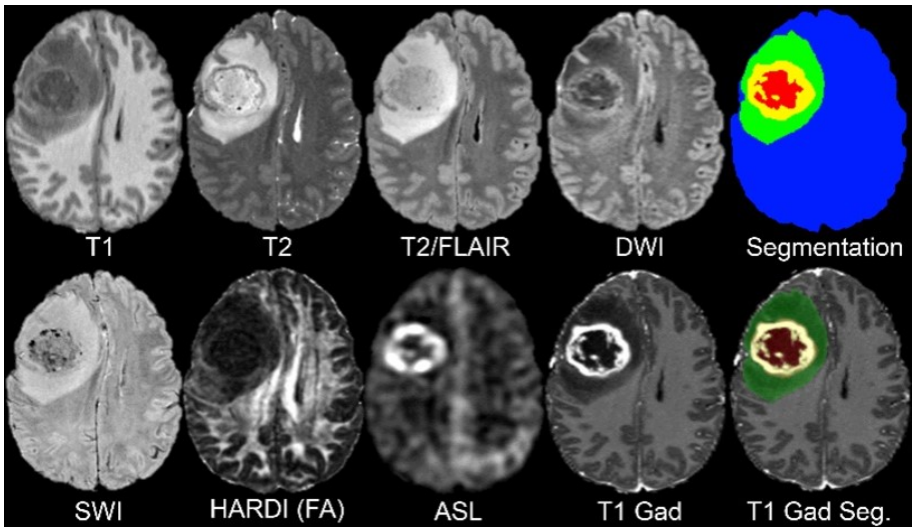


Figure 4.5: MRI Sequences included in the UCSF-PDGM dataset. From [13]

- 2D 55-direction high angular resolution diffusion imaging (HARDI)

Throughout the study period, two gadolinium-based contrast agents were utilized: gadobutrol (Gadovist, Bayer, LOC) at a dose of 0.1 mL/kg and gadoterate (Dotarem, Guerbet, Aulnay-sous-Bois, France) at a dose of 0.2 mL/kg. These agents were used to enhance the visibility of internal structures in MRI scans, with gadobutrol administered at 0.1 mL per kilogram of body weight and gadoterate at 0.2 mL per kilogram.

Each MRI contrast was aligned and resampled to the 3D space defined by the T2/FLAIR image, achieving a uniform isotropic resolution of 1 mm using automated non-linear registration techniques. This process ensures that all images are accurately aligned, enabling precise comparisons and analyses. Following this, the resampled and co-registered data were processed to remove the skull using a publicly available deep-learning algorithm [14]. This "skull-stripping" technique isolates the brain from other cranial structures, providing clean, standardized images for further analysis.

Additionally, the MRI volumes are provided in NIfTI and DICOM format, ensuring compatibility with a wide range of imaging software and facilitating easy integration into research workflows.

4.1.4 Tumor Segmentation

Multicompartment tumor segmentation of the study data was performed as part of the 2021 BraTS challenge. Initially, the image data underwent automated segmentation using an ensemble model that incorporated algorithms from previous BraTS challenge winners. This automated segmentation was then manually reviewed and corrected by trained radiologists. The final segmentations were approved by two expert reviewers to ensure accuracy [13].

4.2 UPENN-GBM Dataset

The Multi-parametric Magnetic Resonance Imaging (mpMRI) scans for de novo Glioblastoma patients from the University of Pennsylvania Health System (UPENN-GBM) dataset [4] are part of TCIA. This dataset, developed at the University of Pennsylvania Health System, includes extensive mpMRI scans for patients diagnosed with de novo glioblastoma, providing a valuable resource for research and clinical studies.

4.2.1 Patient Population

The UPENN-GBM dataset includes imaging and clinical data from 630 adult patients diagnosed with glioblastoma (grade 4 glioma), making it a comprehensive resource for research. Additionally, follow-up images bring the total to 671 cases. This dataset covers a diverse range of patient demographics and clinical characteristics, offering valuable insights into glioblastoma. Figure 4.6 illustrates the distribution of gender and age among the patients in this dataset.

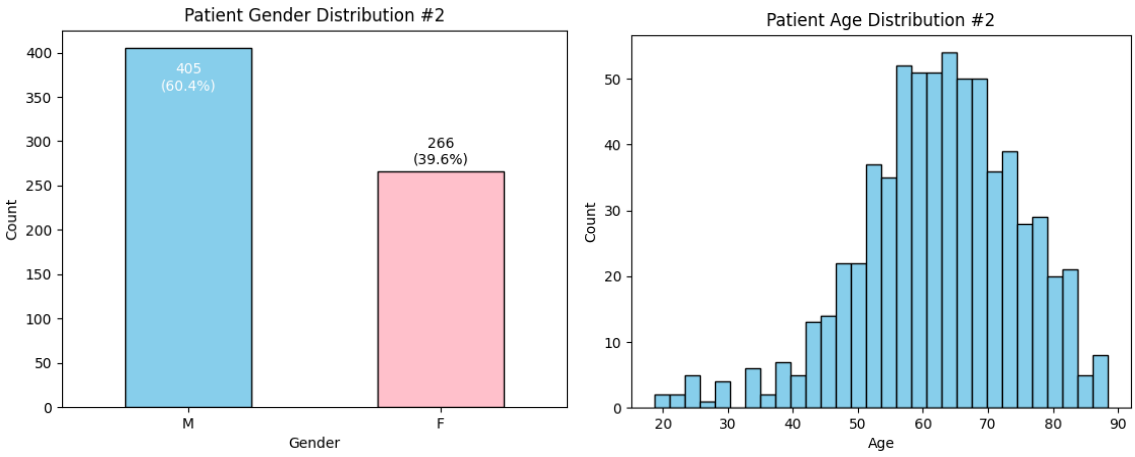


Figure 4.6: Gender and age distributions in UPENN-GBM. Left: Gender. Right: Age.

4.2.2 Clinical Data

In the UPENN-GBM dataset, tumors were tested for IDH1 mutations using genetic sequencing of tissue samples obtained during biopsy or resection. The analysis revealed that IDH1 mutations were present in 2.8% of the cases. However, IDH1 status was not available for 15.8% of the cases. Figure 4.7 provides a summary of the distribution of IDH1 mutations across the patient cohort, illustrating the proportion of patients with wildtype IDH, mutant IDH, and those with unavailable IDH status.

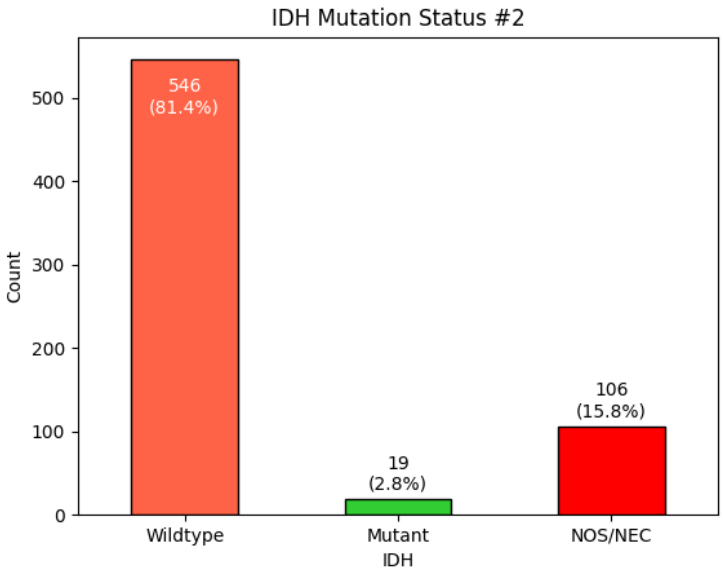


Figure 4.7: Distribution of IDH1 Mutations in the UCSF-PDGM Dataset.

Furthermore, this dataset includes the following additional information:

- **MGMT Status:** Classified as Unmethylated, Methylated, or indeterminate.
- **Survival Status:** The survival status of the patient at the last clinical follow-up.
- **Overall survival:** Measured in days from initial diagnosis to the last clinical follow-up.

4.2.3 Image Acquisition

The UPENN-GBM dataset features multi-parametric MRI scans that have undergone skull-stripping and co-registration processes. These scans offer detailed views of the brain, enhancing the identification and analysis of tumor regions. The mpMRI volumes are provided in NIfTI format, ensuring compatibility with various imaging software and facilitating seamless integration into research workflows.

The imaging protocol employed in this dataset includes the following sequences, each providing unique insights into the brain's structure and pathology (Figure 4.8):

- Pre-contrast T1-weighted images
- Post-contrast T1-weighted images
- T2-weighted images
- FLAIR-weighted images
- Diffusion-Tensor Imaging (DTI) and Dynamic Susceptibility Contrast (DSC) derivatives

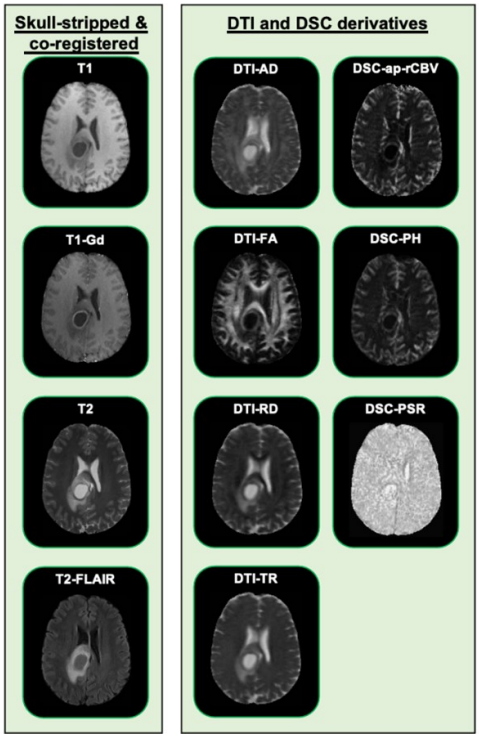


Figure 4.8: Examples of the different sequences available in UPENN-GBM. From [4]

4.2.4 Tumor Segmentation

The dataset includes segmentation labels for multiple histologically distinct tumor sub-regions as well as for the entire brain, both computer-aided and manually corrected. Initially, the segmentation labels were generated using an automated computational method. To ensure precision, these labels were then manually corrected and validated by expert board-certified neuroradiologists [4]. This rigorous process ensures the segmentation labels are of high quality, which is crucial for accurate and reliable analysis.

4.3 Summary

This chapter provides a detailed description of the two datasets utilized for analysis. The UCSF-PDGM and UPENN-GBM datasets offer essential information for building and validating IDH mutation prediction models. The characterization of the data included the origin, patient population, types of imaging data, format, and image pre-processing, as well as tumor segmentation. This deep understanding of the data is crucial for the successful implementation of the machine learning methods described in the following chapters.

Chapter 5

Methodology

The current chapter provides a detailed description of the proposed methodology to perform IDH mutation status prediction. It begins with an overview of the entire methodology. Subsequent sections delve into each step of the process, including dataset collection (Section 5.1), preprocessing (Section 5.2), feature extraction (Section 5.3), feature selection and dimensionality reduction (Section 5.4), the deployment and configuration of learning models (Section 5.5), the training and evaluation phase (Section 5.6), and finally, the development of an interactive tool (Section 5.7).

Method Overview

As previously mentioned, the proposed method aims to develop a classification model for IDH mutation status assessment using MRI data combined with labeled clinical data. The methodology is structured into several key stages: dataset collection, preprocessing, feature extraction, feature selection, and classification using machine learning algorithms. Figure 5.1 illustrates the methodology pipeline.

The following sections provide a detailed description of each stage, highlighting the specific techniques and algorithms employed to ensure accurate and reliable prediction of IDH mutation status.

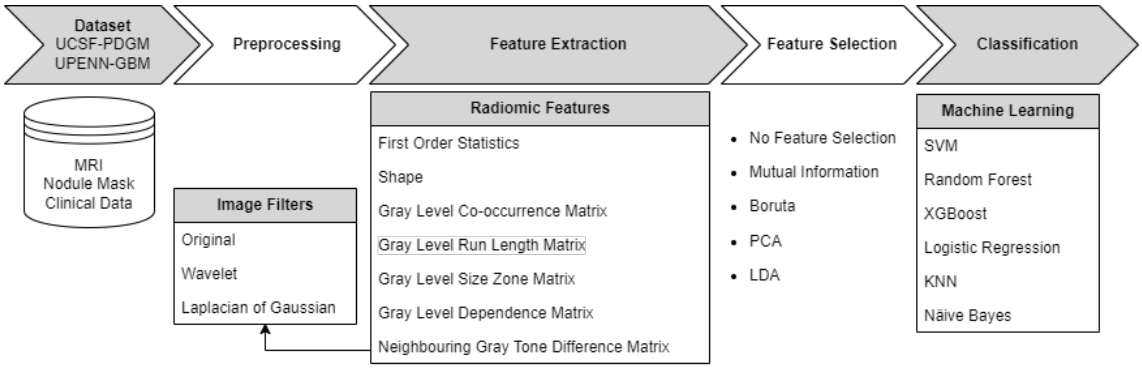


Figure 5.1: Schematic representation of the classification methodology for predicting IDH mutation status. The procedure starts with the input of MRI scans, brain segmentation masks, and clinical data into the software. Next, preprocessing is performed. The original images are enhanced using Wavelet and Laplacian of Gaussian (LoG) filters, followed by the extraction of radiomic features from the regions of interest via the PyRadiomics package. This feature extraction is applied to all images from both datasets. Subsequently, various combinations of six different classification models, alongside several feature selection techniques and a baseline model without feature selection, are evaluated to identify the most effective approach for accurately predicting IDH mutation status.

5.1 Dataset Collection

This study leverages two datasets, the UCSF-PDGM and UPENN-GBM, as detailed in Section 4.1, and a third one resulting from the combination of the two.

Initially, all available images in NIfTI format from both datasets were loaded. To facilitate the exploration and understanding of these datasets, an interactive visualization tool was developed. This tool, described in Section 5.7, is essential for examining each slice of the 3D MRI images. It allows for detailed inspection of specific slices and the application of tumor masks, as illustrated in Figure 5.2.

After analyzing the various sequences available in the datasets, four key MRI sequences were selected for further analysis. These sequences were chosen based on their prevalence in the literature and their effectiveness in highlighting tumor regions visually:

1. T1-Weighted
2. T1GD
3. T2-Weighted
4. FLAIR

These sequences were selected to ensure the inclusion of the most informative and diagnostically relevant images for subsequent feature extraction and model training.

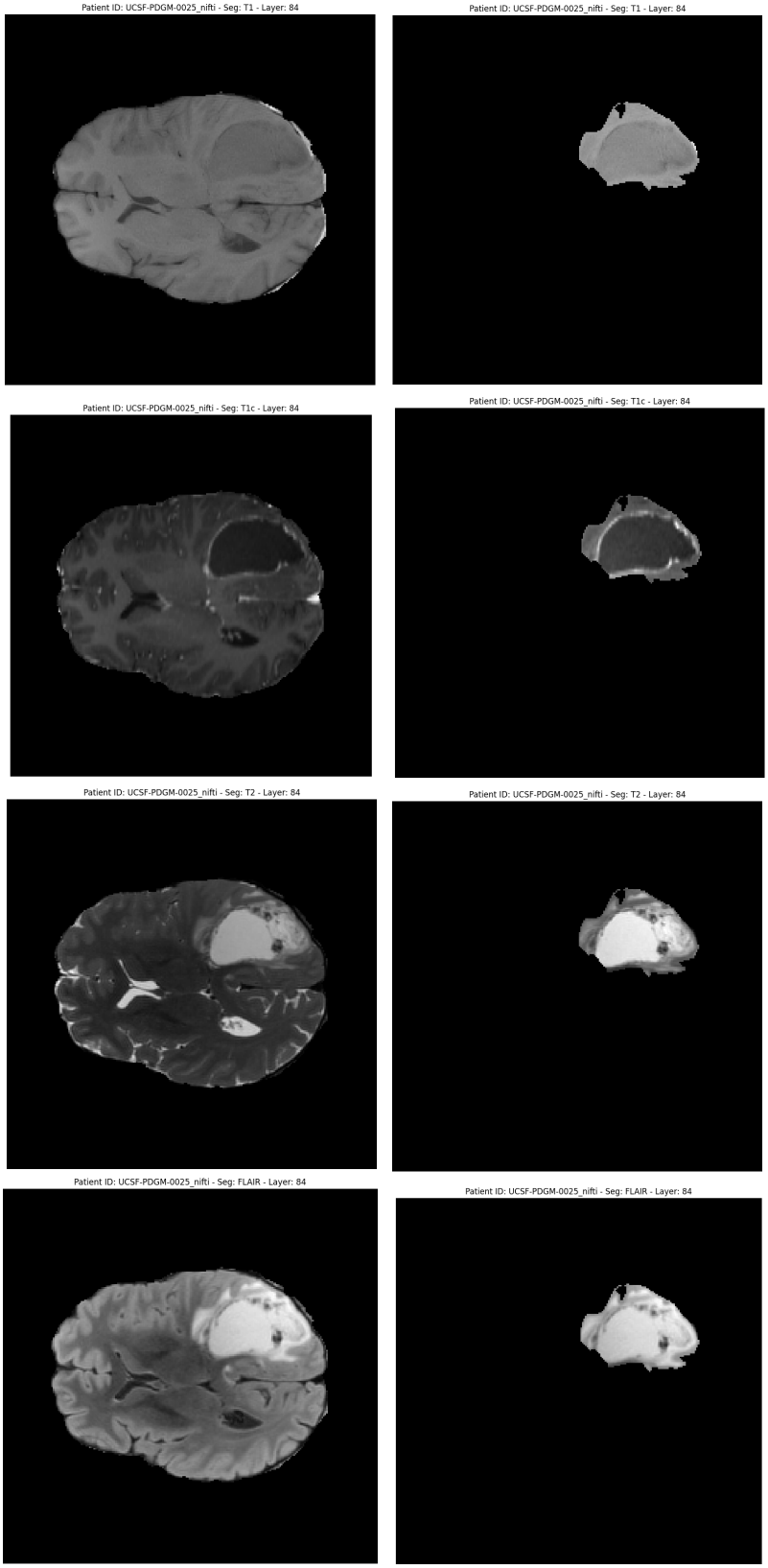


Figure 5.2: T1, T1GD, T2, and FLAIR sequences, respectively, Slice 84 of Patient 0025's Brain, from UCSF-PDGM Dataset. **Left Column:** Full Brain MRI showing the entire brain structure. **Right Column:** ROI highlighting the area with abnormal vascularity.

Patient Inclusion and Exclusion Criteria

Patients were included in the study if they met the following criteria:

1. Diagnosed with grade 2, 3, or 4 glioma with known IDH status confirmed by surgical pathology.
2. Availability of preoperative MRI scans, including T1, T1GD, T2, and FLAIR sequences.

Patients were excluded from the study if they met any of the following conditions:

1. Absence of available IDH status.
2. Insufficient quality of MRI scans.
3. Lack of available tumor segmentation.

The selection process for the UCSF-PDGM dataset identified 493 patient cases, comprising 100 cases with mutated IDH and 393 cases with IDH wildtype. For the UPENN-GBM dataset, the selection process resulted in a final cohort of 514 patient cases. Within this cohort, only 16 cases exhibited mutated IDH. The selection criteria were stringent to ensure the integrity of the MRI data for precise and reliable analysis.

Figure 5.3 illustrates the inclusion and exclusion criteria for both datasets. The left side details the process for the UCSF-PDGM dataset, while the right side outlines the process for the UPENN-GBM dataset.

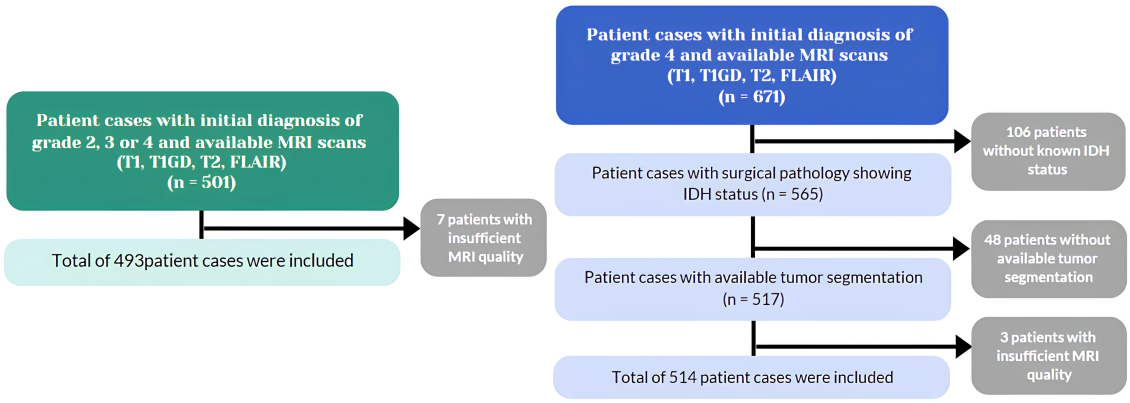


Figure 5.3: Flowchart of patient inclusion and exclusion of UCSF-PDGM dataset on the left and UPENN-GBM dataset on the right.

5.2 Pre-processing

Following the pipeline, Volumes of Interest (VOIs) were extracted by stacking ROIs delineated slice by slice from the selected sequences using the tumor masks provided in both datasets. This step resulted in eight subdatasets (four for each Dataset), allowing the analysis to concentrate exclusively on the relevant tumor areas, thereby enhancing the performance of subsequent processing stages.

To ensure the quality and consistency of the data, a preprocessing phase was implemented:

1. **Normalization:** The images were normalized to standardize the data across different patients and imaging modalities. Normalization is critical for facilitating reliable comparisons and analyses by adjusting the pixel intensity values to a common scale. This reduces variability and ensures that the features extracted from the images are consistent and comparable.

2. **Resampling:** All voxels were isotropically resampled to a resolution of $2 \times 2 \times 2\ mm^3$. During normalization, gray values below the mean were adjusted to avoid negative values. By shifting the intensity values by 300 (equivalent to 3 standard deviations multiplied by 100), the majority of voxels became positive, ensuring that only outliers (more than 3 standard deviations below the mean) remained negative.

The original image dimensions of 240x240 pixels were preserved throughout the preprocessing steps, ensuring that the spatial integrity of the tumor regions was maintained. This resolution was chosen to balance the need for detailed imaging with computational efficiency.

5.3 Feature Extraction

After preprocessing, the next step involved the extraction of radiomics features from the VOIs. For this purpose, a popular radiomics computational toolbox, PyRadiomics [77], was employed. PyRadiomics is an open-source Python package that provides a robust platform for extracting a large number of engineered features from medical imaging. This toolbox includes built-in wavelet and Laplacian of Gaussian (LoG) filters for image processing. A total of 1223 radiomic features were computed, including first-order statistics, shape, and texture classes:

1. **First Order Statistics** (18 features)

   First order statistics are basic statistical measures that describe the intensity values of voxels within an image without considering spatial relationships. These statistics provide a summary of the voxel value distribution and are essential for understanding the overall intensity characteristics of an image.

2. **Shape** (14 features)

3. **Texture classes:**

   (a) Gray Level Co-occurrence Matrix (24 features)

   (b) Gray Level Run Length Matrix (16 features)

   (c) Gray Level Size Zone Matrix (16 features)

   (d) Gray Level Dependence Matrix (14 features)

   (e) Neighbouring Gray Tone Difference Matrix (5 features)

Moreover, radiomic features were extracted not only from the original images (107 features) but also from images processed with wavelet (744 features) and Laplacian of Gaussian (LoG) filters (372 features).

- **Wavelet Filtering:** This technique decomposes the image into different frequency components. Each level of decomposition results in 8 sub-bands, corresponding to different combinations of high and low pass filters applied in three dimensions. This allows for multi-resolution analysis, capturing both coarse and fine details in the image [21].

- **Laplacian of Gaussian:** LoG is used for edge detection and texture analysis. It highlights regions of rapid intensity change, effectively identifying edges. The sigma parameter controls the filter's sensitivity to texture changes: low sigma values focus on fine details, while high sigma values highlight broader, coarser structures [21]. This study applied LoG filters with sigma values of 2.0 mm, 3.0 mm, 4.0 mm, and 5.0 mm to capture a range of texture details.

This thorough extraction of radiomic features from multiple filtered images ensures a rich and diverse dataset, enabling the development of robust predictive models for IDH mutation status.

5.4 Feature Selection / Dimensionality Reduction

Following the extraction of radiomic features, four techniques were employed to decrease the number of features used while training the model, by selecting the most relevant features and excluding redundant features or by transforming the feature space into a more compact one. This process helped to decrease the computational cost of training the models and, theoretically, enhanced model performance, leading to better results.

The techniques used involved one filter method, one wrapper method and two dimensionality reduction methods, previously detailed in Chapter 2, Section 2.5.7:

1. k-Best Features through Mutual Information (Filter Method)

2. Boruta (Wrapper Method)

3. Principal Components Analysis (PCA) (Dimensionality Reduction)

4. Linear Discriminant Analysis (LDA) (Dimensionality Reduction)

Additionally, a baseline model without any feature selection technique was also implemented to compare the performance of the predictive models with and without this process. This comparison allows the evaluation of the effectiveness of the feature selection process and its impact on model performance.

Unlike methods such as Boruta and LDA, which operate without the need for manual parameter specification, techniques like Mutual Information, and PCA require the number of features to be specified. This parameter, denoted as k, can either be set manually or determined automatically during the process of searching for optimal hyperparameters, as will be discussed in later sections. Boruta, being a wrapper method, relies on classifiers with built-in feature importance measures. In this study, a Random Forest classifier was employed to calculate and determine the optimal k features.

5.5 Classification Models and Hyperparameter Tuning

To predict the IDH status using the provided datasets, this study employs a variety of machine learning models, including SVM, Random Forest, XGBoost, Logistic Regression, KNN, and Naive Bayes. Several hyperparameters were trained for each algorithm. Each algorithm's performance is highly dependent on its hyperparameters, which need to be fine-tuned to achieve optimal results. This section outlines the hyperparameter tuning process and specifies the range of each hyperparameter for the different learning models used in this study.

The following subsections detail the hyperparameters considered for each algorithm and the strategies employed to fine-tune them.

5.5.1 Hyperparameters Search

To ensure optimal performance of the machine learning models, a thorough hyperparameter search was conducted using Bayesian optimization. This technique allows for efficient exploration of the hyperparameter space and improves model performance by identifying the best combination of hyperparameters.

The hyperparameter search was implemented using the 'BayesSearchCV' method from the scikit-optimize library, combined with a stratified 10-fold cross-validation strategy. This approach ensured that the class balance was maintained across all folds, providing a reliable assessment of model performance. The primary evaluation metric used to optimize the models was the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which offers a robust measure of model effectiveness.

Bayesian optimization was selected for its efficiency in handling complex hyperparameter spaces and its ability to converge more rapidly compared to traditional grid search or random search methods. This approach leverages probabilistic models to guide the search process, focusing on regions of the hyperparameter space that are more likely to yield better performance.

Specifically, it utilizes Gaussian Process regression to model the objective function and predict promising regions to explore [59]. This targeted approach allows for efficient optimization over a broader range of values, ultimately enhancing the model's performance.

In addition to fine-tuning the learning models, Bayesian optimization was employed to determine the optimal number of features to retain / transform for feature selection techniques such as Mutual Information, and PCA. This ensured that the most relevant features were selected, enhancing the overall predictive power of the models.

5.5.2 Support Vector Machine

- **C (Regularization parameter):** Determines the trade-off between achieving a low error on the training data and minimizing the margin. A range of values from 0.001 to 100 was tested.
- **Kernel:** Different kernel types (linear, polynomial, radial basis function (RBF), and sigmoid) were evaluated to identify the best fit for the data.
- **Gamma:** For non-linear kernels like RBF, the gamma parameter was set to 'scale' or 'auto'.
- **Class weight:** Weights associated with classes to handle class imbalance. The 'balanced' option uses the values of the labels to automatically adjust weights inversely proportional to class frequencies in the input data, while 'None' means no weights are applied.

Table 5.1 summarizes the range of hyperparameters for the SVM model.

Table 5.1: Support Vector Machine Hyperparameter Values

Hyperparameter	Values
C	$0.001 - 100$
kernel	'linear', 'rbf', 'poly', 'sigmoid'
gamma	'scale', 'auto'
class_weight	'balanced', None

5.5.3 Random Forest

- **Criterion:** The function to measure the quality of a split. Supported criteria are:
  1. **'gini'**: Gini impurity.
  2. **'entropy'**: Information gain.
- **Max Features**: The number of features to consider when looking for the best split. Options include:
  1. **'log2'**: Logarithm of the number of features.

2. **'sqrt'**: Square root of the number of features.
3. **None**: All features are considered.

Table 5.2 summarizes the range of hyperparameters for the RF model.

Table 5.2: Random Forest Hyperparameter Values

Hyperparameter	Values
n_estimators	10 – 500
max_depth	5 – 200
criterion	'gini', 'entropy'
min_sample_split	2 – 20
min_sample_leaf	1 – 10
max_features	'log2', 'sqrt', None
class_weight	'balanced', None

5.5.4 XGBoost

- **Subsample:** The proportion of samples used for training each tree was set between 0.4 and 1.
- **Colsample Bytree:** The fraction of features used for each tree was varied from 0.3 to 1.

Table 5.3 summarizes the range of hyperparameters for the XGBoost model.

Table 5.3: XGBoost Hyperparameter Values

Hyperparameter	Values
learning_rate	0.01 – 1
max_depth	3 – 15
n_estimators	100 – 1000
gamma	0 – 5
colsample_bytree	0.3 – 1.0
min_child_weight	1 – 15
subsample	0.4 – 1.0
class_weight	'balanced', None

5.5.5 Logistic Regression

- **Penalty**: The norm used in the penalization. It helps to regularize the model and prevent overfitting. Options include:
  1. **'elasticnet'**: A combination of L1 and L2 penalties.
  2. **'l1'**: Lasso regularization, which can set some coefficients to zero.
  3. **'l2'**: Ridge regularization, which tends to shrink coefficients.
- **C (Inverse of regularization strength):** Regularization parameter values were tested from 0.01 to 100.
- **L1 Ratio**: The Elastic-Net mixing parameter, with 0 <= l1_ratio <= 1. Only used if 'penalty='elasticnet''. It balances the ratio between L1 and L2 penalties.

The 'LogisticRegression' was configured with the 'saga' solver for its ability to handle both L1 and L2 penalties and to work well with large datasets. The 'saga' solver is particularly suited for Elastic-Net regularization, which combines the benefits of both L1 and L2 penalties. Other solvers available include:

- **'lbfgs'** (default): A quasi-Newton method that is efficient for small to medium-sized datasets.
- **'sag':** Stochastic Average Gradient, suitable for large datasets but only supports L2 regularization.
- **'newton-cg':** An optimization algorithm that handles L2 regularization and is suitable for large datasets. Computationally expensive because of the Hessian Matrix.
- **'liblinear':** A library for large linear classification that supports both L1 and L2 regularization but is not efficient for large datasets.

Table 5.4 summarizes the range of hyperparameters for the LogR model.

Table 5.4: Logistic Regression Hyperparameter Values

Hyperparameter	Values
penalty	'elasticnet', 'l1', 'l2'
C	$0.01 - 100$
max_iter	$100 - 3000$
class_weight	'balanced', None
l1_ratio	$0 - 1$ (Only used if penalty='elasticnet')

5.5.6 k-Nearest Neighbours

- **Number of Neighbors**: The number of neighbors to use was varied from 1 to 30. Choosing the optimal number of neighbors is crucial as smaller values make the model sensitive to noise, while larger values can smooth out the prediction too much.
- **Weights**: The weight function used in prediction. Options include:
  1. **'uniform'**: All neighbors are weighted equally.
  2. **'distance'**: Closer neighbors have a greater influence on the prediction than farther ones.
- **Metric**: The distance metric used to find the nearest neighbors. Options include:
  1. **'euclidean'**: The standard Euclidean distance.
  2. **'manhattan'**: The Manhattan distance, also known as L1 distance.
  3. **'minkowski'**: A generalization of both Euclidean and Manhattan distances.

Table 5.5 summarizes the range of hyperparameters for the KNN model.

Table 5.5: k-Nearest Neighbours Hyperparameter Values

Hyperparameter	Values
n_neighbors	$1-30$
weights	'uniform', 'distance'
metric	'euclidean', 'manhattan', 'minkowski'
p	$1-5$ (Only used if metric is 'minkowski')

5.5.7 Näive Bayes

- **GaussianNB():** The Gaussian Naïve Bayes algorithm was used.
- **Var Smoothing**: This parameter adds a small value to the variance of each feature to prevent division by zero and handle numerical stability. It helps to avoid overfitting by ensuring that the algorithm does not become overly sensitive to variations in the training data.

Table 5.6 summarizes the range of hyperparameters for the NB model.

Table 5.6: Näive Bayes Hyperparameter Values

Hyperparameter	Values
var_smoothing	$1e^{-9}-1$

5.6 Training and Evaluation

The process of training and evaluating the machine learning models involved several critical steps to ensure robust and reliable performance, especially given the class imbalance inherent in the datasets.

Initially, the entire dataset was randomly divided into a training set of 80% and a test set of 20%. This split was essential to evaluate the model's performance on unseen data. Table 5.7 provides a summary of the IDH mutation status of patients in the training and test sets across the different datasets.

The training data then underwent a feature selection / dimensionality reduction process to eliminate irrelevant or redundant features.

Given the class imbalance in both datasets, where IDH mutation-positive cases are underrepresented, classifiers would typically exhibit poor predictive performance for the minority class. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset. SMOTE generates synthetic examples for the minority class by selecting existing minority class samples and creating new synthetic points between them and their k nearest neighbors. This approach enhances the representation of the minority class without merely duplicating existing samples, thereby improving the classifier's ability to learn from these examples [11].

It is important to note that SMOTE was not used to create an equal number of mutant and wild-type samples but to increase the representation of the minority class to an optimal level, calculated alongside the optimal hyperparameters.

Finally, once the training data was balanced and the optimal hyperparameters, including the number of features to retain/create, and SMOTE ratio, were determined through Bayesian optimization, the models were evaluated using several performance metrics:

- **Accuracy:** Measures the proportion of correct predictions (both true positives and true negatives) among the total cases.
- **Balanced Accuracy:** Adjusts for imbalanced datasets by averaging the true positive rate (sensitivity) and the true negative rate (specificity).
- **Sensitivity (Recall):** Measures the model's ability to correctly identify positive cases (IDH mutation).
- **Specificity:** Measures the model's ability to correctly identify negative cases (wild type).
- **Precision:** Measures the proportion of true positive predictions among all positive predictions.
- **F1 Score:** The harmonic mean of precision and recall, providing a balanced measure of the model's performance.

- **Area Under the ROC Curve:** Measures the model's ability to discriminate between classes, regardless of the threshold chosen.

These metrics are analysed in Chapter 6, where the results are discussed in detail to evaluate the efficacy of the models in predicting IDH mutation status.

Table 5.7: IDH Mutation Status of Patients in Training and Test Sets Across the Three Datasets.

Dataset	Subset	IDH Status	Total
UCSF-PDGM (Low and High Grade gliomas)	Training (n = 394)	Wildtype	314
		Mutant	80
	Test (n = 99)	Wildtype	79
		Mutant	20
UPENN-GBM (High Grade gliomas)	Training (n = 411)	Wildtype	398
		Mutant	13
	Test (= 103)	Wildtype	100
		Mutant	3
Both Combined	Training (n = 805)	Wildtype	712
		Mutant	93
	Test (n = 201)	Wildtype	178
		Mutant	23

5.7 Interactive Tool

The interactive tool developed as part of this research facilitates the visualization, analysis, and interpretation of radiogenomic data used to predict IDH mutation status in glioma patients. This section details the design, functionality, and significance of the interactive tool. Utilizing Qt Designer (version 5.11.1 - PyQt5), a robust framework comprehensively discussed in Chapter 2, Section 2.8, the tool incorporates a range of features essential for the study's analytical needs. The UI was designed to be intuitive and user-friendly, enabling users to seamlessly navigate through various functionalities. Below is a brief explanation of its key functionalities.

5.7.1 MRI Image and ROI Visualization

Users can select and visualize each MRI image from the two datasets, along with their corresponding ROIs. This feature allows for detailed examination of the tumor structures and facilitates better understanding of the imaging data. Figure 5.4 shows the interface of the interactive tool where users can access this functionality.

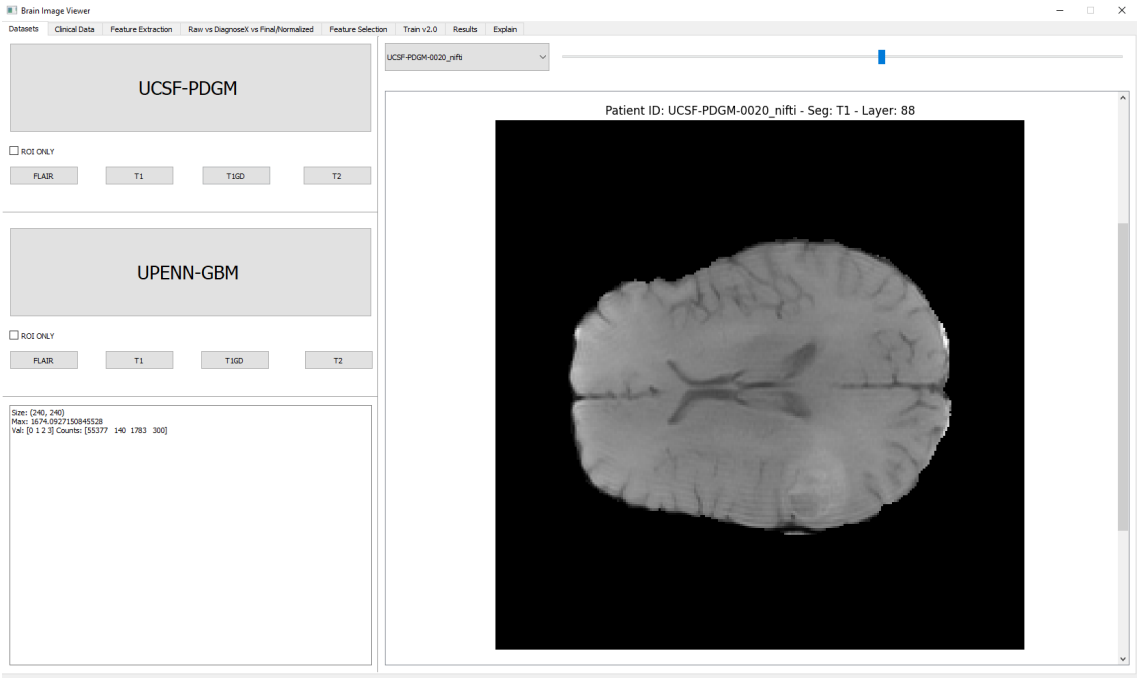


Figure 5.4: Screenshot of the interactive tool's image visualization feature. Displayed is patient number 20 from the UCSF-PDGM dataset, with T1 sequence and slice number 88. Users can select between datasets, choose specific patients, MRI sequences, and slices using a slider. Additionally, users can opt to display either only the ROI or the entire brain.

5.7.2 Clinical Data Visualization

The tool features a section dedicated to visualizing clinical data associated with the datasets. This section presents various graphs and charts that help users understand the distribution and characteristics of the clinical data. Figure 5.5 illustrates the interface of the interactive tool where users can access these visualizations.

5.7.3 Feature Extraction Command Page

The tool provides a command interface to initiate the feature extraction process from the selected dataset. Users can choose the dataset and start the automated extraction process, which is executed efficiently. After feature extraction, the tool displays the extracted features in a tabular format. This table allows users to inspect the features and their values, providing a comprehensive view of the radiomic data. Figure 5.6 depicts the interface of the interactive tool where users can access this functionality.

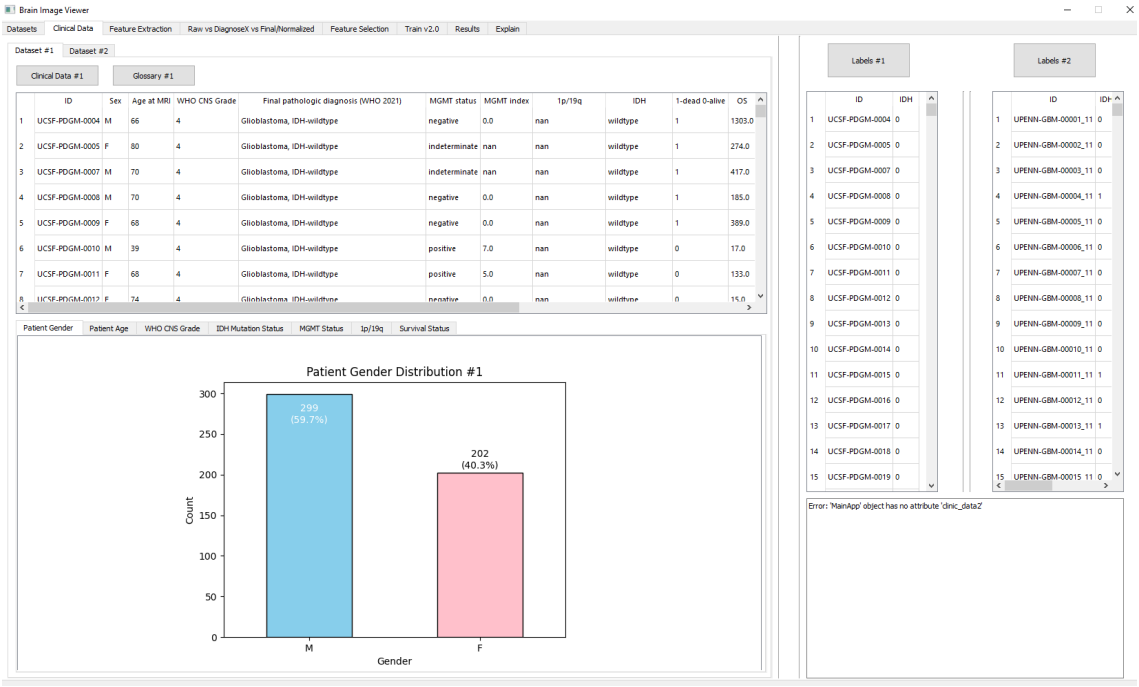


Figure 5.5: Screenshot of the interactive tool's clinical data visualization feature. Shown is clinical data from the UCSF-PDGM dataset, including a graph depicting patient gender distribution and target labels (IDH status) on the right.

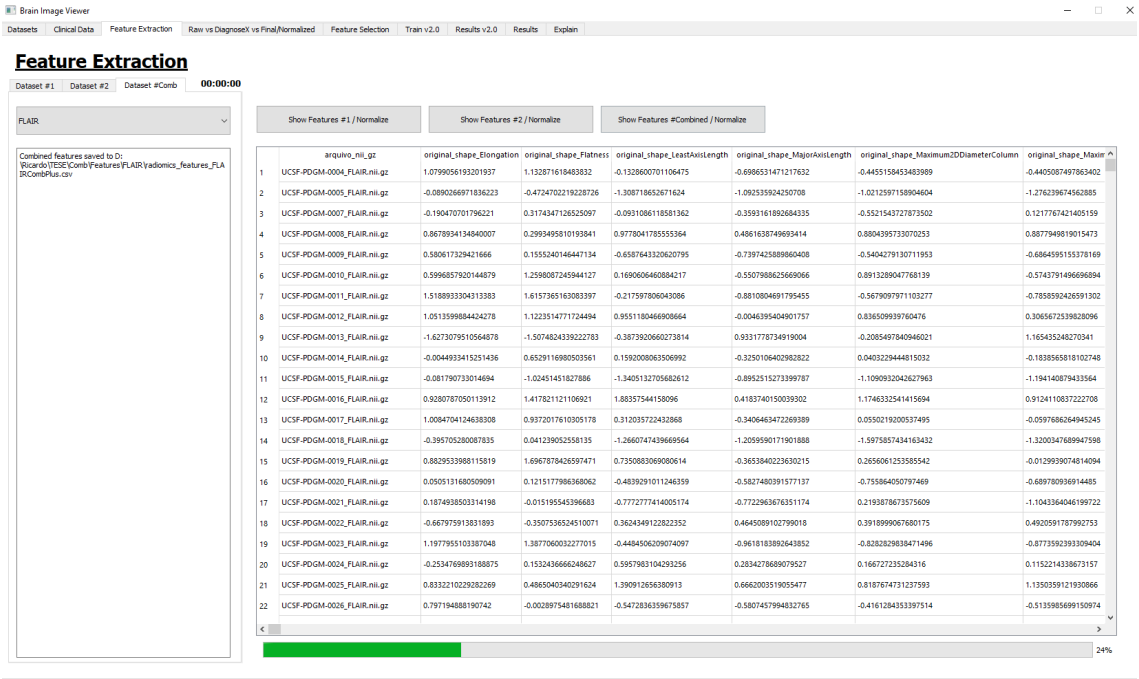


Figure 5.6: Screenshot of the interactive tool's feature extraction command page. Users can select the dataset and initiate the extraction process. The extracted features are then displayed in a table for detailed inspection.

5.7.4 Training Station

The tool features a Training Station where users can access the most important controls for model training. Here users can select from the three datasets (UCSF-PDGM, UPENN-GBM, and the combined dataset) and choose the desired classifier (SVM, RF, XGBoost, LogR, kNN, or Naïve Bayes). Additionally, users can select MRI sequences (T1, T2, T1GD, FLAIR) and feature selection techniques (None, Mutual Information, Boruta, PCA, or LDA). The Training Station also offers users the option to use SMOTE for handling class imbalance, to tune hyperparameters of the selected model, and to enter the k features to retain while using feature selection; otherwise k is automatically selected. Once a model is trained, the tool provides various graphical representations and matrices to evaluate the model's performance. Users can visualize metrics such as accuracy, precision, recall, and ROC curves, which help in assessing the efficacy of the predictive model. Figure 5.7 depicts the interface of the interactive tool where users can access this functionality.

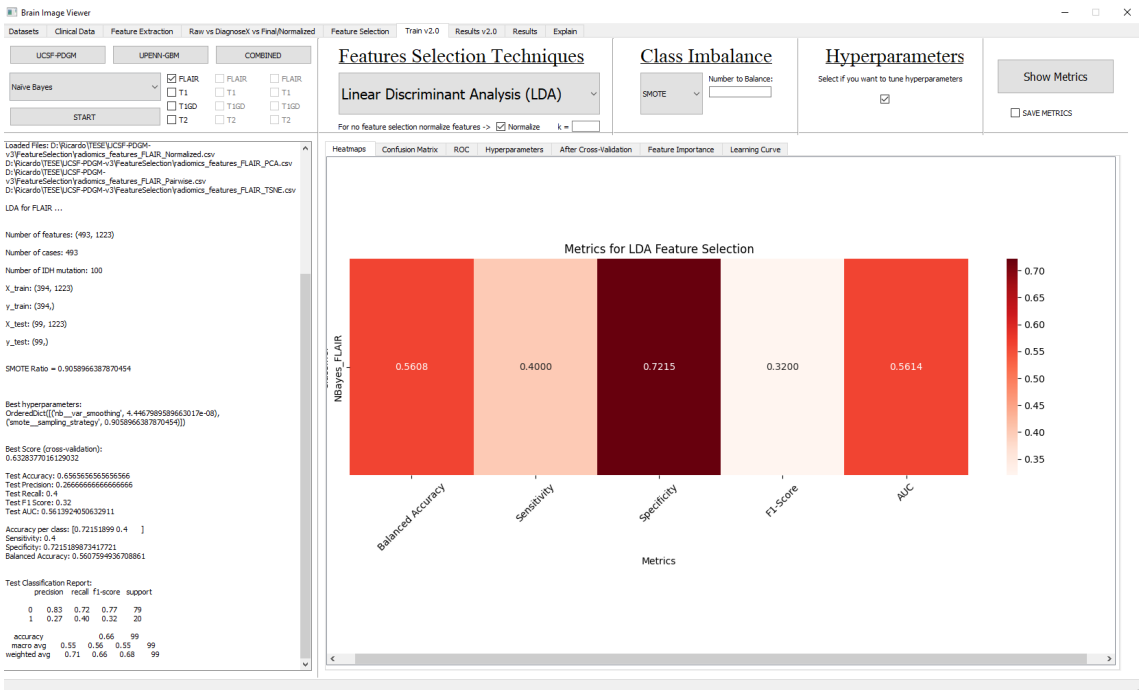


Figure 5.7: Screenshot of the interactive tool's Training Station. Users can select the dataset, choose the classifier, apply feature selection techniques, and initiate the training process. The interface also provides options for handling class imbalance with SMOTE and tuning model hyperparameters.

5.7.5 Importance and Impact

This tool serves as a valuable resource for future researchers by providing a comprehensive and user-friendly platform for analyzing MRI imaging and clinical data. Instead of starting from scratch, researchers can leverage this tool to understand the overall methodology implemented. The intuitive interface facilitates the examination of tumor structures, the visualization of clinical data, and the extraction of significant features, streamlining the research process.

Moreover, the tool's Training Station enables users to experiment with various datasets, classifiers, and feature selection techniques, enabling them to build and evaluate predictive models efficiently.

Overall, this tool not only accelerates the research workflow but also enhances the accuracy and reliability of the findings. Its user-friendly design makes complex data analysis accessible to a broader range of researchers, fostering innovation and advancement in the field of medical imaging and radiomics.

5.8 Summary

Chapter 5 outlines the methodology employed in this study to predict IDH mutation status using MRI analysis and machine learning techniques. The key steps in the methodology are summarized as follows:

1. **Dataset Collection and Preparation:**
   (a) Two primary datasets, UCSF-PDGM and UPENN-GBM, were utilized. Each dataset was divided into training (80%) and test (20%) sets.
   (b) The training data underwent a feature selection and dimensionality reduction process to enhance model efficiency and performance by retaining the most significant features.
2. **Handling Class Imbalance:**
   Given the class imbalance in both datasets, where IDH mutation-positive cases were underrepresented, the SMOTE was applied to the training data.
3. **Model Training and Hyperparameter Optimization:**
   (a) Various machine learning classifiers, including Random Forest, Support Vector Machine, eXtreme Gradient Boosting, Logistic Regression, Naïve Bayes, and K-Nearest Neighbors, were employed.
   (b) Bayesian optimization was used to determine the optimal hyperparameters for each classifier, including the number of features to retain/create and the SMOTE ratio.
4. **Model Evaluation:**
   (a) The performance of the trained models was evaluated using several metrics: Accuracy, Balanced Accuracy, Sensitivity (Recall), Specificity, Precision, F1 Score, and AUC.

Additionally, an interactive tool was developed to enhance the visualization and analysis of MRI images and clinical data. This tool enables users to view MRI images and ROIs, present clinical information comprehensively, extract radiomic features, and train machine learning models directly.

This chapter provides a detailed description of the experimental design and procedures, ensuring a systematic approach to developing and validating machine learning models for predicting IDH mutation status. The subsequent analysis and discussion of results are presented in Chapter 6

Chapter 6

Results and Discussions

This chapter presents the findings of our study on predicting IDH mutation status in gliomas using machine learning algorithms applied to MRI data. It is divided into three main sections, each dedicated to the results and discussions for one of the three datasets used in this study (Sections 6.1, 6.2, 6.3). Within each section, the classification results are provided, the optimal hyperparameters are identified, and the limitations encountered are discussed.

The results detailed in this chapter are focused on showcasing the best-performing models and their configurations. These results were obtained from 10 different runs to ensure statistical significance and robustness. For a complete overview of all the experimental results, including additional metrics and model configurations, please refer to Appendix A. Here, the most promising outcomes from the experiments are highlighted, emphasizing the models that demonstrated the highest robustness in predicting IDH mutation status.

6.1 Results for UCSF-PDGM dataset

This section presents the optimal combinations of classifier, MRI sequence, and feature selection technique, alongside the baseline model without feature selection. Additionally, the SMOTE ratio for each combination is provided, as well as the k features retained / created by the feature selection techniques. Tables 6.1 to 6.6 display the best results for each classifier.

Support Vector Machine

Table 6.1: Best Performances of SVM model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.731±0.145	0.751±0.170	0.620±0.286	0.841±0.110	0.26	1223
	Mutual Info	0.700±0.141	0.739±0.165	0.520±0.289	0.880±0.123	0.64	854
T1	None	0.709±0.151	0.746±0.172	0.550±0.031	0.869±0.134	0.26	1223
	Mutual Info	0.659±0.124	0.736±0.160	0.380±0.271	0.939±0.092	0.35	1086
T1GD	**None**	**0.748±0.141**	**0.744±0.181**	**0.700±0.300**	**0.796±0.118**	**1.00**	**1223**
	Boruta	0.728±0.145	0.747±0.152	0.540±0.294	0.916±0.100	0.55	19
T2	None	0.745±0.159	0.757±0.178	0.570±0.307	0.921±0.097	0.54	1223
	Boruta	0.685±0.091	0.737±0.176	0.300±0.458	0.700±0.458	0.38	21

k-Nearest Neighbors

Table 6.2: Best Performances of KNN model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.696±0.109	0.759±0.151	0.480±0.213	0.913±0.091	0.38	1223
	Boruta	0.698±0.113	0.774±0.151	0.450±0.233	0.946±0.064	0.36	21
T1	None	0.655±0.111	0.705±0.188	0.400±0.232	0.910±0.099	0.27	1223
	Mutual Info	0.639±0.095	0.695±0.158	0.350±0.175	0.928±0.099	0.36	624
T1GD	None	0.674±0.108	0.757±0.162	0.430±0.253	0.918±0.099	0.26	1223
	Boruta	**0.733±0.126**	**0.751±0.203**	**0.550±0.269**	**0.915±0.106**	**0.26**	**19**
T2	None	0.681±0.117	0.765±0.188	0.420±0.227	0.941±0.065	0.26	1223
	Mutual Info	0.701±0.159	0.750±0.177	0.570±0.293	0.831±0.123	0.26	21

XGBoost

Table 6.3: Best Performances of XGBoost model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.684±0.122	0.749±0.145	0.470±0.245	0.897±0.111	1.0	1223
	PCA	0.707±0.149	0.749±0.143	0.550±0.304	0.864±0.100	1.00	7
T1	None	0.529±0.033	0.722±0.175	0.090±0.070	0.969±0.056	0.26	1223
	Boruta	0.700±0.139	0.717±0.191	0.530±0.287	0.870±0.094	0.93	4
T1GD	None	0.500±0.000	0.785±0.172	0.000±0.000	1.000±0.000	0.26	1223
	Mutual Info	**0.752±0.160**	**0.776±0.190**	**0.630±0.320**	**0.875±0.111**	**0.27**	**68**
T2	None	0.662±0.113	0.728±0.185	0.380±0.248	0.944±0.080	0.26	1223
	Boruta	0.706±0.128	0.746±0.191	0.550±0.262	0.862±0.091	1.00	8

Random Forest

Table 6.4: Best Performances of RF model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.691±0.148	0.757±0.176	0.490±0.277	0.893±0.101	1.0	1223
	Mutual Info	0.687±0.146	0.756±0.186	0.460±0.276	0.913±0.103	0.64	419
T1	None	0.630±0.091	0.726±0.188	0.300±0.195	0.959±0.075	0.35	1223
	Mutual Info	0.657±0.097	0.731±0.188	0.360±0.206	0.954±0.084	0.31	863
T1GD	None	0.685±0.130	0.775±0.161	0.470±0.272	0.900±0.093	0.62	1223
	Boruta	**0.762±0.166**	**0.776±0.186**	**0.660±0.326**	**0.864±0.115**	**0.49**	**19**
T2	None	0.690±0.123	0.771±0.189	0.460±0.262	0.921±0.073	0.58	1223
	Mutual Info	0.691±0.140	0.756±0.153	0.490±0.270	0.893±0.076	0.81	699

Logistic Regression

Table 6.5: Best Performances of LogR model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.731±0.147	0.732±0.200	0.560±0.332	0.903±0.099	0.48	1223
	PCA	0.748±0.158	0.758±0.183	0.670±0.310	0.826±0.130	1.00	38
T1	None	0.728±0.138	0.757±0.177	0.660±0.291	0.796±0.115	0.35	1223
	Boruta	0.726±0.147	0.731±0.171	0.620±0.309	0.831±0.127	0.35	16
T1GD	None	0.747±0.178	0.770±0.189	0.660±0.326	0.833±0.109	0.87	1223
	Boruta	**0.760±0.146**	**0.767±0.180**	**0.660±0.317**	**0.859±0.122**	**0.35**	**19**
T2	None	0.754±0.163	0.741±0.211	0.670±0.335	0.838±0.108	0.76	1223
	PCA	0.703±0.118	0.764±0.175	0.460±0.262	0.946±0.071	0.35	19

Naïve Bayes

Table 6.6: Best Performances of NB model Across different MRI Sequences and Feature Selection Methods for UCSF-PDGM. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.649±0.137	0.686±0.167	0.650±0.220	0.649±0.137	0.42	1223
	Mutual Info	0.662±0.136	0.687±0.156	0.510±0.270	0.814±0.097	0.35	158
T1	None	0.622±0.119	0.674±0.169	0.390±0.238	0.854±0.106	0.69	1223
	Mutual Info	0.623±0.109	0.677±0.171	0.350±0.250	0.895±0.084	0.63	196
T1GD	None	0.754±0.153	0.748±0.175	0.680±0.309	0.829±0.106	0.77	1223
	Mutual Info	**0.737±0.166**	**0.773±0.185**	**0.660±0.326**	**0.813±0.115**	**0.35**	**43**
T2	None	0.703±0.177	0.730±0.173	0.600±0.332	0.806±0.099	0.35	1223
	Mutual Info	0.627±0.124	0.721±0.158	0.380±0.223	0.875±0.098	0.37	9

Comparison Between Models

The comparative analysis of various classifiers for identifying IDH mutation status in gliomas reveals that the Random Forest classifier outperforms other models in terms of overall performance metrics. Specifically, the RF classifier achieved a Balanced Accuracy of 0.762 ± 0.166 and an AUC value of 0.776 ± 0.186, indicating its robust capability in classifying IDH mutation status in gliomas (Table 6.7).

The relatively high AUC value signifies the model's proficiency in distinguishing between positive and negative samples, underscoring its reliability for this specific application. The standard deviation of 0.166 for balanced accuracy and 0.186 for AUC in the Random Forest classifier indicates a moderate variability in performance across different cross-validation folds. This suggests that while the model performs well on average, its performance can fluctuate depending on the specific training and validation splits. This variability needs to be considered when deploying the model in real-world scenarios, as it implies potential inconsistency in predictions.

In the context of data augmentation, the results demonstrate that optimal values were generally found with higher SMOTE ratios. This means that adding more synthetic positive samples to balance the dataset tends to enhance model performance. Notably, for some models like K-Nearest Neighbors the average SMOTE ratio was around 0.26, translating to 314 negative cases and approximately 80 positive cases, which is similar to the original training set. This suggests that while higher SMOTE ratios are beneficial in most cases, there are exceptions where a lower ratio might suffice or even be optimal. Research has shown that oversampling can sometimes lead to overgeneralization, creating false decision boundaries due to the introduction of too many synthetic samples. This indicates that a balanced approach, potentially with a lower SMOTE ratio, might sometimes yield better results by avoiding these pitfalls [46].

Regarding feature selection techniques, mutual information and Boruta consistently achieved the best results across various classifiers. However, in some instances (LogR and XGBoost), PCA also showed superior performance. This indicates that while classifiers generally benefit from feature selection techniques, the choice of technique can significantly impact the outcome. Interestingly, for SVM almost every experiment indicated that feature selection did not provide substantial benefits, suggesting that the baseline model's performance might already be optimal. Conversely, LDA was never the best feature selection method in any scenario, implying its limited utility for this particular classification task.

Analyzing the MRI sequences used, it is evident that the T1GD sequence yields the best results across all classifiers. This sequence consistently provided the highest performance metrics, highlighting its importance in accurately determining IDH mutation status. The superior performance of the T1GD sequence could be attributed to its enhanced imaging capabilities, which provide clearer and more distinct features crucial for the classification task.

Table 6.7: Comparison of best-performing classifiers for UCSF-PDGM.

Classifier	MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
SVM	T1GD	None	0.748±0.141	0.744±0.181	0.700±0.300	0.796±0.118
KNN	T1GD	Boruta	0.733±0.126	0.751±0.203	0.550±0.269	0.915±0.106
XGBoost	T1GD	Mutual Info	0.752±0.160	0.776±0.190	0.630±0.320	0.875±0.111
RF	T1GD	Boruta	0.762±0.166	0.776±0.186	0.660±0.326	0.864±0.115
LogR	T1GD	Boruta	0.760±0.146	0.767±0.180	0.660±0.317	0.859±0.122
NB	T1GD	None	0.754±0.153	0.748±0.175	0.680±0.309	0.829±0.106

6.1.1 Best Hyperparameters

This subsection presents the best hyperparameters found for each of the best-performing classifiers used in the experiments (Table 6.8).

Table 6.8: Best Hyperparameters found for each of the Best-Performing Classifiers for UCSF-PDGM

Classifier	Hyperparameter	Value
SVM	C	0.1423
	kernel	sigmoid
	gamma	auto
	class_weight	balanced
KNN	n_neighbors	10
	weights	distance
	metric	manhattan
	p	1
XGBoost	learning_rate	0.108
	max_depth	15
	n_estimators	1000
	gamma	1.937
	colsample_bytree	1.0
	min_child_weight	15
	subsample	0.4
	class_weight	balanced
RF	n_estimator	475
	max_depth	128
	criterion	gini
	min_sample_split	2
	min_sample_leaf	10
	max_features	sqrt
	class_weight	balanced
LogR	C	0.01
	penalty	L2
	max_iter	802
	l1_ratio	None
	class_weight	balanced
NB	var_smoothing	1.0

6.1.2 Limitations

While the study demonstrates promising results in classifying IDH mutation status in gliomas, several limitations must be acknowledged to provide a comprehensive understanding of the challenges encountered. These limitations arise from the dataset composition, the nature of IDH mutations, and the inherent variability in the data.

The dataset used in this study comprises both low-grade (19.8%) and high-grade (80.2%) gliomas, with a significant overrepresentation of high-grade gliomas. This imbalance can potentially bias the model's performance, as the classifier might be more attuned to the characteristics of high-grade gliomas, possibly overlooking nuances present in low-grade gliomas. The skewed distribution necessitates careful consideration during model training and evaluation to ensure that the classifier can generalize across different glioma grades.

A notable limitation is the presence of unspecified IDH mutations within the dataset. The study aims to classify whether an IDH mutation is present or not, without distinguishing between specific types of mutations such as IDH1 or IDH2. This generalization can introduce difficulties because the dataset does not just encompass a single type of mutation but multiple types, each with its own characteristics. IDH2 mutations are much less common than IDH1 mutations and are mutually exclusive with IDH1 mutations. This heterogeneity within the dataset can complicate the classification task, as the model needs to learn to recognize a broader range of mutation types.

There is a significant imbalance between the number of positive (mutated) and negative (non-mutated) cases in the dataset. Such class imbalance can adversely affect the model's performance, leading to a bias towards the majority class (non-mutated cases). Although techniques like SMOTE were employed to mitigate this issue by augmenting the dataset with synthetic positive samples, the inherent imbalance still poses a challenge. It can lead to scenarios where the model performs well on negative cases but struggles with correctly identifying positive cases, ultimately affecting the overall reliability of the classifier.

Another limitation is the inclusion of cases with prior biopsies that were not excluded from the dataset. These prior biopsies could influence the results, as they may affect the MRI features used for classification. However, this factor is present in both positive and negative cases, indicating that its influence is distributed across the dataset. Despite this, the potential impact of prior biopsies should be considered, as it might introduce additional variability into the model's performance.

6.2 Results for UPENN-GBM dataset

This section presents the optimal combinations of classifier, MRI sequence, and feature selection technique, alongside the baseline model without feature selection. Additionally, the SMOTE ratio for each combination is provided, as well as the k features retained / created by the feature selection techniques. Tables 6.9 to 6.14 display the best results for each classifier.

Support Vector Machine

Table 6.9: Best Performances of SVM model Across different MRI Sequences and Feature Selection Methods. The best combination is highlighted in bold for UPENN-GBM.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.569±0.161	0.560±0.315	0.250±0.335	0.888±0.032	0.36	1223
	Mutual Info	0.640±0.199	0.644±0.359	0.300±0.400	0.980±0.013	0.72	954
T1	None	0.539±0.147	0.679±0.244	0.100±0.300	0.978±0.017	0.47	1223
	PCA	0.662±0.196	0.774±0.168	0.350±0.391	0.974±0.016	0.63	37
T1GD	**None**	**0.816±0.161**	**0.850±0.174**	**0.850±0.320**	**0.783±0.070**	**0.24**	**1223**
	PCA	0.721±0.199	0.843±0.186	0.600±0.374	0.842±0.224	0.35	24
T2	None	0.679±0.236	0.796±0.213	0.400±0.490	0.958±0.026	0.37	1223
	Mutual Info	0.695±0.245	0.841±0.199	0.400±0.490	0.990±0.010	0.34	1004

k-Nearest Neighbors

Table 6.10: Best Performances of KNN model Across different MRI Sequences and Feature Selection Methods. The best combination is highlighted in bold for UPENN-GBM.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.421±0.026	0.478±0.185	0.000±0.000	0.841±0.051	0.45	1223
	Mutual Info	0.639±0.208	0.685±0.281	0.550±0.415	0.729±0.069	1.0	432
T1	None	0.655±0.235	0.746±0.256	0.550±0.472	0.759±0.055	1.0	1223
	Mutual Info	0.702±0.236	0.747±0.273	0.550±0.472	0.853±0.046	0.35	1203
T1GD	**None**	**0.779±0.196**	**0.771±0.282**	**0.750±0.403**	**0.807±0.064**	**0.74**	**1223**
	Mutual Info	0.759±0.201	0.787±0.246	0.750±0.403	0.767±0.078	0.79	288
T2	None	0.676±0.225	0.750±0.224	0.550±0.472	0.801±0.068	0.69	1223
	Boruta	0.675±0.224	0.739±0.253	0.500±0.447	0.849±0.062	0.36	21

XGBoost

Table 6.11: Best Performances of XGBoost model Across different MRI Sequences and Feature Selection Methods. The best combination is highlighted in bold for UPENN-GBM.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.639±0.201	0.643±0.288	0.300±0.400	0.978±0.027	0.84	1223
	PCA	0.634±0.213	0.628±0.297	0.450±0.415	0.818±0.074	0.49	3
T1	None	0.642±0.204	0.800±0.198	0.300±0.400	0.984±0.018	0.42	1223
	Mutual Info	0.623±0.200	0.778±0.216	0.250±0.403	0.996±0.008	0.99	754
T1GD	None	0.689±0.222	0.908±0.154	0.400±0.436	0.978±0.019	0.72	1223
	PCA	**0.754±0.208**	**0.840±0.183**	**0.750±0.403**	**0.759±0.088**	**0.98**	**1**
T2	None	0.548±0.151	0.772±0.214	0.100±0.300	0.996±0.008	0.47	1223
	Mutual Info	0.641±0.201	0.709±0.221	0.300±0.400	0.982±0.011	1.0	640

Random Forest

Table 6.12: Best Performances of RF model Across different MRI Sequences and Feature Selection Methods. The best combination is highlighted in bold for UPENN-GBM.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.598±0.166	0.632±0.316	0.200±0.332	0.996±0.008	0.35	1223
	Boruta	0.593±0.169	0.641±0.326	0.200±0.332	0.986±0.016	0.28	11
T1	None	0.548±0.151	0.774±0.265	0.100±0.300	0.996±0.008	0.36	1223
	Boruta	0.547±0.151	0.750±0.269	0.100±0.300	0.994±0.009	0.72	7
T1GD	None	0.595±0.200	0.788±0.261	0.200±0.400	0.990±0.013	0.36	1223
	PCA	**0.744±0.178**	**0.754±0.289**	**0.650±0.391**	**0.837±0.083**	**0.31**	**9**
T2	None	0.597±0.199	0.749±0.306	0.200±0.400	0.994±0.009	0.77	1223
	Boruta	0.597±0.199	0.749±0.306	0.200±0.400	0.994±0.009	0.24	16

Logistic Regression

Table 6.13: Best Performances of LogR model Across different MRI Sequences and Feature Selection Methods. The best combination is highlighted in bold for UPENN-GBM.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.612±0.180	0.710±0.272	0.350±0.391	0.873±0.068	0.44	1223
	PCA	0.677±0.194	0.728±0.315	0.400±0.374	0.954±0.033	0.37	6
T1	None	0.757±0.216	0.797±0.257	0.600±0.436	0.914±0.043	0.33	1223
	Mutual Info	0.753±0.185	0.790±0.258	0.600±0.374	0.906±0.048	0.23	1124
T1GD	None	0.756±0.199	0.852±0.193	0.700±0.400	0.811±0.033	0.66	1223
	PCA	**0.771±0.204**	**0.848±0.157**	**0.650±0.391**	**0.892±0.051**	**0.49**	**9**
T2	None	0.712±0.234	0.869±0.179	0.450±0.472	0.974±0.025	0.64	1223
	PCA	0.728±0.218	0.842±0.188	0.500±0.447	0.956±0.023	1.0	50

Naïve Bayes

Table 6.14: Best Performances of NB model Across different MRI Sequences and Feature Selection Methods. The best combination is highlighted in bold for UPENN-GBM.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.643±0.177	0.645±0.194	0.750±0.403	0.536±0.144	0.35	1223
	PCA	0.643±0.192	0.718±0.218	0.350±0.391	0.936±0.040	1.0	1
T1	None	0.675±0.179	0.666±0.225	0.750±0.403	0.601±0.124	0.46	1223
	Mutual Info	0.697±0.215	0.754±0.270	0.600±0.436	0.793±0.084	0.35	162
T1GD	None	0.781±0.151	0.826±0.164	0.750±0.403	0.811±0.078	0.57	1223
	PCA	**0.833±0.169**	**0.822±0.186**	**0.850±0.320**	**0.815±0.078**	**0.35**	**11**
T2	None	0.635±0.173	0.669±0.196	0.800±0.332	0.470±0.134	0.36	1223
	PCA	0.617±0.162	0.744±0.262	0.250±0.335	0.984±0.018	0.47	1

Comparison Between Models

The comparative analysis of various classifiers using the UPENN-GBM dataset reveals that SVM and Naïve Bayes classifiers demonstrated the best performance. Specifically, the SVM classifier achieved a balanced accuracy of 0.816 ± 0.161 and an AUC of 0.850 ± 0.174, while the NB classifier obtained a balanced accuracy of 0.833 ± 0.169 and an AUC of 0.822 ± 0.186 (Table 6.15). These results indicate that, overall, this dataset shows lower performance metrics compared to the first dataset. However, the best-performing models in the UPENN-GBM dataset managed to surpass the top results from the initial dataset.

The standard deviation observed in the performance metrics for this dataset is higher compared to the first dataset, indicating greater fluctuations in model performance across different cross-validation folds. This higher variability suggests that the models' performances are more sensitive to the specific training and validation splits in this dataset, which could be attributed to underlying data inconsistencies or inherent complexity.

Regarding data augmentation, the results indicate that models generally benefit from the addition of synthetic positive samples. This pattern aligns with observations from the first dataset, reinforcing the utility of techniques like SMOTE in balancing the dataset and enhancing model performance.

Contrary to the findings from the first dataset, here PCA emerged as the feature selection technique yielding the best results. PCA not only improved model performance but also was the method that most reduced the number of features, thereby decreasing the execution time. However, there are notable exceptions where the baseline models without any feature selection outperformed those with feature selection, particularly for SVM and KNN. This inconsistency highlights the importance of selecting appropriate feature selection techniques based on the specific classifier and dataset characteristics. Once more, LDA never proved to be the optimal feature selection method.

The T1GD MRI sequence once again demonstrated superior performance across all classifiers, consistent with findings from the initial dataset. This sequence provided the highest accuracy and AUC values, underscoring its importance in accurately determining IDH mutation status.

Table 6.15: Comparison of best-performing classifiers for UPENN-GBM.

Classifier	**MRI Seq**	**Feature Selection**	**Balanced Accuracy**	**AUC**	**Sensitivity**	**Specificity**
SVM	T1GD	None	0.816±0.161	0.850±0.174	0.850±0.320	0.783±0.070
KNN	T1GD	None	0.779±0.196	0.771±0.282	0.750±0.403	0.807±0.064
XGBoost	T1GD	PCA	0.754±0.208	0.840±0.183	0.750±0.403	0.759±0.088
RF	T1GD	PCA	0.744±0.178	0.754±0.289	0.650±0.391	0.837±0.083
LogR	T1GD	PCA	0.771±0.204	0.848±0.157	0.650±0.391	0.892±0.051
NB	T1GD	PCA	0.833±0.169	0.822±0.186	0.850±0.320	0.815±0.078

6.2.1 Best Hyperparameters

This subsection presents the best hyperparameters found for each of the best-performing classifiers used in the experiments (Table 6.16).

Table 6.16: Best Hyperparameters found for each of the Best-Performing Classifiers for UPENN-GBM

Classifier	Hyperparameter	Value
SVM	C	11.727
	kernel	linear
	gamma	scale
	class_weight	balanced
KNN	n_neighbors	30
	weights	uniform
	metric	minkowski
	p	5
XGBoost	learning_rate	0.1746
	max_depth	3
	n_estimators	229
	gamma	3.340
	colsample_bytree	1.0
	min_child_weight	6
	subsample	0.819
	class_weight	balanced
RF	n_estimator	112
	max_depth	200
	criterion	entropy
	min_sample_split	20
	min_sample_leaf	10
	max_features	None
	class_weight	balanced
LogR	C	0.907
	penalty	l2
	max_iter	100
	l1_ratio	None
	class_weight	balanced
NB	var_smoothing	1.0

6.2.2 Limitations

While the analysis of the UPENN-GBM dataset has yielded promising results, several limitations must be considered to understand the constraints and challenges faced in this study. These limitations impact the generalizability, performance, and robustness of the developed classifiers.

The UPENN-GBM dataset exclusively comprises cases of high-grade gliomas (glioblastomas). This specificity might contribute to the better performance metrics observed; however, it also limits the generalizability of the models. Since the models are trained only on high-grade gliomas, their applicability to low-grade gliomas may be compromised.

A significant limitation in the UPENN-GBM dataset is the pronounced imbalance between positive (mutation) and negative (non-mutation) cases. Approximately 3% of the cases are positive, compared to about 20% in the UCSF-PDGM dataset. This extreme imbalance can bias the model towards the majority class, reducing its ability to accurately classify positive cases. Although data augmentation techniques like SMOTE were used to address this issue, the inherent imbalance still poses a challenge and may affect the reliability of the model's predictions.

Unlike the first dataset, the UPENN-GBM dataset contains only one type of mutation, classified as IDH1. This limitation simplifies the classification task but also reduces the model's ability to generalize to other IDH mutation types.

Similar to the first dataset, the inclusion of cases with prior biopsies that were not excluded presents another limitation. Prior biopsies can influence the MRI features, potentially affecting the classification results. Future studies should consider excluding or separately analyzing cases with prior biopsies to minimize this confounding factor.

6.3 Results for Combined Dataset

This section presents the optimal combinations of classifier, MRI sequence, and feature selection technique, alongside the baseline model without feature selection. Additionally, the SMOTE ratio for each combination is provided, as well as the *k* features retained / created by the feature selection techniques. Tables 6.17 to 6.22 display the best results for each classifier.

Support Vector Machine

Table 6.17: Best Performances of SVM model Across different MRI Sequences and Feature Selection Methodsfor Combined Dataset. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.789±0.115	0.859±0.113	0.688±0.274	0.891±0.104	0.35	1223
	Mutual Info	0.778±0.116	0.853±0.123	0.636±0.268	0.920±0.094	0.58	723
T1	None	0.750±0.122	0.865±0.098	0.561±0.290	0.939±0.078	0.35	1223
	Mutual Info	0.737±0.135	0.866±0.105	0.501±0.289	0.974±0.030	0.35	952
T1GD	**None**	**0.848±0.108**	**0.932±0.059**	**0.824±0.222**	**0.871±0.061**	**0.91**	**1223**
	Boruta	0.831±0.094	0.927±0.072	0.695±0.202	0.966±0.038	0.57	19
T2	None	0.802±0.094	0.880±0.091	0.653±0.221	0.952±0.055	0.49	1223
	Boruta	0.766±0.084	0.838±0.100	0.661±0.205	0.872±0.100	0.43	11

k-Nearest Neighbors

Table 6.18: Best Performances of KNN model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.765±0.105	0.822±0.117	0.617±0.237	0.912±0.097	0.36	1223
	Boruta	0.771±0.111	0.822±0.135	0.601±0.244	0.942±0.075	0.35	21
T1	None	0.751±0.109	0.799±0.135	0.567±0.244	0.935±0.071	0.27	1223
	Boruta	0.760±0.112	0.814±0.129	0.567±0.253	0.952±0.064	0.36	16
T1GD	None	0.817±0.100	0.917±0.077	0.705±0.222	0.929±0.045	0.26	1223
	PCA	**0.825±0.091**	**0.915±0.077**	**0.739±0.208**	**0.910±0.049**	**0.53**	**30**
T2	None	0.759±0.083	0.849±0.086	0.721±0.212	0.796±0.108	0.92	1223
	Mutual Info	0.809±0.075	0.851±0.074	0.685±0.168	0.934±0.068	0.44	427

XGBoost

Table 6.19: Best Performances of XGBoost model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.762±0.123	0.871±0.115	0.577±0.278	0.948±0.066	0.53	1223
	PCA	0.766±0.138	0.863±0.096	0.568±0.289	0.963±0.041	0.36	18
T1	None	0.749±0.099	0.864±0.116	0.541±0.232	0.957±0.059	0.59	1223
	Mutual Info	0.749±0.085	0.891±0.096	0.523±0.190	0.975±0.035	0.35	1196
T1GD	None	0.834±0.068	0.935±0.066	0.702±0.156	0.965±0.037	0.35	1223
	PCA	**0.835±0.098**	**0.946±0.048**	**0.705±0.216**	**0.966±0.037**	**0.36**	**39**
T2	None	0.787±0.105	0.893±0.079	0.627±0.233	0.947±0.052	0.97	1223
	Boruta	0.779±0.088	0.879±0.086	0.652±0.206	0.907±0.078	1.00	8

Random Forest

Table 6.20: Best Performances of RF model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.746±0.107	0.875±0.108	0.516±0.235	0.975±0.035	0.45	1223
	Mutual Info	0.746±0.109	0.888±0.106	0.515±0.232	0.978±0.033	0.43	1006
T1	None	0.736±0.106	0.869±0.107	0.497±0.227	0.975±0.035	0.35	1223
	Mutual Info	0.839±0.063	0.924±0.061	0.719±0.147	0.959±0.039	0.34	863
T1GD	None	0.685±0.130	0.775±0.161	0.470±0.272	0.900±0.093	0.39	1223
	PCA	**0.856±0.091**	**0.930±0.049**	**0.781±0.207**	**0.931±0.049**	**0.82**	**26**
T2	None	0.717±0.79	0.832±0.0101	0.525±0.236	0.971±0.039	0.58	1223
	Mutual Info	0.739±0.82	0.858±0.092	0.524±0.232	0.974±0.037	0.81	699

Logistic Regression

Table 6.21: Best Performances of LogR model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.766±0.122	0.849±0.124	0.585±0.267	0.947±0.058	0.48	1223
	Mutual Info	0.776±0.135	0.862±0.108	0.601±0.293	0.952±0.056	0.36	809
T1	None	0.757±0.089	0.874±0.109	0.566±0.215	0.947±0.055	0.35	1223
	Mutual Info	0.776±0.095	0.874±0.095	0.731±0.243	0.821±0.170	1.0	757
T1GD	None	0.857±0.100	0.935±0.052	0.772±0.216	0.941±0.044	0.87	1223
	Mutual Info	**0.853±0.095**	**0.937±0.051**	**0.764±0.209**	**0.943±0.044**	**0.35**	**651**
T2	None	0.788±0.082	0.877±0.081	0.687±0.205	0.890±0.088	0.76	1223
	Mutual Info	0.802±0.074	0.890±0.072	0.730±0.194	0.873±0.102	1.0	952

Naïve Bayes

Table 6.22: Best Performances of NB model Across different MRI Sequences and Feature Selection Methods for Combined Dataset. The best combination is highlighted in bold.

MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity	SMOTE Ratio	K features
FLAIR	None	0.673±0.122	0.743±0.130	0.719±0.179	0.626±0.254	0.36	1223
	PCA	0.686±0.120	0.831±0.104	0.431±0.244	0.940±0.063	0.40	34
T1	None	0.723±0.071	0.730±0.118	0.545±0.135	0.901±0.069	0.46	1223
	PCA	0.730±0.080	0.816±0.101	0.770±0.193	0.690±0.175	0.89	50
T1GD	None	0.828±0.082	0.885±0.067	0.790±0.179	0.866±0.066	0.34	1223
	Mutual Info	**0.836±0.070**	**0.917±0.062**	**0.875±0.155**	**0.797±0.063**	**0.39**	**38**
T2	None	0.763±0.085	0.768±0.116	0.617±0.195	0.910±0.083	1.0	1223
	PCA	0.732±0.103	0.841±0.104	0.545±0.216	0.919±0.063	0.55	44

Comparison Between Models

The combination of the UCSF-PDGM and UPENN-GBM datasets resulted in a larger sample size, which significantly enhanced the performance of the classifiers. As shown in Table 6.23, this dataset combination yielded the best results, with the highest balanced accuracy of 0.856 ± 0.091 achieved by the Random Forest classifier. This was obtained using PCA with 26 features and a SMOTE ratio of 0.82, which translates to an addition of 488 synthetic positive samples during training. The highest AUC of 0.946 ± 0.048 was achieved by the XGBoost classifier, also utilizing PCA with 39 features and a SMOTE ratio of 0.36, resulting in 162 new synthetic positive samples in the training set. The Random Forest model took 2.546 ± 0.173 seconds to run, while the XGBoost model took 0.736 ± 0.019 seconds.

The combined dataset also exhibited more stable performance metrics, as indicated by lower standard deviations in the results. The balanced accuracy and AUC values showed smoother fluctuations, suggesting that the classifiers performed more consistently across the 10 different runs. This enhanced stability can be attributed to the increased sample size, which provides a more comprehensive representation of the data, thereby reducing the impact of anomalies and variations in individual folds.

Consistent with previous findings, the T1GD MRI sequence continued to deliver the best results across all classifiers, reinforcing its importance in accurately determining IDH mutation status.

Table 6.23: Comparison of best-performing classifiers for Combined Dataset.

Classifier	MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
SVM	T1GD	None	0.848±0.108	0.932±0.059	0.824±0.222	0.871±0.061
KNN	T1GD	PCA	0.825±0.091	0.915±0.077	0.739±0.208	0.910±0.049
XGBoost	T1GD	PCA	0.835±0.098	0.946±0.048	0.705±0.216	0.966±0.037
RF	T1GD	PCA	0.856±0.091	0.930±0.049	0.781±0.207	0.931±0.049
LogR	T1GD	Mutual Info	0.853±0.095	0.937±0.051	0.764±0.209	0.943±0.044
NB	T1GD	Mutual Info	0.836±0.070	0.917±0.062	0.875±0.155	0.797±0.063

6.3.1 Best Hyperparameters

This subsection presents the best hyperparameters found for each of the best-performing classifiers used in the experiments (Table 6.24).

Table 6.24: Best Hyperparameters found for each of the Best-Performing Classifiers for Combined Dataset

Classifier	Hyperparameter	Value
SVM	C	0.016
	kernel	sigmoid
	gamma	auto
	class_weight	balanced
KNN	n_neighbors	28
	weights	distance
	metric	manhattan
	p	5
XGBoost	learning_rate	0.01
	max_depth	3
	n_estimators	1000
	gamma	0.0
	colsample_bytree	1.0
	min_child_weight	1
	subsample	0.641
	class_weight	balanced
RF	n_estimator	429
	max_depth	165
	criterion	entropy
	min_sample_split	12
	max_sample_leaf	9
	max_features	sqrt
	class_weight	balanced
LogR	C	0.107
	penalty	l1
	max_iter	100
	l1_ratio	None
	class_weight	balanced
NB	var_smoothing	1.0

6.3.2 Limitations

One of the main limitations of this study is the absence of deep learning techniques in our comparative analysis. While traditional machine learning methods were extensively explored, deep learning models, which have shown considerable success in medical image analysis due to their ability to automatically extract and learn hierarchical features, were not employed. The inclusion of DL methods could provide additional insights and potentially improve predictive performance. Future research should include a comparative analysis involving state-of-the-art deep learning architectures, such as CNNs, to evaluate their efficacy in predicting IDH mutation status.

In this study, both public datasets (UCSF-PDGM and UPENN-GBM) were combined to form a single, larger dataset for training and testing purposes. Although this approach was chosen to enhance the training process by increasing the number of available samples, it presents a limitation in terms of generalizability. We did not experiment with training the model on one dataset and testing it on the other, which could have provided a more robust evaluation of the model's performance across different patient populations and imaging protocols. Future studies should explore this approach to better understand the model's generalization capabilities.

While multiple MRI sequences were used in this study, not all possible sequences were included. Additional sequences, such as Diffusion-Weighted Imaging and Perfusion-Weighted Imaging, were not explored. These sequences could potentially provide valuable information and improve the accuracy of IDH mutation predictions.

In summary, while the study presents promising results, these limitations highlight the need for further research to enhance the robustness and generalizability of the findings. Addressing these limitations in future work will contribute to the development of more reliable and effective predictive models for IDH mutation status in glioblastoma multiforme.

6.4 Summary

This chapter presented the results and discussions of the study on predicting IDH mutation status in gliomas using machine learning algorithms applied to MRI data. The analysis was conducted across three datasets: UCSF-PDGM, UPENN-GBM, and a combined dataset of the two.

For the UCSF-PDGM dataset, the best AUC of 0.776±0.186 was achieved by the RF classifier using the T1GD MRI sequence with Boruta. This combination yielded a balanced accuracy of 0.762 ± 0.166 with a SMOTE ratio of 0.49 and retained all 19 features.

In the UPENN-GBM dataset, the highest AUC of 0.850 ± 0.174 was obtained by the SVM classifier using the T1GD MRI sequence without feature selection. This model configuration provided a balanced accuracy of 0.816 ± 0.161, with a SMOTE ratio of 0.24 and 1223 features.

The combined dataset demonstrated the best overall performance, with the XGBoost classifier achieving an AUC of 0.946 ± 0.048 using the T1GD MRI sequence and PCA for feature selection. This combination resulted in a balanced accuracy of 0.835 ± 0.098, with a SMOTE ratio of 0.36 and 39 features retained.

The main limitations encountered included dataset imbalances, potential biases due to prior biopsies, the specificity of the datasets to certain glioma grades, and the exclusion of deep learning techniques. Addressing these limitations in future research could enhance the robustness and generalizability of predictive models for IDH mutation status in gliomas.

Chapter 7

Conclusion

This study focused on developing an artificial intelligence system capable of predicting IDH mutation status in glioblastoma multiforme using machine learning techniques applied to magnetic resonance imaging analysis. Various state-of-the-art methods were studied and implemented aiming to achieve accurate and non-invasive predictions.

Two public datasets from The Cancer Imaging Archive (TCIA), namely UCSF-PDGM and UPENN-GBM, were utilized for this study, containing 493 and 514 usable patient cases, respectively. One of the major challenges encountered was class imbalance, as the number of IDH-positive cases was significantly lower compared to IDH-negative cases. This imbalance posed a considerable obstacle in training the machine learning models and necessitated data augmentation techniques to mitigate its effects.

To tackle the problem, a comprehensive radiogenomic approach was employed, leveraging the extraction of radiomic features from MRI scans and integrating them with machine learning algorithms. Radiomic features capture the quantitative information from medical images, which reflect the tumor's phenotypic characteristics. This data was then fed into various machine learning classifiers, including Random Forest, Support Vector Machine, eXtreme Gradient Boosting, Logistic Regression, Naïve Bayes, and K-Nearest Neighbors. This approach aimed to explore the predictive power of these features and identify the most effective combination of MRI sequences and classifiers for accurately predicting IDH mutation status.

Despite these challenges, the study achieved highly satisfactory results in predicting IDH mutations, with a maximum AUC of 0.946 ± 0.048. However, it is important to note that these promising results do not necessarily indicate good performance with external samples. The limited dataset size, particularly the low number of IDH-positive cases, may affect the model's generalizability and performance on independent datasets.

In conclusion, this work demonstrates the significant potential of integrating machine learning and radiogenomics in predicting genetic mutations in GBM. The successful application of these techniques suggests a promising avenue for non-invasive diagnostics and personalized treatment planning. For future work, it is recommended to experiment with different MRI sequences, such as Diffusion-Weighted Imaging and Perfusion-Weighted Imaging, to evaluate whether the inclusion

of these modalities can further improve the accuracy and robustness of the developed models. Additionally, incorporating larger and more diverse datasets could enhance the generalizability of the findings, ultimately contributing to better clinical outcomes for GBM patients.

References

[1] D. Alis, M. H. Yilmaz, S. B. Gorkem, B. Kocak, M. D. Bulut, M. S. Uluer, and T. Yurtseven. Machine learning-based quantitative texture analysis of conventional mri combined with adc maps for assessment of idh1 mutation in high-grade gliomas. *Japanese Journal of Radiology*, 38(2):135–143, Nov. 18 2019.

[2] Jonys Arcanjo. How to select the best variables for your model with boruta. https://medium.com/jonys-arcanjo/how-to-select-the-best-variables-for-your-model-with-boruta-53df05d74b78, 2024. Accessed: 10 Jul 2024.

[3] M. Avanzo, J. Stancanello, and I. El Naqa. Machine and deep learning methods for radiomics. *Medical Physics*, 47(5), May 2020.

[4] S. Bakas et al. 'multi-parametric magnetic resonance imaging (mpmri) scans for de novo glioblastoma (gbm) patients from the university of pennsylvania health system (upenn-gbm)', 2021. doi: 10.7937/TCIA.709X-DN49.

[5] N. Beig, K. Bera, and P. Tiwari. Introduction to radiomics and radiogenomics in neuro-oncology: implications and challenges, Dec. 01 2020.

[6] A. Bettinelli, L. Cozzi, G. Veronesi, G. Petralia, M. Caballo, C. Resteghini, S. Raimondi, E. Voulaz, C. N. De Cecco, M. Bellomi, and P. Summers. A novel benchmarking approach to assess the agreement among radiomic tools. *Radiology*, 303(3):533–541, Jun 2022.

[7] W. L. et al. Bi. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*, 69(2):127–157, Feb 2019.

[8] S. Bisdas, B. Baessler, A. M. Wang, and S. Thust. Texture analysis- and support vector machine-assisted diffusional kurtosis imaging may allow in vivo gliomas grading and idh-mutation status prediction: a preliminary study. *Scientific Reports*, 8(1):24438, Apr. 17 2018.

[9] T. Bosquez. Neuroimaging: Three brain imaging techniques, 2022. [Online] Available at https://blogs.iu.edu/sciu/2022/02/05/three-brain-imaging-techniques/, last accessed in June 2024.

[10] JW Britton, LC Frey, JL Hopp, et al. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. American Epilepsy Society, Chicago, 2016. Appendix 6. A Brief History of EEG.

[11] J. Brownlee. Smote for imbalanced classification with python, 2021. [Online] Available at https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/, last accessed in June 2024.

[12] Y. H. Byun and C.-K. Park. Classification and diagnosis of adult glioma: A scoping review. *Brain & Neurorehabilitation*, 15(3), 2022.

[13] E. Calabrese et al. The university of california san francisco preoperative diffuse glioma mri (ucsf-pdgm). *The Cancer Imaging Archive*, 2022. doi: 10.7937/TCIA.BDGF-8V37.

[14] E. Calabrese, J. E. Villanueva-Meyer, and S. Cha. A fully automated artificial intelligence method for non-invasive, imaging-based identification of genetic alterations in glioblastomas. *Scientific Reports*, 10(1), Jul. 16 2020.

[15] G. Carpenter and J. Jones. Ct head. https://radiopaedia.org/articles/ct-head, Mar. 13 2010.

[16] Center for Biomedical Image Computing and Analytics. Cancer imaging phenomics toolkit (captk). https://www.med.upenn.edu/cbica/captk/, 2024. Accessed: 10 Jul 2024.

[17] A. Chaddad, C. Desrosiers, and T. Niazi. Radiomics in glioblastoma: Current status and challenges facing clinical implementation. *Frontiers in Oncology*, 9:374, May 21 2019.

[18] Y. S. Choi, S. Bae, J. H. Chang, S. G. Kang, S. T. Kim, J. H. Kim, I. S. Hamm, and S. K. Lee. Fully automated hybrid approach to predict the idh mutation status of gliomas via deep learning and radiomics. *Neuro-Oncology*, 23(2):304–313, Jul. 24 2020.

[19] Mayo Clinic. Glioma. Mayo Clinic. [Online] Available at https://www.mayoclinic.org/diseases-conditions/glioma/symptoms-causes/syc-20350251, last accessed in June 2024.

[20] Shanghai United Imaging Healthcare Co. Article 2, 2023. [Online] Available at https://eu.united-imaging.com/en/news-center/article/article-2, last accessed in June 2024.

[21] Pyradiomics community. Radiomics documentation: getwaveletimage, 2016. [Online] Available at https://pyradiomics.readthedocs.io/en/latest/radiomics.html#radiomics.imageoperations.getWaveletImage, last accessed in June 2024.

[22] C. Davatzikos, S. Bakas, H. Akbari, M. Bilello, M. Boccardi, F. Colleoni, A. Crimi, J. Lee, T. Melcher, and R. T. et al. Shinohara. Ai-based prognostic imaging biomarkers for precision neuro-oncology: the respond consortium. *Neuro-Oncology*, 22(6):886–888, Mar. 10 2020.

[23] C. Davatzikos, S. Rathore, S. Bakas, S. Pati, M. Bergman, R. Kalarot, P. Sridharan, A. Gastounioti, N. Jahani, M. Bilello, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *Journal of Medical Imaging*, 5(01):1, Jan. 11 2018.

[24] XGBoost developers. Xgboost documentation, 2022. [Online] Available at https://xgboost.readthedocs.io/en/stable/, last accessed in June 2024.

[25] T. H. Dou, T. P. Coroller, J. J. M. van Griethuysen, R. H. Mak, and H. J. W. L. Aerts. Peritumoral radiomics features predict distant metastasis in locally advanced nsclc. *PLOS ONE*, 13(11):e0206108, Nov. 02 2018.

[26] D. Du, Z. Fang, D. Wang, W. Zhao, C. Yan, P. Manickam, J. Tian, and C. Li. Machine learning methods for optimal radiomics-based differentiation between recurrence and inflammation: Application to nasopharyngeal carcinoma post-therapy pet/ct images. *Molecular Imaging and Biology*, 22(3):730–738, Jul. 23 2019.

[27] GeeksforGeeks. Difference between pca and t-sne, 2023. [Online] Available at https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/, last accessed in June 2024.

[28] GeeksforGeeks. Feature selection techniques in machine learning, 2024. [Online] Available at https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/, last accessed in June 2024.

[29] GeeksforGeeks. Random forest algorithm in machine learning. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/, 2024. Accessed: 10 Jul 2024.

[30] Y. Glick and F. Gaillard. Isocitrate dehydrogenase. https://radiopaedia.org/articles/isocitrate-dehydrogenase, Aug. 22 2015.

[31] J. GRILL. Orphanet: Glioblastoma, 2007. Orphanet. [Online] Available at https://www.orpha.net/en/disease/detail/360?name=Glioblastoma%20multiforme&mode=name, last accessed in may 2024.

[32] M. Götz, M. Nolden, and K. Maier-Hein. Mitk phenotyping: An open-source toolchain for image-based personalized medicine with radiomics. *Radiotherapy and Oncology*, 131:108–111, Feb 2019.

[33] C. Hartmann, B. Hentschel, W. Wick, D. Capper, J. Felsberg, M. Simon, M. Westphal, G. Schackert, R. Meyermann, and T. et al. Pietsch. Patients with idh1 wild type anaplastic astrocytomas exhibit worse prognosis than idh1-mutated glioblastomas, and idh1 mutation status accounts for the unfavorable prognostic effect of higher age: implications for classification of gliomas. *Acta Neuropathologica*, 120(6):707–718, Nov 2010.

[34] C. Hartmann, B. Hentschel, W. Wick, D. Capper, J. Felsberg, M. Simon, M. Westphal, G. Schackert, R. Meyermann, and T. et al. Pietsch. Patients with idh1 wild type anaplastic astrocytomas exhibit worse prognosis than idh1-mutated glioblastomas, and idh1 mutation status accounts for the unfavorable prognostic effect of higher age: implications for classification of gliomas. *Acta Neuropathologica*, 120(6):707–718, Nov. 19 2010.

[35] T. Hines. Brain anatomy, anatomy of the human brain. Mayfield Brain Spine Cincinnati, Ohio. [Online] Available at https://mayfieldclinic.com/pe-anatbrain.htm, last accessed in June 2024.

[36] M. Häggström. Mri sequences. https://radlines.org/MRI_sequences, 2024. Accessed on 10 Jul 2024.

[37] IBM. K-nearest neighbors (knn). [Online] Available at https://www.ibm.com/topics/knn, last accessed in June 2024.

[38] IBM. Logistic regression. [Online] Available at https://www.ibm.com/topics/logistic-regression, last accessed in June 2024.

[39] IBM. Naive bayes. [Online] Available at https://www.ibm.com/topics/naive-bayes, last accessed in June 2024.

[40] IBM. Random forest. [Online] Available at https://www.ibm.com/topics/random-forest, last accessed in June 2024.

[41] IBM. Support vector machine (svm), 2023. [Online] Available at https://www.ibm.com/topics/support-vector-machine, last accessed in June 2024.

[42] Image Biomarker Standardisation Initiative. Ibsi: Image features. https://ibsi.readthedocs.io/en/latest/03_Image_features.html, 2021. Accessed: 2024-07-06.

[43] V. Prabhu J. Thakkar, P. Peruzzi and American Association of Neurological Surgeons. Glioblastoma multiforme – symptoms, diagnosis and treatment options. [Online] Available at https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Glioblastoma-Multiforme, last accessed in may 2024.

[44] V. Kanade. What is logistic regression?, 2022. [Online] Available at https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/, last accessed in June 2024.

[45] A. Fathi Kazerooni, M. Bax, G. S. Young, and W. Choy. Applications of radiomics and radiogenomics in high-grade gliomas in the era of precision medicine. *Cancers*, 13(23):5921, Nov. 25 2021.

[46] A. Kim and I. Jung. Optimal selection of resampling methods for imbalanced data with high complexity. *PLOS ONE*, 18(7):e0288540, 2023.

[47] B. Kocak, E. A. Kus, A. E. Kavur, and O. Kilickesmez. Radiomics with artificial intelligence: a practical guide for beginners. *Diagnostic and Interventional Radiology*, 25(6):485–495, Nov. 05 2019.

[48] N. Q. K. Le, T. N. K. Hung, D. T. Do, L. H. T. Lam, L. H. Dang, and T.-T. Huynh. Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from mri. *Computers in Biology and Medicine*, 132:104320, May 2021.

[49] Scikit learn developers. sklearn.decomposition.pca — scikit-learn 0.24.1 documentation. [Online] Available at https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html, last accessed in June 2024.

[50] Y. Lei, S. Tian, H. Zhu, S. Vijayakumar, W. J. Curran, and T. Liu. Benchmarking various radiomic toolkit features while applying the image biomarker standardization initiative toward clinical translation of radiomic analysis. *Journal of Digital Imaging*, 34(5):1156–1170, Sep. 20 2021.

[51] Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao. Deep learning based radiomics (dlr) and its usage in noninvasive idh1 prediction for low grade glioma. *Scientific Reports*, 7(1), Jul 2017.

[52] Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao. Multiregional radiomics profiling from multiparametric mri: Identifying an imaging predictor of idh1 mutation status in glioblastoma. *Cancer Medicine*, 7(12):5999–6009, Nov. 13 2018.

[53] Z. et al. Li. Multiregional radiomics profiling from multiparametric mri: Identifying an imaging predictor of idh1 mutation status in glioblastoma. *Cancer Medicine*, 7(12):5999–6009, Nov 2018.

[54] M. Alex Lobera. Glioblastoma (multiforme) imaging. Practice Essentials, Computed Tomography, Magnetic Resonance Imaging. [Online] Available at https://emedicine.medscape.com/article/340870-overview?form=fpf, last accessed in June 2024.

[55] E. Lotan, R. Jain, N. Razavian, G. M. Fatterpekar, and Y. W. Lui. State of the art: Machine learning applications in glioma imaging. *American Journal of Roentgenology*, 212(1):26–37, Jan 2019.

[56] P. Martin-Gonzalez, A. Reyes-Palomares, B. Kostov, S. Aguirre, C. Jimenez-Luna, J. Bandres-Meriz, S. Ortega, M. J. Lecumberri, M. Villalba, and M. et al. Calvo. Integrative radiogenomics for virtual biopsy and treatment monitoring in ovarian cancer. *Insights into Imaging*, 11(1):1–15, Aug. 17 2020.

[57] J. R. McFaline-Figueroa and E. Q. Lee. Brain tumors. *The American Journal of Medicine*, 131(8):874–882, Aug 2018.

[58] C. C. medical professional. Cerebellum: What it is, function anatomy. Cleveland Clinic. [Online] Available at https://my.clevelandclinic.org/health/body/23418-cerebellum#overview, last accessed in June 2024.

[59] R. Michael. Optimizing hyperparameters: The right way, 2020. [Online] Available at https://towardsdatascience.com/optimizing-hyperparameters-the-right-way-3c9cafc279cc, last accessed in June 2024.

[60] C. F. Muccio, G. Esposito, A. Bartolini, L. Brunese, F. Cademartiri, and C. Andreula. Magnetic resonance features of pyogenic brain abscesses and differential diagnosis using morphological and functional imaging studies: A pictorial essay. *Journal of Neuroradiology*, 41(3):153–167, Jul 2014.

[61] V. Nardone, G. Battaglia, R. M. Angelillo, V. Ravo, G. Vacca, P. Tini, R. Corvò, and G. Sanguineti. Texture analysis as a predictor of radiation-induced xerostomia in head and neck patients undergoing imrt. *La radiologia medica*, 123(6):415–423, Jan. 24 2018.

[62] Qt Documentation. *Qt Designer Manual*, 2024. Accessed: 10 Jul 2024.

[63] Sruthi E R. Understanding random forest, 2024. [Online] Available at https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/, last accessed in June 2024.

[64] Sachinsoni. Mastering t-sne: t-distributed stochastic neighbor embedding. https://medium.com/@sachinsoni600517/mastering-t-sne-t-distributed-stochastic-neighbor-embedding-0e365ee898ea, 2024. Accessed: 10 Jul 2024.

[65] Y. Sakai et al. Mri radiomic features to predict idh1 mutation status in gliomas: A machine learning approach using gradient tree boosting. *International Journal of Molecular Sciences*, 21(21):8004, Oct 2020.

[66] A. Sanjyal. Dimensionality reduction vs feature selection, 2022. [Online] Available at https://medium.com/@asanjyal81/dimensionality-reduction-vs-feature-selection-e68f91aa8724, last accessed in June 2024.

[67] Saurabh. t-sne algorithm implementation in r and python, 2024. [Online] Available at https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/, last accessed in June 2024.

[68] A. Saxena, I. Ghosh, A. Gaurav, R. Srivastava, and A. Pandey. Role of artificial intelligence in radiogenomics for cancers in the era of precision medicine. *Cancers*, 14(12):2860, Jun. 09 2022.

[69] Sanjay Saxena, Biswajit Jena, Neha Gupta, Suchismita Das, Deepaneeta Sarmah, Pallab Bhattacharya, Tanmay Nath, Sudip Paul, Mostafa M. Fouda, Manudeep Kalra, Luca Saba, Gyan Pareek, and Jasjit S. Suri. Role of artificial intelligence in radiogenomics for cancers in the era of precision medicine. *Cancers*, 14(12), 2022.

[70] ScienceDirect. Logistic regression. [Online] Available at https://www.sciencedirect.com/topics/computer-science/logistic-regression, last accessed in June 2024.

[71] scikit-learn. sklearn feature selection rfe. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html, 2024. Accessed: 10 Jul 2024.

[72] Deepak Sethi. From data to insights: How mutual information revolutionizes feature engineering. https://medium.com/@deepaksethionly/from-data-to-insights-how-mutual-information-revolutionizes-feature-enginee 2024. Accessed: 10 Jul 2024.

[73] T. Srivastava. Introduction to k-nearest neighbors algorithm: Clustering, 2024. [Online] Available at https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/, last accessed in June 2024.

[74] R. Su, X. Liu, Q. Jin, X. Liu, and L. Wei. Identification of glioblastoma molecular subtype and prognosis based on deep mri features. *Knowledge-Based Systems*, 232:107490, Nov 2021.

[75] 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: A Cancer Journal for Clinicians*, 70(4):299–312, Jun 2020.

[76] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, Sep 2008.

[77] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillon-Robin, S. Pieper, and H. J. W. L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017.

[78] Analytics Vidhya. An end-to-end guide to understand the math behind xgboost, 2024. [Online] Available at https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/?utm_source=blog&utm_medium=4-boosting-algorithms-machine-learning, last accessed in June 2024.

[79] M. et al. Weller. Eano guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. *Nature Reviews Clinical Oncology*, 18(3):170–186, Dec 2020.

[80] S. Wu, J. Meng, Q. Yu, P. Li, and S. Fu. Radiomics-based machine learning methods for isocitrate dehydrogenase genotype prediction of diffuse gliomas. *Journal of Cancer Research and Clinical Oncology*, 145(3):543–550, Feb 2019.

[81] H. Yan, D. W. Parsons, G. Jin, R. McLendon, B. A. Rasheed, W. Yuan, I. Kos, I. Batinic-Haberle, S. Jones, and G. J. et al. Riggins. Idh1 and idh2 mutations in gliomas. *New England Journal of Medicine*, 360(8):765–773, Feb. 19 2009.

[82] X. et al. Zhang. Radiomics strategy for molecular subtype stratification of lower-grade glioma: Detecting idh and tp53 mutations based on multimodal mri. *Journal of Magnetic Resonance Imaging*, 48(4):916–926, Feb 2018.

[83] A. Zwanenburg et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, May 2020.

Appendix A

Detailed Experimental Results

This appendix provides a detailed overview of all the experimental results, including additional metrics and model configurations that were explored during the study. The results presented here supplement the highlights discussed in Chapterr 6, providing a comprehensive understanding of the performance and behavior of different machine learning models and configurations used to predict IDH mutation status in gliomas.

A.1 Dataset 1: UCSF-PDGM

Support Vector Machine

Table A.1: Performance of SVM Model for UCSD-PDGM.

ML Model	MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
SVM	FLAIR	None	0.731±0.145	0.751±0.170	0.620±0.286	0.841±0.110
		Mutual Info	0.700±0.141	0.739±0.165	0.520±0.289	0.880±0.123
		Boruta	0.641±0.105	0.657±0.135	0.460±0.229	0.821±0.063
		PCA	0.665±0.117	0.658±0.171	0.470±0.249	0.860±0.076
		LDA	0.627±0.140	0.642±0.134	0.560±0.276	0.695±0.079
	T1	None	0.709±0.151.	0.746±0.172	0.550±0.031	0.869±0.134
		Mutual Info	0.659±0.124	0.736±0.160	0.380±0.271	0.939±0.092
		Boruta	0.686±0.088	0.718±0.148	0.280±0.218	0.892±0.118
		PCA	0.631±0.097	0.685±0.160	0.330±0.215	0.931±0.080
		LDA	0.644±0.087	0.647±0.072	0.560±0.211	0.728±0.078
	T1GD	None	0.748±0.141	0.744±0.181	0.700±0.300	0.796±0.118
		Mutual Info	0.500±0.000	0.747±0.195	0.000±0.000	1.000±0.000
		Boruta	0.728±0.145	0.747±0.152	0.540±0.294	0.916±0.100
		PCA	0.501±0.023	0.267±0.150	0.700±0.438	0.303±0.457
		LDA	0.558±0.099	0.575±0.153	0.320±0.209	0.797±0.060
	T2	None	0.745±0.159	0.757±0.178	0.570±0.307	0.921±0.097
		Mutual Info	0.500±0.000	0.745±0.194	0.000±0.000	1.000±0.000
		Boruta	0.685±0.091	0.737±0.176	0.300±0.458	0.700±0.458
		PCA	0.500±0.000	0.418±0.171	0.300±0.458	0.700±0.458
		LDA	0.575±0.083	0.580±0.122	0.420±0.147	0.730±0.096

Table A.2: SVM SMOTE ratio and K features for UCSD-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
SVM	FLAIR	None	0.35	1223
		Mutual Info	0.64	854
		Boruta	0.73	21
		PCA	0.55	44
		LDA	0.62	1
	T1	None	0.35	1223
		Mutual Info	0.35	1086
		Boruta	0.41	16
		PCA	0.35	37
		LDA	0.62	1
	T1GD	None	1.00	1223
		Mutual Info	0.35	1223
		Boruta	0.55	19
		PCA	0.35	24
		LDA	0.62	1
	T2	None	0.54	1223
		Mutual Info	0.35	518
		Boruta	0.38	21
		PCA	0.35	19
		LDA	0.62	1

XGBoost

Table A.3: Performance of XGBoost Model for UCSD-PDGM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
XGB	FLAIR	None	0.684±0.122	0.749±0.145	0.470±0.245	0.897±0.111
		Mutual Info	0.705±0.147	0.723±0.195	0.500±0.293	0.910±0.103
		Boruta	0.706±0.127	0.744±0.163	0.530±0.265	0.882±0.102
		PCA	0.707±0.149	0.749±0.143	0.550±0.304	0.864±0.100
		LDA	0.633±0.116	0.650±0.153	0.520±0.248	0.746±0.082
	T1	None	0.529±0.033	0.722±0.175	0.090±0.070	0.969±0.056
		Mutual Info	0.647±0.107	0.679±0.173	0.370±0.224	0.923±0.087
		Boruta	0.700±0.139	0.717±0.191	0.530±0.287	0.870±0.094
		PCA	0.639±0.096	0.681±0.159	0.380±0.204	0.898±0.054
		LDA	0.628±0.088	0.664±0.066	0.490±0.197	0.766±0.070
	T1GD	None	0.500±0.000	0.785±0.172	0.000±0.000	1.000±0.000
		Mutual Info	0.752±0.160	0.776±0.190	0.630±0.320	0.875±0.111
		Boruta	0.655±0.113	0.725±0.150	0.440±0.211	0.870±0.100
		PCA	0.628±0.101	0.772±0.167	0.310±0.212	0.946±0.066
		LDA	0.605±0.094	0.597±0.116	0.410±0.207	0.799±0.060
	T2	None	0.662±0.113	0.728±0.185	0.380±0.248	0.944±0.080
		Mutual Info	0.600±0.086	0.731±0.168	0.230±0.210	0.969±0.069
		Boruta	0.706±0.128	0.746±0.191	0.550±0.262	0.862±0.091
		PCA	0.678±0.137	0.738±0.159	0.520±0.260	0.836±0.093
		LDA	0.583±0.082	0.607±0.105	0.420±0.209	0.745±0.104

Table A.4: XGBoost SMOTE ratio and K features for UCSD-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
XGB	FLAIR	None	0.	1223
		Mutual Info	1.00	896
		Boruta	1.00	7
		PCA	0.74	15
		LDA	0.35	1
	T1	None	0.26	1223
		Mutual Info	0.35	848
		Boruta	0.93	4
		PCA	0.73	43
		LDA	0.47	1
	T1GD	None	0.26	1223
		Mutual Info	0.27	68
		Boruta	0.42	3
		PCA	0.56	36
		LDA	0.35	1
	T2	None	0.26	1223
		Mutual Info	0.24	513
		Boruta	1.00	8
		PCA	1.00	35
		LDA	0.62	1

Random Forest

Table A.5: Performance of RF Model for UCSD-PDGM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
RF	FLAIR	None	0.691±0.148	0.757±0.176	0.490±0.277	0.893±0.101
		Mutual Info	0.687±0.146	0.756±0.186	0.460±0.276	0.913±0.103
		Boruta	0.638±0.084	0.756±0.170	0.320±0.209	0.956±0.075
		PCA	0.588±0.114	0.724±0.186	0.230±0.245	0.946±0.077
		LDA	0.640±0.129	0.647±0.148	0.530±0.257	0.751±0.064
	T1	None	0.630±0.091	0.726±0.188	0.300±0.195	0.959±0.075
		Mutual Info	0.657±0.097	0.731±0.188	0.360±0.206	0.954±0.084
		Boruta	0.643±0.097	0.732±0.188	0.330±0.215	0.956±0.080
		PCA	0.628±0.079	0.740±0.153	0.310±0.151	0.946±0.049
		LDA	0.618±0.081	0.651±0.083	0.470±0.185	0.766±0.064
	T1GD	None	0.685±0.130	0.775±0.161	0.470±0.272	0.900±0.093
		Mutual Info	0.724±0.139	0.795±0.155	0.530±0.290	0.918±0.097
		Boruta	0.762±0.166	0.776±0.186	0.660±0.326	0.864±0.115
		PCA	0.633±0.079	0.764±0.179	0.310±0.192	0.956±0.071
		LDA	0.575±0.101	0.577±0.113	0.390±0.197	0.761±0.059
	T2	None	0.690±0.123	0.771±0.189	0.460±0.262	0.921±0.073
		Mutual Info	0.691±0.140	0.756±0.153	0.490±0.270	0.893±0.076
		Boruta	0.628±0.077	0.754±0.202	0.310±0.176	0.946±0.072
		PCA	0.672±0.158	0.730±0.194	0.510±0.281	0.834±0.106
		LDA	0.583±0.090	0.599±0.100	0.420±0.204	0.745±0.088

Table A.6: RF SMOTE ratio and K features for UCSD-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
RF	FLAIR	None	1.0	1223
		Mutual Info	0.64	419
		Boruta	0.35	21
		PCA	0.59	50
		LDA	0.35	1
	T1	None	0.35	1223
		Mutual Info	0.21	863
		Boruta	0.35	16
		PCA	0.36	38
		LDA	0.44	1
	T1GD	None	0.62	1223
		Mutual Info	0.47	1215
		Boruta	0.49	19
		PCA	0.35	48
		LDA	0.70	1
	T2	None	0.58	1223
		Mutual Info	0.80	699
		Boruta	0.35	21
		PCA	0.99	24
		LDA	0.83	1

Logistic Regression

Table A.7: Performance of LogR Model for UCSD-PDGM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
LogR	FLAIR	None	0.731±0.147	0.732±0.200	0.560±0.332	0.903±0.099
		Mutual Info	0.746±0.144	0.733±0.196	0.660±0.301	0.831±0.112
		Boruta	0.672±0.096	0.696±0.160	0.480±0.209	0.865±0.079
		PCA	0.748±0.158	0.758±0.183	0.670±0.310	0.826±0.130
		LDA	0.629±0.128	0.636±0.167	0.550±0.262	0.707±0.077
	T1	None	0.728±0.138	0.757±0.177	0.660±0.291	0.796±0.115
		Mutual Info	0.683±0.095	0.742±0.148	0.460±0.180	0.906±0.102
		Boruta	0.726±0.147	0.731±0.171	0.620±0.309	0.831±0.127
		PCA	0.614±0.080	0.667±0.131	0.310±0.181	0.918±0.085
		LDA	0.625±0.083	0.650±0.099	0.500±0.179	0.751±0.062
	T1GD	None	0.747±0.178	0.770±0.189	0.660±0.326	0.833±0.109
		Mutual Info	0.733±0.155	0.770±0.174	0.670±0.323	0.796±0.120
		Boruta	0.760±0.146	0.767±0.180	0.660±0.317	0.859±0.122
		PCA	0.719±0.134	0.773±0.192	0.520±0.296	0.918±0.089
		LDA	0.579±0.098	0.592±0.143	0.390±0.197	0.769±0.061
	T2	None	0.754±0.163	0.741±0.211	0.670±0.335	0.838±0.108
		Mutual Info	0.690±0.126	0.719±0.184	0.470±0.272	0.910±0.093
		Boruta	0.665±0.121	0.675±0.167	0.510±0.305	0.819±0.088
		PCA	0.703±0.118	0.764±0.175	0.460±0.262	0.946±0.071
		LDA	0.566±0.074	0.591±0.134	0.410±0.192	0.722±0.098

Table A.8: LogR SMOTE ratio and K features for UCSD-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
LogR	FLAIR	None	0.48	1223
		Mutual Info	0.72	392
		Boruta	0.35	21
		PCA	1.00	38
		LDA	0.78	1
	T1	None	0.35	1223
		Mutual Info	0.35	669
		Boruta	0.35	16
		PCA	0.39	43
		LDA	0.79	1
	T1GD	None	0.87	1223
		Mutual Info	0.35	10
		Boruta	0.35	19
		PCA	0.35	29
		LDA	0.79	1
	T2	None	0.76	1223
		Mutual Info	0.41	383
		Boruta	0.35	21
		PCA	0.35	19
		LDA	0.79	1

Naïve Bayes

Table A.9: Performance of NB Model for UCSD-PDGM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
NB	FLAIR	None	0.649±0.137	0.686±0.167	0.650±0.220	0.649±0.137
		Mutual Info	0.662±0.136	0.687±0.156	0.510±0.270	0.814±0.097
		Boruta	0.609±0.109	0.652±0.145	0.720±0.172	0.498±0.166
		PCA	0.652±0.150	0.751±0.173	0.370±0.293	0.933±0.068
		LDA	0.616±0.133	0.659±0.182	0.400±0.257	0.832±0.064
	T1	None	0.622±0.119	0.674±0.169	0.390±0.238	0.854±0.106
		Mutual Info	0.623±0.109	0.677±0.171	0.350±0.250	0.895±0.084
		Boruta	0.630±0.135	0.660±0.183	0.530±0.237	0.770±0.160
		PCA	0.562±0.068	0.716±0.136	0.160±0.136	0.964±0.037
		LDA	0.638±0.093	0.661±0.101	0.510±0.197	0.766±0.078
	T1GD	None	0.754±0.153	0.748±0.175	0.680±0.309	0.829±0.106
		Mutual Info	0.737±0.166	0.773±0.185	0.660±0.326	0.813±0.115
		Boruta	0.741±0.161	0.743±0.177	0.710±0.321	0.788±0.114
		PCA	0.546±0.080	0.758±0.144	0.120±0.172	0.972±0.029
		LDA	0.571±0.092	0.590±0.138	0.320±0.194	0.822±0.045
	T2	None	0.703±0.177	0.730±0.173	0.600±0.332	0.806±0.099
		Mutual Info	0.627±0.124	0.721±0.158	0.380±0.223	0.875±0.098
		Boruta	0.645±0.122	0.672±0.118	0.660±0.254	0.630±0.109
		PCA	0.561±0.059	0.745±0.172	0.150±0.136	0.972±0.031
		LDA	0.579±0.095	0.576±0.137	0.370±0.210	0.789±0.070

Table A.10: NB SMOTE ratio and K features for UCSD-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
NB	FLAIR	None	0.42	1223
		Mutual Info	0.35	158
		Boruta	0.90	21
		PCA	0.46	12
		LDA	0.82	1
	T1	None	0.69	1223
		Mutual Info	0.63	196
		Boruta	0.42	16
		PCA	0.35	37
		LDA	0.82	1
	T1GD	None	0.77	1223
		Mutual Info	0.35	43
		Boruta	0.95	19
		PCA	0.35	36
		LDA	0.82	1
	T2	None	0.35	1223
		Mutual Info	0.37	9
		Boruta	0.95	21
		PCA	0.35	17
		LDA	0.82	1

k-Nearest Neighbors

Table A.11: Performance of KNN Model for UCSD-PDGM.

ML Model	MRI Seq	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
KNN	FLAIR	None	0.696±0.109	0.759±0.151	0.480±0.213	0.913±0.091
		Mutual Info	0.685±0.107	0.743±0.174	0.430±0.228	0.941±0.072
		Boruta	0.698±0.113	0.774±0.151	0.450±0.233	0.946±0.064
		PCA	0.509±0.050	0.641±0.132	0.050±0.092	0.969±0.025
		LDA	0.634±0.128	0.651±0.149	0.520±0.248	0.748±0.082
	T1	None	0.655±0.111	0.705±0.188	0.400±0.232	0.910±0.099
		Mutual Info	0.639±0.095	0.695±0.158	0.350±0.175	0.928±0.099
		Boruta	0.639±0.110	0.682±0.195	0.340±0.201	0.939±0.097
		PCA	0.539±0.042	0.703±0.093	0.090±0.083	0.990±0.021
		LDA	0.629±0.074	0.659±0.084	0.480±0.160	0.779±0.053
	T1GD	None	0.674±0.108	0.757±0.162	0.430±0.253	0.918±0.099
		Mutual Info	0.721±0.127	0.744±0.190	0.560±0.280	0.882±0.100
		Boruta	0.733±0.126	0.751±0.203	0.550±0.269	0.915±0.106
		PCA	0.662±0.130	0.754±0.147	0.390±0.263	0.933±0.073
		LDA	0.577±0.102	0.583±0.111	0.390±0.197	0.763±0.059
	T2	None	0.681±0.117	0.765±0.188	0.420±0.227	0.941±0.065
		Mutual Info	0.701±0.159	0.750±0.177	0.570±0.293	0.831±0.123
		Boruta	0.681±0.137	0.746±0.164	0.420±0.268	0.941±0.056
		PCA	0.591±0.049	0.669±0.125	0.200±0.089	0.982±0.026
		LDA	0.590±0.087	0.602±0.103	0.410±0.192	0.771±0.074

Table A.12: KNN SMOTE ratio and K features for UCSD-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
KNN	FLAIR	None	0.38	1223
		Mutual Info	0.35	352
		Boruta	0.36	21
		PCA	0.35	39
		LDA	0.73	1
	T1	None	0.27	1223
		Mutual Info	0.36	624
		Boruta	0.35	16
		PCA	0.37	50
		LDA	0.78	1
	T1GD	None	0.26	1223
		Mutual Info	0.51	1223
		Boruta	0.26	19
		PCA	0.47	19
		LDA	0.47	1
	T2	None	0.26	1223
		Mutual Info	0.59	508
		Boruta	0.26	21
		PCA	0.47	17
		LDA	0.78	1

A.2 Dataset 2: UPENN-GBM

Support Vector Machine

Table A.13: Performance of SVM Model for UPENN-GBM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
SVM	FLAIR	None	0.569±0.161	0.560±0.315	0.250±0.335	0.888±0.032
		Mutual Info	0.640±0.199	0.644±0.359	0.300±0.400	0.980±0.013
		Boruta	0.588±0.169	0.609±0.351	0.200±0.332	0.976±0.020
		PCA	0.553±0.197	0.549±0.339	0.300±0.400	0.805±0.053
		LDA	0.549±0.158	0.624±0.193	0.150±0.320	0.948±0.038
	T1	None	0.539±0.147	0.679±0.244	0.100±0.300	0.978±0.017
		Mutual Info	0.646±0.197	0.831±0.208	0.300±0.400	0.992±0.010
		Boruta	0.615±0.203	0.607±0.261	0.500±0.447	0.729±0.100
		PCA	0.662±0.196	0.774±0.168	0.350±0.391	0.974±0.016
		LDA	0.605±0.192	0.659±0.281	0.300±0.400	0.910±0.031
	T1GD	None	0.816±0.161	0.850±0.174	0.850±0.320	0.783±0.070
		Mutual Info	0.588±0.167	0.763±0.224	0.200±0.332	0.976±0.024
		Boruta	0.517±0.075	0.688±0.226	0.050±0.150	0.984±0.015
		PCA	0.721±0.199	0.843±0.186	0.600±0.374	0.842±0.224
		LDA	0.558±0.163	0.646±0.240	0.200±0.332	0.916±0.034
	T2	None	0.679±0.236	0.796±0.213	0.400±0.490	0.958±0.026
		Mutual Info	0.695±0.245	0.841±0.199	0.400±0.490	0.990±0.010
		Boruta	0.459±0.094	0.646±0.289	0.050±0.150	0.867±0.266
		PCA	0.500±0.000	0.778±0.206	0.000±0.000	1.000±0.000
		LDA	0.676±0.212	0.710±0.212	0.550±0.415	0.801±0.067

Table A.14: SVM SMOTE ratio and K features for UPENN-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
SVM	FLAIR	None	0.36	1223
		Mutual Info	0.72	954
		Boruta	0.73	21
		PCA	0.55	44
		LDA	0.37	1
	T1	None	0.47	1223
		Mutual Info	0.63	978
		Boruta	0.41	16
		PCA	0.63	37
		LDA	0.99	1
	T1GD	None	0.24	1223
		Mutual Info	0.82	614
		Boruta	0.55	19
		PCA	0.35	24
		LDA	0.35	1
	T2	None	0.37	1223
		Mutual Info	0.34	1004
		Boruta	0.38	21
		PCA	0.35	19
		LDA	0.37	1

XGBoost

Table A.15: Performance of XGBoost Model for UPENN-GBM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
XGBoost	FLAIR	None	0.639±0.201	0.643±0.288	0.300±0.400	0.978±0.027
		Mutual Info	0.653±0.227	0.633±0.314	0.350±0.450	0.956±0.022
		Boruta	0.635±0.199	0.661±0.262	0.300±0.400	0.970±0.029
		PCA	0.634±0.213	0.628±0.297	0.450±0.415	0.818±0.074
		LDA	0.617±0.198	0.616±0.198	0.300±0.400	0.933±0.083
	T1	None	0.642±0.204	0.800±0.198	0.300±0.400	0.984±0.018
		Mutual Info	0.623±0.200	0.778±0.216	0.250±0.403	0.996±0.008
		Boruta	0.563±0.167	0.645±0.246	0.200±0.332	0.926±0.044
		PCA	0.633±0.200	0.709±0.191	0.300±0.400	0.966±0.026
		LDA	0.591±0.195	0.657±0.209	0.200±0.400	0.982±0.019
	T1GD	None	0.689±0.222	0.908±0.154	0.400±0.436	0.978±0.019
		Mutual Info	0.618±0.201	0.829±0.207	0.250±0.403	0.986±0.016
		Boruta	0.757±0.234	0.833±0.185	0.550±0.472	0.964±0.025
		PCA	0.754±0.208	0.840±0.183	0.750±0.403	0.759±0.088
		LDA	0.549±0.147	0.546±0.263	0.150±0.320	0.947±0.066
	T2	None	0.548±0.151	0.772±0.214	0.100±0.300	0.996±0.008
		Mutual Info	0.641±0.201	0.709±0.221	0.300±0.400	0.982±0.011
		Boruta	0.622±0.207	0.673±0.296	0.300±0.400	0.944±0.038
		PCA	0.582±0.205	0.775±0.165	0.200±0.400	0.964±0.033
		LDA	0.615±0.195	0.668±0.230	0.350±0.450	0.880±0.099

Table A.16: XGBoost SMOTE ratio and K features for UPENN-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
XGB	FLAIR	None	0.84	1223
		Mutual Info	0.72	983
		Boruta	1.00	7
		PCA	0.49	3
		LDA	0.67	1
	T1	None	0.42	1223
		Mutual Info	0.99	754
		Boruta	0.93	4
		PCA	25	0.59
		LDA	0.98	1
	T1GD	None	0.72	1223
		Mutual Info	0.82	614
		Boruta	0.42	3
		PCA	0.98	1
		LDA	0.39	1
	T2	None	0.47	1223
		Mutual Info	1.0	640
		Boruta	1.00	8
		PCA	0.82	21
		LDA	0.95	1

Random Forest

Table A.17: Performance of RF Model for UPENN-GBM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
RF	FLAIR	None	0.598±0.166	0.632±0.316	0.200±0.332	0.996±0.008
		Mutual Info	0.582±0.171	0.636±0.324	0.200±0.332	0.964±0.033
		Boruta	0.593±0.169	0.641±0.326	0.200±0.332	0.986±0.016
		PCA	0.526±0.115	0.656±0.225	0.150±0.229	0.902±0.034
		LDA	0.619±0.193	0.615±0.197	0.300±0.400	0.938±0.045
	T1	None	0.548±0.151	0.774±0.265	0.100±0.300	0.996±0.008
		Mutual Info	0.498±0.004	0.743±0.302	0.000±0.000	0.996±0.008
		Boruta	0.547±0.151	0.750±0.269	0.100±0.300	0.994±0.009
		PCA	0.521±0.070	0.663±0.323	0.050±0.150	0.992±0.013
		LDA	0.625±0.195	0.621±0.196	0.300±0.400	0.950±0.035
	T1GD	None	0.595±0.200	0.788±0.261	0.200±0.400	0.990±0.013
		Mutual Info	0.620±0.200	0.826±0.208	0.250±0.403	0.990±0.013
		Boruta	0.734±0.224	0.808±0.241	0.500±0.447	0.968±0.029
		PCA	0.744±0.178	0.754±0.289	0.650±0.391	0.837±0.083
		LDA	0.546±0.156	0.568±0.163	0.150±0.320	0.942±0.033
	T2	None	0.597±0.199	0.749±0.306	0.200±0.400	0.994±0.009
		Mutual Info	0.562±0.160	0.672±0.275	0.150±0.320	0.974±0.027
		Boruta	0.597±0.199	0.749±0.306	0.200±0.400	0.994±0.009
		PCA	0.498±0.004	0.638±0.208	0.000±0.000	0.996±0.008
		LDA	0.570±0.198	0.564±0.201	0.200±0.400	0.940±0.024

Table A.18: RF SMOTE ratio and K features for UPENN-GBM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
RF	FLAIR	None	0.35	1223
		Mutual Info	0.35	509
		Boruta	0.28	11
		PCA	0.66	4
		LDA	0.91	1
	T1	None	0.36	1223
		Mutual Info	1.0	443
		Boruta	0.72	7
		PCA	0.78	24
		LDA	0.99	1
	T1GD	None	0.36	1223
		Mutual Info	0.71	919
		Boruta	0.49	19
		PCA	0.31	9
		LDA	0.43	1
	T2	None	0.77	1223
		Mutual Info	0.40	645
		Boruta	0.24	16
		PCA	0.74	48
		LDA	0.37	1

Logistic Regression

Table A.19: Performance of LogR Model for UPENN-GBM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
LogR	FLAIR	None	0.612±0.180	0.710±0.272	0.350±0.391	0.873±0.068
		Mutual Info	0.604±0.165	0.689±0.270	0.400±0.374	0.807±0.073
		Boruta	0.609±0.166	0.613±0.364	0.250±0.335	0.968±0.016
		PCA	0.677±0.194	0.728±0.315	0.400±0.374	0.954±0.033
		LDA	0.622±0.191	0.608±0.295	0.350±0.391	0.893±0.053
	T1	None	0.757±0.216	0.797±0.257	0.600±0.436	0.914±0.043
		Mutual Info	0.753±0.185	0.790±0.258	0.600±0.374	0.906±0.048
		Boruta	0.726±0.196	0.795±0.260	0.500±0.387	0.952±0.031
		PCA	0.690±0.193	0.798±0.164	0.450±0.415	0.930±0.038
		LDA	0.630±0.187	0.657±0.290	0.350±0.391	0.910±0.038
	T1GD	None	0.756±0.199	0.852±0.193	0.700±0.400	0.811±0.033
		Mutual Info	0.679±0.215	0.858±0.150	0.400±0.436	0.958±0.021
		Boruta	0.738±0.207	0.866±0.160	0.550±0.415	0.926±0.032
		PCA	0.771±0.204	0.848±0.157	0.650±0.391	0.892±0.051
		LDA	0.565±0.162	0.546±0.292	0.200±0.332	0.930±0.036
	T2	None	0.712±0.234	0.869±0.179	0.450±0.472	0.974±0.025
		Mutual Info	0.713±0.233	0.872±0.191	0.450±0.472	0.976±0.029
		Boruta	0.678±0.220	0.845±0.192	0.400±0.436	0.956±0.015
		PCA	0.728±0.218	0.842±0.188	0.500±0.447	0.956±0.023
		LDA	0.558±0.200	0.650±0.248	0.200±0.400	0.916±0.042

Table A.20: LogR SMOTE ratio and K features for UPENN-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
LogR	FLAIR	None	0.44	1223
		Mutual Info	0.68	264
		Boruta	0.36	11
		PCA	0.37	6
		LDA	0.92	1
	T1	None	0.33	1223
		Mutual Info	0.23	1124
		Boruta	0.25	7
		PCA	0.35	37
		LDA	0.98	1
	T1GD	None	0.66	1223
		Mutual Info	1.0	531
		Boruta	0.21	19
		PCA	0.49	9
		LDA	0.73	1
	T2	None	0.64	1223
		Mutual Info	0.50	1055
		Boruta	0.34	16
		PCA	1.0	50
		LDA	0.94	1

Naïve Bayes

Table A.21: Performance of NB Model for UPENN-GBM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
NB	FLAIR	None	0.643±0.177	0.645±0.194	0.750±0.403	0.536±0.144
		Mutual Info	0.614±0.191	0.686±0.287	0.350±0.391	0.877±0.078
		Boruta	0.573±0.181	0.584±0.203	0.800±0.400	0.346±0.121
		PCA	0.643±0.192	0.718±0.218	0.350±0.391	0.936±0.040
		LDA	0.493±0.010	0.563±0.166	0.000±0.000	0.986±0.020
	T1	None	0.675±0.179	0.666±0.225	0.750±0.403	0.601±0.124
		Mutual Info	0.697±0.215	0.754±0.270	0.600±0.436	0.793±0.084
		Boruta	0.666±0.177	0.659±0.227	0.750±0.403	0.583±0.126
		PCA	0.666±0.192	0.756±0.283	0.350±0.391	0.982±0.019
		LDA	0.627±0.197	0.661±0.295	0.300±0.400	0.954±0.035
	T1GD	None	0.781±0.151	0.826±0.164	0.750±0.403	0.811±0.078
		Mutual Info	0.749±0.194	0.835±0.157	0.700±0.400	0.797±0.085
		Boruta	0.668±0.225	0.863±0.166	0.350±0.450	0.986±0.018
		PCA	0.833±0.169	0.822±0.186	0.850±0.320	0.815±0.078
		LDA	0.484±0.011	0.444±0.208	0.000±0.000	0.968±0.023
	T2	None	0.635±0.173	0.669±0.196	0.800±0.332	0.470±0.134
		Mutual Info	0.546±0.212	0.510±0.311	0.450±0.415	0.641±0.146
		Boruta	0.572±0.239	0.595±0.233	0.600±0.436	0.544±0.088
		PCA	0.617±0.162	0.744±0.262	0.250±0.335	0.984±0.018
		LDA	0.534±0.156	0.606±0.247	0.100±0.300	0.968±0.029

Table A.22: NB SMOTE ratio and K features for UPENN-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
NB	FLAIR	None	0.35	1223
		Mutual Info	0.35	565
		Boruta	0.91	21
		PCA	1.0	1
		LDA	0.99	1
	T1	None	0.46	1223
		Mutual Info	0.35	162
		Boruta	0.42	16
		PCA	0.42	29
		LDA	0.93	1
	T1GD	None	0.57	1223
		Mutual Info	1.0	569
		Boruta	0.35	19
		PCA	0.35	11
		LDA	0.35	1
	T2	None	0.36	1223
		Mutual Info	0.68	23
		Boruta	0.95	21
		PCA	0.47	1
		LDA	0.85	1

k-Nearest Neighbors

Table A.23: Performance of KNN Model for UPENN-GBM.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
KNN	FLAIR	None	0.421±0.026	0.478±0.185	0.000±0.000	0.841±0.051
		Mutual Info	0.639±0.208	0.685±0.281	0.550±0.415	0.729±0.069
		Boruta	0.647±0.222	0.621±0.305	0.400±0.436	0.893±0.059
		PCA	0.590±0.214	0.661±0.226	0.550±0.415	0.630±0.085
		LDA	0.619±0.193	0.619±0.193	0.300±0.400	0.938±0.045
	T1	None	0.655±0.235	0.746±0.256	0.550±0.472	0.759±0.055
		Mutual Info	0.702±0.236	0.747±0.273	0.550±0.472	0.853±0.046
		Boruta	0.652±0.224	0.743±0.254	0.450±0.472	0.853±0.048
		PCA	0.644±0.223	0.739±0.245	0.350±0.450	0.938±0.024
		LDA	0.619±0.193	0.619±0.197	0.300±0.400	0.938±0.041
	T1GD	None	0.779±0.196	0.771±0.282	0.750±0.403	0.807±0.064
		Mutual Info	0.759±0.201	0.787±0.246	0.750±0.403	0.767±0.078
		Boruta	0.696±0.228	0.751±0.260	0.500±0.447	0.892±0.062
		PCA	0.764±0.190	0.753±0.300	0.650±0.391	0.878±0.060
		LDA	0.567±0.163	0.565±0.164	0.200±0.332	0.934±0.038
	T2	None	0.676±0.225	0.750±0.224	0.550±0.472	0.801±0.068
		Mutual Info	0.641±0.235	0.725±0.231	0.400±0.490	0.882±0.053
		Boruta	0.675±0.224	0.739±0.253	0.500±0.447	0.849±0.062
		PCA	0.567±0.197	0.691±0.232	0.200±0.400	0.934±0.031
		LDA	0.570±0.198	0.570±0.198	0.200±0.400	0.940±0.024

Table A.24: KNN SMOTE ratio and K features for UPENN-PDGM

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
KNN	FLAIR	None	0.45	1223
		Mutual Info	1.0	432
		Boruta	0.36	21
		PCA	1.0	1
		LDA	0.96	1
	T1	None	1.0	1223
		Mutual Info	0.35	1203
		Boruta	0.31	16
		PCA	0.32	40
		LDA	0.99	1
	T1GD	None	0.74	1223
		Mutual Info	0.79	288
		Boruta	0.26	19
		PCA	0.48	44
		LDA	0.44	1
	T2	None	0.69	1223
		Mutual Info	0.33	1032
		Boruta	0.36	21
		PCA	0.31	50
		LDA	0.37	1

A.3 Dataset 3: Combined

Support Vector Machine

Table A.25: Performance of SVM Model for Combined Dataset.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
SVM	FLAIR	None	0.789±0.115	0.859±0.113	0.688±0.274	0.891±0.104
		Mutual Info	0.778±0.116	0.853±0.123	0.636±0.268	0.920±0.094.
		Boruta	0.724±0.098	0.787±0.123	0.542±0.246	0.906±0.068
		PCA	0.713±0.110	0.695±0.178	0.496±0.229	0.930±0.076
		LDA	0.638±0.114	0.653±0.124	0.558±0.197	0.719±0.049
	T1	None	0.750±0.122	0.865±0.098	0.561±0.290	0.939±0.078
		Mutual Info	0.737±0.135	0.866±0.105	0.501±0.289	0.974±0.030
		Boruta	0.735±0.105	0.802±0.107	0.627±0.247	0.843±0.108
		PCA	0.669±0.081	0.729±0.149	0.392±0.168	0.946±0.029
		LDA	0.609±0.074	0.669±0.059	0.511±0.255	0.708±0.176
	T1GD	None	0.848±0.108	0.932±0.059	0.824±0.222	0.871±0.061
		Mutual Info	0.804±0.108	0.924±0.063	0.669±0.236	0.939±0.053
		Boruta	0.831±0.094	0.927±0.072	0.695±0.202	0.966±0.038
		PCA	0.747±0.070	0.886±0.064	0.548±0.172	0.945±0.047
		LDA	0.568±0.076	0.712±0.097	0.262±0.133	0.875±0.032
	T2	None	0.802±0.094	0.880±0.091	0.653±0.221	0.952±0.055
		Mutual Info	0.500±0.000	0.501±0.100	0.000±0.000	1.000±0.000
		Boruta	0.766±0.084	0.838±0.100	0.661±0.205	0.872±0.100
		PCA	0.500±0.000	0.505±0.147	0.500±0.500	0.500±0.500
		LDA	0.642±0.102	0.657±0.108	0.566±0.237	0.718±0.068

Table A.26: SVM SMOTE ratio and K features for Combined Dataset

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
SVM	FLAIR	None	0.35	1223
		Mutual Info	0.58	723
		Boruta	0.73	24
		PCA	0.61	44
		LDA	0.92	1
	T1	None	0.35	1223
		Mutual Info	0.35	952
		Boruta	0.43	17
		PCA	0.23	37
		LDA	0.82	1
	T1GD	None	0.91	1223
		Mutual Info	0.46	841
		Boruta	0.57	19
		PCA	0.33	24
		LDA	1.0	1
	T2	None	0.49	1223
		Mutual Info	0.26	518
		Boruta	0.43	11
		PCA	0.37	19
		LDA	0.73	1

XGBoost

Table A.27: Performance of XGBoost Model for Combined Dataset.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
XGB	FLAIR	None	0.762±0.123	0.871±0.115	0.577±0.278	0.948±0.066
		Mutual Info	0.745±0.117	0.873±0.115	0.542±0.266	0.947±0.065
		Boruta	0.756±0.126	0.858±0.118	0.592±0.278	0.919±0.092
		PCA	0.766±0.138	0.863±0.096	0.568±0.289	0.963±0.041
		LDA	0.615±0.094	0.669±0.132	0.367±0.178	0.863±0.028
	T1	None	0.749±0.099	0.864±0.116	0.541±0.232	0.957±0.059
		Mutual Info	0.749±0.085	0.891±0.096	0.523±0.190	0.975±0.035
		Boruta	0.761±0.120	0.858±0.119	0.611±0.287	0.910±0.105
		PCA	0.755±0.119	0.862±0.101	0.568±0.272	0.942±0.070
		LDA	0.649±0.075	0.684±0.086	0.633±0.142	0.664±0.051
	T1GD	None	0.834±0.068	0.935±0.066	0.702±0.156	0.965±0.037
		Mutual Info	0.834±0.070	0.934±0.062	0.702±0.156	0.966±0.036
		Boruta	0.822±0.088	0.906±0.076	0.694±0.189	0.949±0.037
		PCA	0.835±0.098	0.946±0.048	0.705±0.216	0.966±0.037
		LDA	0.626±0.061	0.664±0.074	0.615±0.113	0.637±0.046
	T2	None	0.787±0.105	0.893±0.079	0.627±0.233	0.947±0.052
		Mutual Info	0.751±0.090	0.893±0.076	0.557±0.215	0.945±0.071
		Boruta	0.779±0.088	0.879±0.086	0.652±0.206	0.907±0.078
		PCA	0.620±0.084	0.860±0.092	0.245±0.175	0.994±0.009
		LDA	0.644±0.117	0.673±0.106	0.412±0.268	0.877±0.059

Table A.28: XGBoost SMOTE ratio and K features for Combined Dataset

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
XGB	FLAIR	None	0.53	1223
		Mutual Info	1.0	684
		Boruta	0.74	7
		PCA	0.36	18
		LDA	0.58	1
	T1	None	0.59	1223
		Mutual Info	0.35	1196
		Boruta	0.49	4
		PCA	0.71	50
		LDA	0.45	1
	T1GD	None	0.35	1223
		Mutual Info	0.38	1102
		Boruta	0.52	3
		PCA	0.36	39
		LDA	0.69	1
	T2	None	0.97	1223
		Mutual Info	0.70	1106
		Boruta	1.0	8
		PCA	0.35	44
		LDA	0.36	1

Random Forest

Table A.29: Performance of RF Model for Combined Dataset.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
RF	FLAIR	None	0.746±0.107	0.875±0.108	0.516±0.235	0.975±0.035
		Mutual Info	0.746±0.109	0.888±0.106	0.515±0.232	0.978±0.033
		Boruta	0.731±0.093	0.870±0.109	0.488±0.202	0.974±0.034
		PCA	0.734±0.115	0.853±0.112	0.517±0.268	0.951±0.074
		LDA	0.634±0.123	0.635±0.122	0.507±0.225	0.762±0.038
	T1	None	0.736±0.106	0.869±0.107	0.497±0.227	0.975±0.035
		Mutual Info	0.735±0.107	0.860±0.109	0.498±0.233	0.972±0.036
		Boruta	0.731±0.099	0.867±0.106	0.488±0.212	0.974±0.028
		PCA	0.725±0.102	0.874±0.081	0.490±0.242	0.960±0.059
		LDA	0.649±0.047	0.657±0.046	0.485±0.109	0.814±0.037
	T1GD	None	0.837±0.064	0.884±0.072	0.720±0.132	0.933±0.046
		Mutual Info	0.839±0.063	0.924±0.061	0.719±0.147	0.959±0.039
		Boruta	0.849±0.075	0.933±0.062	0.772±0.182	0.927±0.057
		PCA	0.856±0.091	0.930±0.049	0.781±0.207	0.931±0.049
		LDA	0.667±0.075	0.676±0.073	0.478±0.140	0.855±0.036
	T2	None	0.717±0.79	0.832±0.0101	0.525±0.236	0.971±0.039
		Mutual Info	0.739±0.82	0.858±0.092	0.524±0.232	0.974±0.037
		Boruta	0.729±0.103	0.878±0.093	0.482±0.226	0.976±0.035
		PCA	0.695±0.091	0.881±0.090	0.402±0.193	0.988±0.018
		LDA	0.637±0.127	0.644±0.121	0.499±0.302	0.774±0.068

Table A.30: RF SMOTE ratio and K features for Combined Dataset.

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
RF	FLAIR	None	0.45	1223
		Mutual Info	0.43	1006
		Boruta	0.69	21
		PCA	0.94	41
		LDA	0.53	1
	T1	None	0.35	1223
		Mutual Info	0.34	863
		Boruta	0.38	16
		PCA	0.98	42
		LDA	0.77	1
	T1GD	None	0.9	1223
		Mutual Info	0.48	1071
		Boruta	0.63	19
		PCA	0.82	26
		LDA	0.68	1
	T2	None	0.58	1223
		Mutual Info	0.81	699
		Boruta	0.40	21
		PCA	0.35	38
		LDA	0.72	1

Logistic Regression

Table A.31: Performance of LogR Model for Combined Dataset.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
LogR	FLAIR	None	0.766±0.122	0.849±0.124	0.585±0.267	0.947±0.058
		Mutual Info	0.776±0.135	0.862±0.108	0.601±0.293	0.952±0.056
		Boruta	0.756±0.114	0.804±0.122	0.594±0.268	0.919±0.071
		PCA	0.769±0.124	0.867±0.130	0.732±0.280	0.805±0.186
		LDA	0.631±0.121	0.668±0.143	0.549±0.211	0.713±0.048
	T1	None	0.757±0.089	0.874±0.109	0.566±0.215	0.947±0.055
		Mutual Info	0.776±0.095	0.874±0.095	0.731±0.243	0.821±0.170
		Boruta	0.768±0.123	0.858±0.109	0.646±0.301	0.890±0.102
		PCA	0.763±0.100	0.855±0.092	0.695±0.231	0.831±0.165
		LDA	0.649±0.063	0.676±0.091	0.607±0.105	0.692±0.054
	T1GD	None	0.857±0.100	0.935±0.052	0.772±0.216	0.941±0.044
		Mutual Info	0.853±0.095	0.937±0.051	0.764±0.209	0.943±0.044
		Boruta	0.818±0.081	0.899±0.078	0.722±0.195	0.913±0.050
		PCA	0.845±0.096	0.933±0.058	0.748±0.216	0.941±0.048
		LDA	0.664±0.054	0.663±0.041	0.614±0.094	0.714±0.050
	T2	None	0.788±0.082	0.877±0.081	0.687±0.205	0.890±0.088
		Mutual Info	0.802±0.074	0.890±0.072	0.730±0.194	0.873±0.102
		Boruta	0.782±0.100	0.820±0.095	0.645±0.226	0.920±0.053
		PCA	0.800±0.075	0.869±0.068	0.730±0.205	0.869±0.108
		LDA	0.635±0.076	0.668±0.112	0.634±0.185	0.636±0.059

Table A.32: LogR SMOTE ratio and K features for Combined Dataset

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
LogR	FLAIR	None	0.48	1223
		Mutual Info	0.36	809
		Boruta	0.37	21
		PCA	0.97	13
		LDA	0.87	1
	T1	None	0.35	1223
		Mutual Info	1.0	757
		Boruta	0.29	16
		PCA	0.41	48
		LDA	0.40	1
	T1GD	None	0.87	1223
		Mutual Info	0.35	651
		Boruta	0.25	19
		PCA	0.41	50
		LDA	0.37	1
	T2	None	0.76	1223
		Mutual Info	1.0	952
		Boruta	0.31	21
		PCA	1.0	27
		LDA	0.72	1

Naïve Bayes

Table A.33: Performance of NB Model for Combined Dataset.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
NB	FLAIR	None	0.673±0.122	0.743±0.130	0.719±0.179	0.626±0.254
		Mutual Info	0.683±0.082	0.778±0.124	0.439±0.160	0.928±0.073
		Boruta	0.662±0.132	0.667±0.141	0.856±0.187	0.468±0.262
		PCA	0.686±0.120	0.831±0.104	0.431±0.244	0.940±0.063
		LDA	0.633±0.126	0.666±0.155	0.524±0.223	0.741±0.051
	T1	None	0.723±0.071	0.730±0.118	0.545±0.135	0.901±0.069
		Mutual Info	0.658±0.069	0.741±0.116	0.334±0.133	0.982±0.015
		Boruta	0.655±0.079	0.651±0.129	0.626±0.199	0.684±0.139
		PCA	0.730±0.080	0.816±0.101	0.770±0.193	0.690±0.175
		LDA	0.644±0.071	0.637±0.104	0.452±0.140	0.836±0.033
	T1GD	None	0.828±0.082	0.885±0.067	0.790±0.179	0.866±0.066
		Mutual Info	0.836±0.070	0.917±0.062	0.875±0.155	0.797±0.063
		Boruta	0.837±0.080	0.859±0.063	0.832±0.174	0.842±0.073
		PCA	0.810±0.098	0.915±0.055	0.696±0.215	0.925±0.051
		LDA	0.661±0.064	0.662±0.041	0.426±0.116	0.896±0.028
	T2	None	0.763±0.085	0.768±0.116	0.617±0.195	0.910±0.083
		Mutual Info	0.731±0.087	0.802±0.107	0.511±0.189	0.951±0.055
		Boruta	0.738±0.090	0.739±0.120	0.695±0.197	0.780±0.146
		PCA	0.732±0.103	0.841±0.104	0.545±0.216	0.919±0.063
		LDA	0.650±0.128	0.669±0.111	0.464±0.285	0.836±0.058

Table A.34: NB SMOTE ratio and K features for Combined Dataset.

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
NB	FLAIR	None	0.36	1223
		Mutual Info	0.69	73
		Boruta	0.91	21
		PCA	0.40	34
		LDA	0.78	1
	T1	None	0.46	1223
		Mutual Info	0.65	204
		Boruta	0.45	16
		PCA	0.89	50
		LDA	0.35	1
	T1GD	None	0.34	1223
		Mutual Info	0.39	38
		Boruta	0.97	19
		PCA	0.53	20
		LDA	0.36	1
	T2	None	1.0	1223
		Mutual Info	1.0	180
		Boruta	0.89	21
		PCA	0.55	44
		LDA	0.36	1

k-Nearest Neighbors

Table A.35: Performance of KNN Model for Combined Dataset.

ML Model	MRI Sequence	Feature Selection	Balanced Accuracy	AUC	Sensitivity	Specificity
KNN	FLAIR	None	0.765±0.105	0.822±0.117	0.617±0.237	0.912±0.097
		Mutual Info	0.769±0.101	0.818±0.114	0.609±0.227	0.928±0.084
		Boruta	0.771±0.111	0.822±0.135	0.601±0.244	0.942±0.075
		PCA	0.753±0.109	0.820±0.131	0.566±0.229	0.940±0.073
		LDA	0.626±0.122	0.630±0.128	0.498±0.229	0.754±0.048
	T1	None	0.751±0.109	0.799±0.135	0.567±0.244	0.935±0.071
		Mutual Info	0.724±0.122	0.800±0.127	0.627±0.260	0.820±0.139
		Boruta	0.760±0.112	0.814±0.129	0.567±0.253	0.952±0.064
		PCA	0.756±0.123	0.810±0.131	0.558±0.264	0.954±0.059
		LDA	0.655±0.048	0.659±0.050	0.502±0.104	0.809±0.035
	T1GD	None	0.817±0.100	0.917±0.077	0.705±0.222	0.929±0.045
		Mutual Info	0.821±0.079	0.910±0.065	0.815±0.177	0.827±0.054
		Boruta	0.818±0.081	0.899±0.078	0.722±0.195	0.913±0.050
		PCA	0.825±0.091	0.915±0.077	0.739±0.208	0.910±0.049
		LDA	0.675±0.065	0.690±0.072	0.503±0.124	0.847±0.030
	T2	None	0.759±0.083	0.849±0.086	0.721±0.212	0.796±0.108
		Mutual Info	0.809±0.075	0.851±0.074	0.685±0.168	0.934±0.068
		Boruta	0.785±0.089	0.825±0.092	0.635±0.214	0.935±0.063
		PCA	0.778±0.094	0.878±0.067	0.617±0.225	0.939±0.069
		LDA	0.646±0.127	0.646±0.130	0.516±0.287	0.775±0.060

Table A.36: KNN SMOTE ratio and K features for Combined Dataset.

ML Model	MRI Sequence	Feature Selection	SMOTE Ratio	K
KNN	FLAIR	None	.36	1223
		Mutual Info	0.35	701
		Boruta	0.35	21
		PCA	0.34	16
		LDA	0.87	1
	T1	None	0.27	1223
		Mutual Info	0.74	725
		Boruta	0.36	16
		PCA	0.36	50
		LDA	0.83	1
	T1GD	None	0.26	1223
		Mutual Info	1.0	463
		Boruta	0.24	19
		PCA	0.53	30
		LDA	0.42	1
	T2	None	0.92	1223
		Mutual Info	0.44	427
		Boruta	0.28	21
		PCA	0.54	50
		LDA	0.37	1