People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
University of Setif 1 - Ferhat Abbas

University of Setif 1 - Ferhat Abbas

THESIS

Faculty of Sciences

Presented at the Faculty of Sciences
Department of Computer Science
For the Award of the Degree of

DOCTORATE IN SCIENCES

Doctoral School
Information and Communication Sciences and Technologies
Option: Computer Systems Engineering
By:

Youssef BOULKHIOUT

TITLE

Knowledge Extraction with Machine Learning Techniques from Multi-modal MRI Data: Application to Gliomas Classification

Presented on **February 17, 2026**, before a jury composed of:

President:	SLIMANI Yacine	MCA	University of Setif 1 - Ferhat Abbas
Supervisor:	MOUSSAOUI Abdelouahab	Prof	University of Setif 1 - Ferhat Abbas
Examiners:	BABAHENINI Mohamed Chaouki	Prof	University of Biskra - Mohamed Khider
	ATTIA Abdelouahab	Prof	University of Bordj Bou Arreridj - Mohamed El Bachir El Ibrahimi
	BOUAITA Bilal	MCA	University of Constantine 3 - Salah Boubnider
Guests:	KECHADI Mohand-Tahar	Prof	University of College Dublinen Irlande

Acknowledgment

First and foremost, praise be to God for granting us the mental and physical abilities that enabled us to complete this modest thesis, and we beseech Him to crown our work with success.

I would like to express my deepest gratitude to my supervisor, Pr. MOUSSAOUI Abdelouahab, for initiating this project and for his continuous support, encouragement, and constructive feedback that enriched this work. I also warmly thank the esteemed jury members for evaluating this humble work; their review, insights, and considerate perspectives are deeply appreciated.

A special thank you:

I would also like to express my sincere thanks to K. NASRI, S. BALBAL, M. SEMCHEDDINE, and all those who, through numerous discussions, contributed to shaping some of the ideas presented in this thesis. I am truly grateful for their support, their cooperation during the realization of this modest work, and the enriching exchanges that greatly benefited my reflection.

I also thank the various individuals who shared this passion with me and, indirectly but undeniably, contributed to this thesis.

Finally, I thank my teachers and everyone who has taught me even a single letter since birth.

Youssef BOULKHIOUT
February 17, 2026

To my brothers and sisters
To my wife and children

Contents

Acknowledgment . . . . . **i**

Contents . . . . . **iii**

List of Figures . . . . . **vii**

List of Tables . . . . . **ix**

List of Algorithms . . . . . **x**

List of Abbreviations . . . . . **xi**

Introduction **1**

Thesis Goals and Results 2

Thesis Outline 3

Academic Publication 4

I Fundamentals of Brain Anatomy, Tumors, and Imaging **5**

I.1 Introduction 5

I.2 Cerebral Anatomy Overview 5

I.2.1 Central Nervous System 6

I.2.2 Major Brain Tissues 9

I.2.3 Ventricular System 11

I.3 Brain Tumors Overview 12

I.3.1 Types of Brain Tumors 13

I.3.1.1 Benign Brain Tumors 13

I.3.1.2 Malignant Brain Tumors 16

I.3.2 Diagnosis 17

I.3.3 Management and Treatment 18

I.3.4 Epidemiology 19

I.4 Planes for Brain Imaging 20

I.5 Magnetic Resonance Imaging 21

I.5.1 Fundamental Principles 22

I.5.1.1 Static Magnetic Field and Proton Alignment Mechanism . 22

I.5.1.2 Larmor Precession and Resonance Frequency 23

I.5.1.3 Radiofrequency Excitation and Flip Angle Dynamics . . . 23
I.5.1.4 Signal Generation and Image Reconstruction 24
I.5.2 Physical Principles of MRI 25
I.5.2.1 Nuclear Magnetic Resonance Fundamentals 25
I.5.2.2 Magnetic Moment and Magnetization Dynamics 26
I.5.2.3 Relaxation Mechanisms 26
I.5.3 Image Contrast and Weighted Imaging Techniques 27
I.5.4 Spatial Encoding Techniques 29
I.5.5 Characteristics of MRI Images 30
I.5.6 Common MRI Artifacts and Mitigation 31
I.5.7 Advantages and Limitations 32
I.5.8 Clinical Applications 33
I.6 Conclusion 34

II Supervised Machine Learning Algorithms 35
Introduction 35
II.1 Classification 36
II.1.1 Perceptron 36
II.1.2 Support Vector Machines 38
II.1.3 Logistic Regression 41
II.1.4 Naive Bayes 43
II.1.5 Decision Trees 46
II.1.6 Ensemble Methods 48
II.1.6.1 Bagging (Bootstrap Aggregating) 49
II.1.6.2 Stacking 51
II.1.6.3 Voting 52
II.1.6.4 Boosting 53
II.2 Regression 60
II.2.1 Simple Linear Regression 61
II.2.2 Multiple Linear Regression 63
II.2.3 Lasso Regression 64
II.2.4 Ridge Regression 66
II.2.5 Comparative Summary 67
Conclusion 68

III State of the Art on MGMT Promoter Methylation 69
III.1 Introduction 69
III.2 Molecular and Clinical Significance 69
III.2.1 Molecular Basis 70
III.2.2 Clinical Relevance 70

III.2.3 Clinical Integration 70
III.3 Testing Methodologies 71
III.3.1 Tissue-Based Molecular Diagnostics 71
III.3.2 Assay Variability and Cutoff Definition 72
III.3.3 Non-Invasive and Emerging Approaches 72
III.4 Methodological Landscape for Imaging-Based Prediction 73
III.5 Notable Studies 74
III.5.1 Classical Machine Learning on Radiomic Features 74
III.5.2 Multiparametric MRI and Hybrid Approaches 75
III.5.3 PET and Modality Fusion 76
III.5.4 Advanced Machine Learning and Deep Architectures 76
III.5.5 Representative AI Studies 78
III.6 Discussion 79
III.7 Conclusion 80

IV Proposed Approaches for MGMT Detection 81
IV.1 Introduction 81
IV.2 Materials and Methods 82
IV.2.1 Dataset Selection 83
IV.2.2 Feature Extraction 84
IV.2.3 Feature Selection 86
IV.2.4 Dataset Splitting 87
IV.2.5 Data Normalization 88
IV.2.6 Class Balancing 88
IV.2.7 Data Augmentation 89
IV.2.8 Training Models 89
IV.2.9 Evaluation & Metrics 89
IV.3 Contributions 91
IV.3.1 Random Forest 91
IV.3.2 LightGBM 93
IV.3.3 XGBoost 94
IV.3.4 CatBoost 95
IV.3.5 Voting Ensemble 96
IV.3.6 Stacking Ensemble 97
IV.4 Conclusion 98

V Results and Discussions 99
V.1 Introduction 99
V.2 Radiomics Feature Selection 99
V.3 Feature Distribution 101

V.4 Models Performances 102
V.4.1 Base Learners Performances 104
V.4.2 Ensembles Performances 107
V.5 Comparison of all Models 112
V.6 Conclusion 114

Conclusion **115**

Bibliography **123**

Abstract **125**

List of Figures

I.1 The Central Nervous System [1]. 6
I.2 Lateral View of the External Parts of Adult CNS. 7
I.3 Lateral View of the Internal Parts of Adult CNS. 7
I.4 Lateral View of Brainstem and Cranial. 8
I.5 Spinal Cord and Peripheral Nervous System. 9
I.6 White and Gray Matters. 10
I.7 Ventricular System [1]. 11
I.8 Brain Tumors. 12
I.9 Anatomical Distribution of Major Brain Tumors [2] 15
I.10 Standard Anatomical Planes for Brain MRI [3]. 20
I.11 Standard 2D Slice Projection Used in MRI. 21
I.12 MRI Scanner, Note the Coil Over the Patient's Head[4]. 22
I.13 Random Spin Orientations Become Aligned in Presence of B_0. 23
I.14 Larmor Precession & Flip Angle. 24
I.15 T1 Recovery (Blue) and T2 Decay (Red) Processes. 24
I.16 Representation of Image Reconstruction. 25
I.17 Free Induction Decay Following RF Excitation. 27
I.18 Image Contrast and Weighted Imaging Techniques. 28

II.1 Model Complexity Trade-Off Between Underfitting and Overfitting. 36
II.2 Max Margin Solution of SVM Classifier. 38
II.3 The Sigmoid Function and Logistic Function. 42
II.4 Probabilistic Graphical Model for Naive Bayes 44
II.5 Bidimensional Models. 62

III.1 Growth of MGMT Promoter Methylation Publications (2000–2025). 77

IV.1 The Modeling Pipeline. 82
IV.2 Data Selection from Brats 2021. 84
IV.3 Radiomics Features Extraction from IRM Scans. 85
IV.4 Results of Features Selection Steps Using LightGBM and CatBoost. 86
IV.5 Dataset Distribution over Subsets. 87

V.1 AUC of Voting Model Using Different Numbers of Features. 100

V.2 Accuracy of Voting Model Using Different Numbers of Features. 100
V.3 AUC of Stacking Model Using Different Numbers of Features. 100
V.4 Accuracy of Stacking Model Using Different Numbers of Features. 100
V.5 CatBoost ROC Visualization. 104
V.6 CatBoost Confusion Matrix Visualization. 104
V.7 XGBoost ROC Visualization 105
V.8 XGBoost Confusion Matrix Visualization. 105
V.9 LightGBM ROC Visualization 106
V.10 LightGBM Confusion Matrix Visualization. 106
V.11 Random Forest ROC Visualization. 106
V.12 Random Forest Confusion Matrix Visualization. 106
V.13 Voting Ensemble ROC Visualization. 108
V.14 Voting Ensemble Confusion Matrix Visualization. 108
V.15 Voting Ensemble Accuracy vs Threshold Visualization. 108
V.16 Voting Ensemble F1 Score vs Threshold Visualization. 108
V.17 Accuracy and AUC of Voting Ensemble vs Other Models. 109
V.18 Stacking Ensemble ROC Visualization. 110
V.19 Stacking Ensemble Confusion Matrix Visualization. 110
V.20 Stacking Ensemble Accuracy vs Threshold Visualization. 111
V.21 Stacking Ensemble F1 Score vs Threshold Visualization. 111
V.22 Accuracy and AUC of Stacking Ensemble vs Other Models. 112
V.23 AUC Comparison of all Models. 112
V.24 Accuracy Comparison of all Models. 112
V.25 F1 Score Comparison of all Models. 113
V.26 Precision Comparison of all Models. 113
V.27 Recall Comparison of all Models. 113
V.28 Kappa Comparison of all Models. 113

List of Tables

II.1 Comparative Summary of Regression Methods. 67

III.1 Performance of AI Models for MGMT Methylation Prediction. 78

IV.1 Radiomics Feature Distribution. 85
IV.2 Confusion Matrix. 91

V.1 Top 15 Radiomics Features Selected. 101
V.2 Test Performance Metrics (%). 102

List of Algorithms

1 Perceptron Algorithm 37
2 SVM Algorithm 40
3 Logistic Regression Algorithm 43
4 Naive Bayes Algorithm 45
5 CART Algorithm 47
6 Bagging Algorithm 49
7 Random Forest Algorithm 51
8 Stacking Algorithm 52
9 Voting Algorithm 53
10 AdaBoost Algorithm 54
11 Gradient Boosting Algorithm 55
12 LightGBM Algorithm 57
13 XGBoost Algorithm 58
14 CatBoost Algorithm 60
15 Simple Linear Regression Algorithm 62
16 Multiple Linear Regression Algorithm 64
17 Lasso Regression Algorithm 65
18 Ridge Regression Algorithm 67
19 Random Forest for MGMT Status Prediction 92
20 LightGBM for MGMT Status Prediction 93
21 XGBoost for MGMT Status Prediction 94
22 CatBoost for MGMT Status Prediction 95
23 Voting Ensemble for MGMT Status Prediction 96
24 Stacking Ensemble for MGMT Methylation Status Prediction 97

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
AUC	Area Under Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BraTS	Brain Tumor Segmentation
CART	Classification and Regression Trees
CatBoost	Categorical Boosting
CNS	Central Nervous System
CNNs	Convolutional Neural Networks
CpG	Cytosine-Phosphate-Guanine
CSF	Cerebrospinal Fluid
CT	Computed Tomography
DL	Deep Learning
DNA	Deoxyribonucleic Acid
DQN	Deep Q-Networks
DTI	Diffusion Tensor Imaging
DWI	Diffusion-Weighted Imaging
FID	Free Induction Decay
FLAIR	Fluid Attenuated Inversion Recovery
fMRI	Functional MRI
FP-Tree	Frequent Pattern Tree
GBM	Glioblastoma
GLCM	Gray Level Co-Occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
GOSS	Gradient-based One-Side Sampling
GPU	Graphics Processing Unit
IDH	Isocitrate Dehydrogenase

Abbreviation	**Definition**
LightGBM	Light Gradient Boosting Machine
MGMT	O-6-Methylguanine-DNA Methyltransferase
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NGTDM	Neighboring Gray Tone Difference Matrix
NLL	Negative Log-Likelihood
NMR	Nuclear Magnetic Resonance
OLS	Ordinary Least Squares
PCR	Polymerase Chain Reaction
PET	Positron Emission Tomography
ReLU	Rectified Linear Unit
RF	Radiofrequency
RL	Reinforcement Learning
RNNs	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-Sampling Technique
SVMs	Support Vector Machines
T1CE	Contrast-Enhanced T1
T1W	T1-weighted
T2W	T2-weighted
TD	Temporal-Difference
TE	Echo Time
TMZ	Temozolomide
TR	Repetition Time
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

Introduction

Glioblastoma (GBM), classified by the World Health Organization (WHO) as a grade IV astrocytoma, stands as the most prevalent and aggressive form of primary malignant brain tumor in adults, distinguished by rapid progression and poor prognosis. Its marked molecular heterogeneity has made biomarker discovery central to advancing patient management. Among these biomarkers, the methylation status of the O-6-methylguanine-DNA methyltransferase (MGMT) gene promoter plays a pivotal role, serving both as a prognostic indicator and as a predictor of therapeutic response. Specifically, MGMT promoter methylation is strongly associated with increased sensitivity to temozolomide chemotherapy, thereby enabling more personalized and effective treatment strategies tailored to individual patients.

Conventionally, determining MGMT status requires invasive procedures such as biopsy or surgical resection, which pose risks and may delay critical treatment decisions. As a result, non-invasive methods for predicting MGMT status using magnetic resonance imaging (MRI) have garnered significant attention. Multi-modal MRI, encompassing T1-weighted (T1W), T2-weighted (T2W), contrast-enhanced T1 (T1CE), and fluid-attenuated inversion recovery (FLAIR) sequences, captures vital tumor characteristics, making it well-suited for in-depth analysis.

Machine learning (ML) has transformed medical image analysis by enabling automated recognition, segmentation, and classification of complex imaging patterns. In the realm of MRI, ML algorithms utilize annotated datasets to distinguish tumor regions from healthy tissue, detect subtle variations in texture or intensity, and classify tumors based on molecular or clinical characteristics. Traditional supervised methods, including support vector machines (SVMs), random forests, and gradient boosting, are widely used for image classification tasks, while deep learning (DL) approaches, such as convolutional neural networks (CNNs), automatically extract hierarchical features from raw images, delivering highly accurate tumor detection and reliable biomarker prediction. These methods enable precise, reproducible, and non-invasive evaluation of biomarkers, including MGMT promoter methylation status, from multi-modal MRI data.

A pivotal advancement in ML-based image analysis is radiomic feature extraction, which transforms complex imaging data into robust, noise-resistant representations. Radiomics captures subtle phenotypic differences that correlate with molecular biomarkers, enhancing the

predictive power of ML models. Developing reliable ML models for radiomic-based medical image analysis typically involves the following tasks:

- Acquire suitable medical imaging modalities and extract radiomic features to quantitatively describe tumor characteristics.
- Preprocess the dataset by cleaning and imputing missing values to ensure data quality.
- Perform feature selection to retain the most informative variables, reduce dimensionality, and improve model performance.
- Standardize features to maintain consistency across the dataset.
- Split the dataset into training and testing subsets for model development and evaluation.
- Address class imbalances in the training data using balancing strategies and apply data augmentation to increase sample diversity.
- Train machine learning models on the processed features and evaluate their performance on the test set.
- Assess performance using metrics such as accuracy, sensitivity, specificity, and the area under the ROC curve (AUC-ROC).

These steps ensure robust and generalizable ML models capable of leveraging radiomic features for precise medical image analysis.

Thesis Goals and Results

This thesis focuses on the application of machine learning techniques for knowledge extraction from multi-modal MRI data to facilitate the classification of gliomas. Specifically, it proposes a non-invasive ML framework for predicting MGMT promoter methylation status from multi-modal MRI, supporting personalized treatment strategies for GBM patients. The proposed pipeline leverages radiomic feature extraction from the Brain Tumor Segmentation (BraTS) 2021 dataset, employs feature selection using LightGBM and CatBoost, and performs classification via a voting-ensemble and a stacking-ensemble of ML models.

The framework's performance was evaluated using key metrics: accuracy, precision, sensitivity, specificity, and AUC. The voting-ensemble model achieved an accuracy of 92.86% and an AUC of 96.84%, while the stacking-ensemble model recorded an accuracy of 88.10% and an AUC of 97.01%. These results demonstrate strong alignment with clinical outcomes, outperforming traditional approaches. The findings underscore the potential of integrating multi-modal MRI radiomics with machine learning for precise biomarker prediction, advancing precision medicine in glioblastoma management.

Thesis Outline

Chapter 1: Background This chapter provides foundational knowledge on brain anatomy, pathology, and imaging. It covers cerebral structure, including lobes and functional regions, followed by an overview of brain tumors and their clinical significance. Standard imaging planes (axial, sagittal, coronal) are introduced, alongside key MRI modalities (T1W, T1CE, T2W, FLAIR) and their role in tumor detection and treatment planning.

Chapter 2: Supervised Machine Learning Overview This chapter surveys supervised machine learning algorithms, beginning with classification methods such as perceptrons, support vector machines, logistic regression, Naïve Bayes, and decision trees. It then reviews ensemble techniques including bagging, stacking, voting, and boosting that enhance predictive accuracy, before addressing regression approaches like linear, Lasso, and Ridge regression, and concluding with a comparative analysis of their performance.

Chapter 3: State of the Art on MGMT Prompter Methylation This chapter reviews the molecular and clinical significance of MGMT promoter methylation in GBM. It discusses diagnostic methodologies, including tissue-based assays and non-invasive imaging-based approaches. The chapter examines radiomics, multiparametric MRI, Positron Emission Tomography (PET) fusion, and advanced ML techniques, particularly deep learning, while addressing challenges and future directions in MGMT methylation prediction.

Chapter 4: Proposed Approaches MGMT Detection This chapter presents the proposed approaches for MGMT detection, detailing the pipeline from dataset preparation and feature engineering to model training and evaluation. Various classifiers, including Random Forest, LightGBM, XGBoost, CatBoost, and ensemble strategies, are explored to establish a robust framework for accurate and reliable prediction.

Chapter 5: Results and Discussions This chapter presents the results of the proposed approaches for MGMT promoter methylation detection, covering radiomics feature selection, feature distribution, and machine learning model evaluation. Performances of base learners and ensembles are compared, highlighting the most effective strategies and their potential clinical relevance for non-invasive glioma characterization.

Conclusion The final section summarizes the findings, discusses their implications, and proposes directions for future research in non-invasive MGMT status prediction.

Academic Publication

- Y. Boulkhiout, A. Moussaoui, and K. Nasri. “Enhancing Machine Learning Approach for MGMT Promoter Methylation Detection in Glioma from MRI Features.” *International Journal of Computational and Experimental Science and Engineering (IJCESEN)*, 11(3):5768–5776, 2025 [5].

CHAPTER 1 Fundamentals of Brain Anatomy, Tumors, and Imaging

I.1 Introduction

This chapter provides the foundational knowledge required to understand the structural and pathological aspects of the brain in relation to medical imaging. It begins with an overview of cerebral anatomy, describing the Central Nervous System (CNS), the major brain tissues, and the ventricular system. This anatomical framework is essential for interpreting neurological alterations and serves as a basis for studying tumor development and progression.

The chapter then turns to brain tumors, presenting their classification into benign and malignant types, along with key aspects of diagnosis, management, and epidemiology. Finally, it introduces medical imaging with a focus on Magnetic Resonance Imaging (MRI), discussing its physical principles, image contrast mechanisms, artifacts, and clinical applications. Together, these sections establish the background necessary for understanding the role of imaging in brain tumor analysis and for supporting the methodological developments in the subsequent chapters.

I.2 Cerebral Anatomy Overview

This section examines the structural components vital to neurological function: the central nervous system, which processes sensory information and coordinates responses; major brain tissues, crucial for cognition and motor functions; and the ventricular system, which circulates cerebrospinal fluid to cushion the brain, remove waste, and regulate intracranial pressure, collectively supporting neural communication and brain health.

I.2.1 Central Nervous System

The central nervous system forms a complex communication network responsible for regulating physiological processes and behavior [1]. Structurally, it is composed of white and gray matter and consists of three main components: the spinal cord, brainstem, and brain. These components interact dynamically to process sensory input, generate motor responses, and coordinate higher-order cognitive functions.

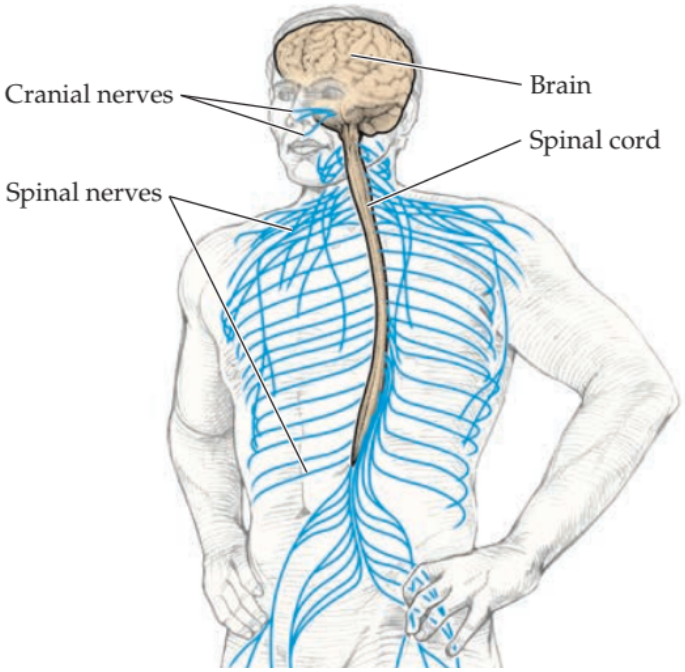


Figure I.1: The Central Nervous System [1].

1. **Brain**: The brain is the most prominent and complex structure of the CNS. It comprises two cerebral hemispheres connected by interhemispheric structures, such as the corpus callosum, which facilitates communication between the hemispheres. The brain is enclosed within the cranial cavity but suspended in a protective layer of cerebrospinal fluid (CSF). Its tissue is divided into grey matter, white matter, and cerebrospinal fluid, each with distinct structural and functional properties contributing to the brain's overall function.

   The brain is organized into interconnected regions with specialized roles. The telencephalon ("end brain") includes the cerebral hemispheres, comprising the cerebral cortex, white matter, and basal ganglia, which support cognition, sensory perception, and motor control. The diencephalon contains the thalamus, a relay for sensory and motor signals, and the hypothalamus, which regulates homeostasis and autonomic functions. The midbrain links the forebrain to the hindbrain, which consists of the metencephalon (pons and cerebellum) and the myelencephalon (medulla). Together, the midbrain,

pons, and medulla form the brainstem, a vital structure connecting the forebrain to the spinal cord and controlling essential functions such as breathing, circulation, and heart rate [1].

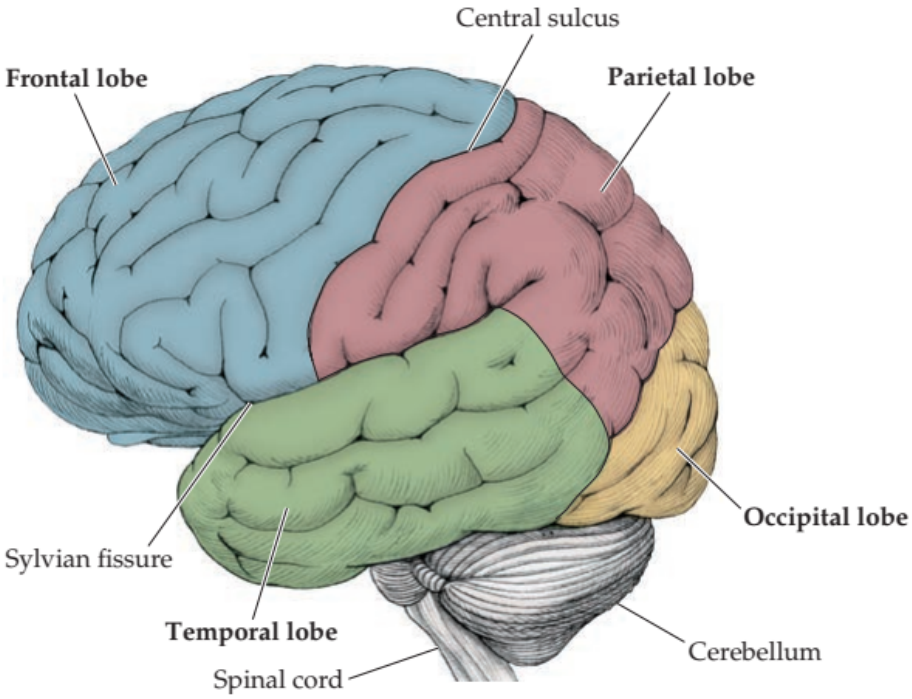


Figure I.2: Lateral View of the External Parts of Adult CNS.

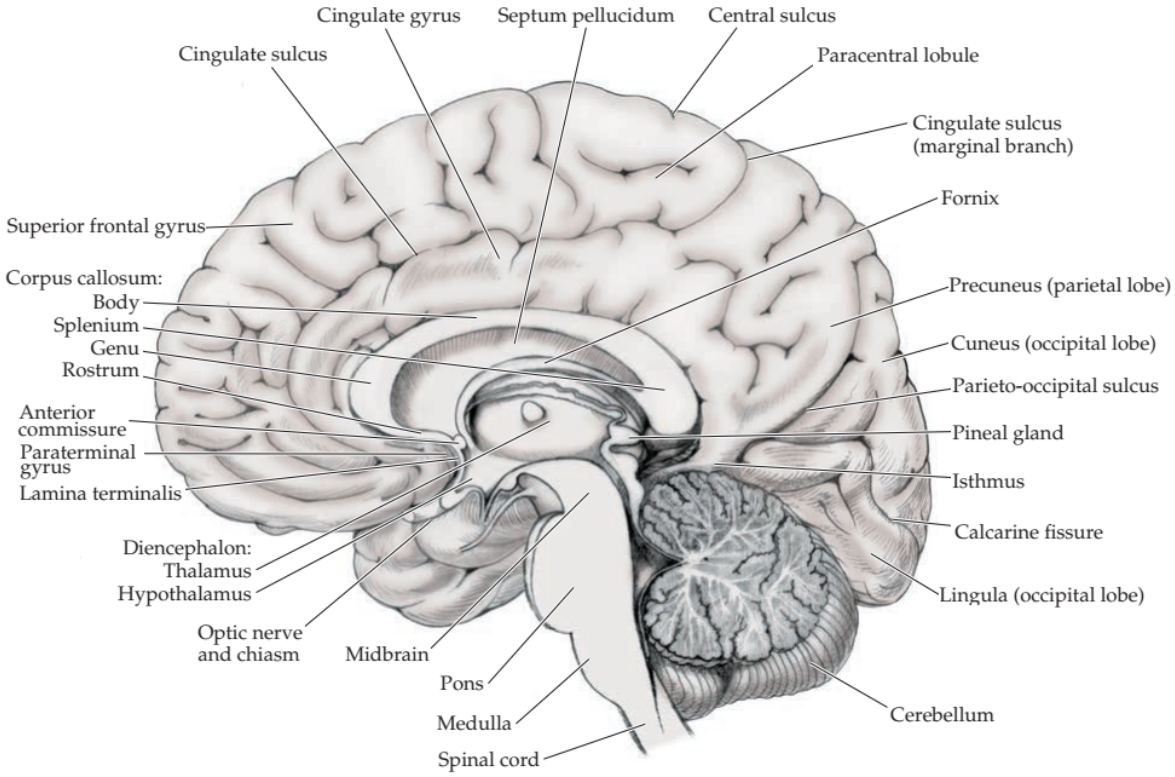


Figure I.3: Lateral View of the Internal Parts of Adult CNS.

2. **Brainstem**: The brainstem includes the medulla oblongata, pons, and midbrain. It plays a critical role in autonomic control and sensory-motor integration. The brainstem also houses various nuclei involved in essential life functions such as respiration and cardiac regulation. Attached to it is the cerebellum, responsible for coordinating voluntary movements and maintaining balance. The cerebellum is characterized by a highly folded surface and a thin cortical layer of grey matter, enhancing its computational capacity in motor control [1].

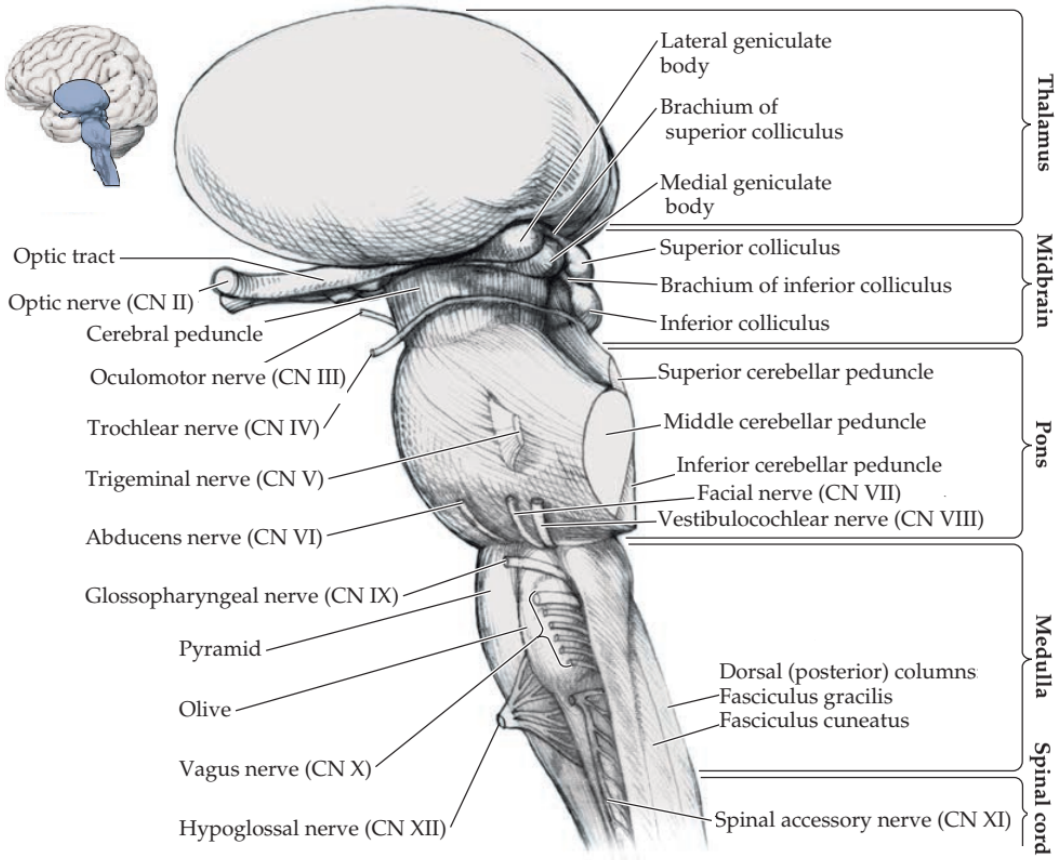


Figure I.4: Lateral View of Brainstem and Cranial.

3. **Spinal Cord and Peripheral Nervous System**: The spinal cord is a cylindrical structure housed within the vertebral canal and protected by meningeal layers. It serves as a conduit for transmitting signals between the brain and the peripheral nervous system. Structurally, it contains white matter (composed of axonal pathways) and grey matter, arranged in an H-shaped configuration. At its center lies the central canal, which connects with the brain's ventricular system and contains CSF that facilitates nutrient transport and provides mechanical protection .

   The human nervous system develops in a segmented pattern, resembling that of simpler organisms like segmented worms. In the head, these segments expand and fuse to form the cerebral hemispheres and brainstem. The spinal cord retains its segmented

structure, giving rise to spinal nerves. Each spinal segment generates paired sensory and motor nerve roots on both sides of the body. Motor systems are typically located ventrally (anteriorly), while sensory systems are positioned dorsally (posteriorly), an organization consistently maintained throughout the spinal cord [1]. Dorsal nerve roots serve as the main pathways for transmitting afferent sensory information from the body into the dorsal region of the spinal cord, allowing the CNS to receive and process external and internal stimuli. In contrast, ventral nerve roots carry efferent motor signals from the ventral spinal cord toward the periphery, enabling voluntary movement and autonomic responses. Each spinal segment and its associated nerve roots are named according to the vertebral level at which they exit the bony vertebral canal, providing an anatomical basis for both clinical localization and neurological assessment.

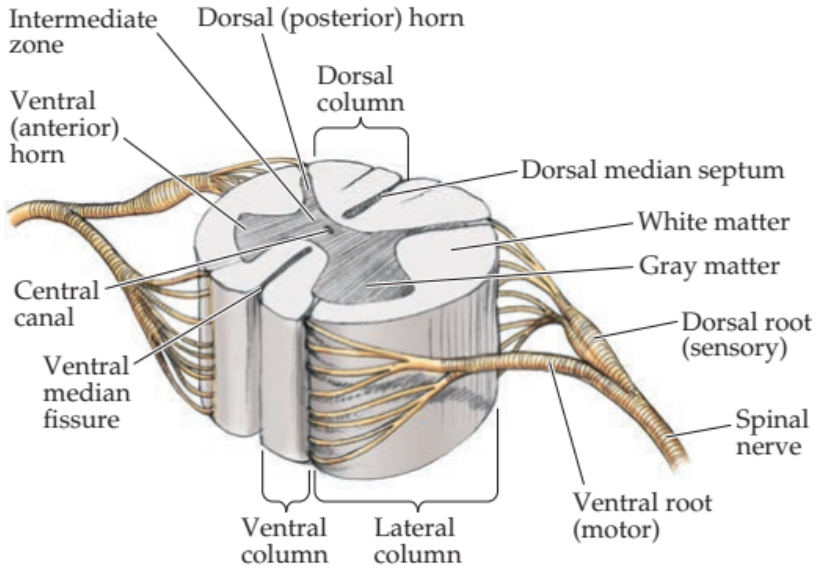


Figure I.5: Spinal Cord and Peripheral Nervous System.

I.2.2 Major Brain Tissues

1. **Cerebrospinal Fluid**: During development, the neural tube cavities form the ventricular system, which contains cerebrospinal fluid produced mainly by the choroid plexus. The lateral ventricles connect to the third ventricle via the foramina of Monro, and the third communicates with the fourth through the cerebral aqueduct. From the fourth ventricle, CSF enters the subarachnoid space through the foramina of Magendie and Luschka, circulating around the brain and spinal cord before being reabsorbed into the venous system by arachnoid granulations. In adults, CSF volume averages 150 mL with a daily production of about 500 mL. This clear fluid cushions the central nervous system, protects it from mechanical injury, and supports metabolic exchange to maintain homeostasis [1].

2. **Grey Matter**: primarily composed of neuronal cell bodies, plays a central role in processing and integrating neural information within the central nervous system. It is found in several regions, including:

    - **Cerebral cortex**: the outermost layer of the brain, essential for higher-order cognitive processes including reasoning, memory, perception, language, and voluntary motor control.
    - **Basal nuclei**: a group of subcortical structures that play a central role in regulating motor activity, coordinating movement, and supporting procedural learning and habit formation.
    - **Brainstem** and **spinal cord nuclei**: clusters of neurons that govern fundamental reflexes and autonomic responses, ensuring vital functions such as respiration, cardiovascular regulation, and posture control.

    The cerebral cortex, the outermost layer of the brain, is highly convoluted to maximize its surface area, with an average thickness of 2.6 mm and up to 4.5 mm in certain regions. It is organized into four lobes (frontal, parietal, temporal, and occipital) defined by prominent sulci and gyri that serve as anatomical landmarks [1].

    The basal nuclei, comprising the caudate nucleus and lentiform nucleus (putamen and globus pallidus), play a key role in regulating voluntary movement and are implicated in disorders such as Parkinson's disease. The thalamus functions as a critical relay station for sensory information, while the hypothalamus is central to endocrine and autonomic functions, contributing to the regulation of bodily homeostasis.

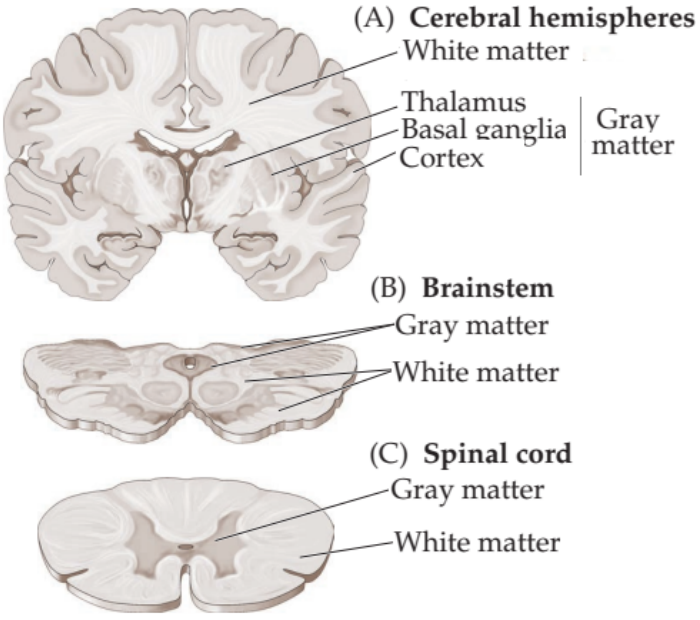


Figure I.6: White and Gray Matters.

3. **White Matter**: White matter consists primarily of bundles of myelinated axons that establish communication networks both within the brain and between the brain and the rest of the body. These axonal tracts are highly organized into distinct fiber systems:

    - **Association fibers**: interconnect cortical regions within the same cerebral hemisphere, supporting the integration of information across functional areas.
    - **Commissural fibers**: link corresponding regions of the two hemispheres, with the corpus callosum being the largest and most prominent commissural pathway, essential for interhemispheric communication.
    - **Projection fibers**: extend vertically to connect the cerebral cortex with subcortical structures, the brainstem, and the spinal cord, thereby transmitting motor commands outward and conveying sensory inputs inward.

    Collectively, these fiber systems form the structural backbone of neural connectivity, enabling the integration of sensory processing, the coordination of voluntary and involuntary motor activity, and the execution of higher cognitive functions [1].

I.2.3 Ventricular System

The brain's ventricular system is a series of interconnected cavities (the lateral, third, and fourth ventricles) through which cerebrospinal fluid circulates via the foramina of Monro and the cerebral aqueduct. From the fourth ventricle, CSF enters the spinal canal and subarachnoid space, where it cushions neural tissue, regulates intracranial pressure, and aids in waste clearance [1].

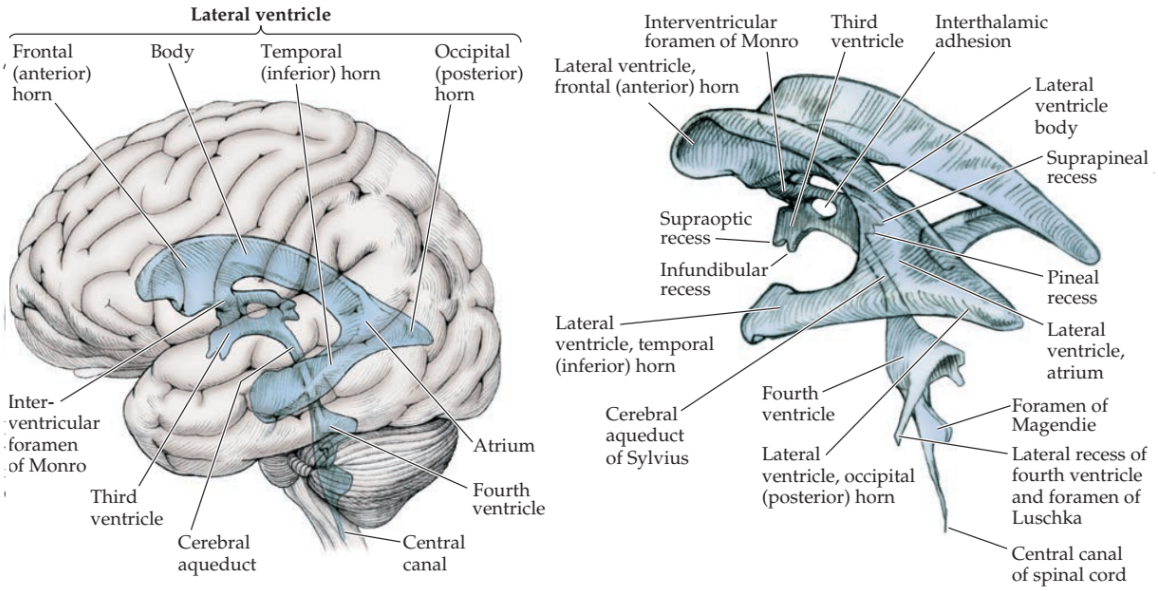


Figure I.7: Ventricular System [1].

I.3 Brain Tumors Overview

Brain tumors are abnormal cell proliferations arising within or near the central nervous system. They are classified as primary, originating in brain parenchyma, meninges, or other intracranial structures, or secondary, representing metastases from systemic cancers. Tumors may be benign, with slower growth and limited invasion, or malignant, marked by aggressive proliferation and recurrence, with about one-third being malignant [6]. Regardless of type, both can impair neurological function by compressing brain tissue, altering neural pathways, reducing blood flow, and affecting cranial nerves [7].

Primary brain tumors are rare, affecting approximately 23 per 100,000 people annually, but their impact is profound due to the CNS's critical role in regulating bodily functions [6]. They occur across all ages, with certain types (e.g., medulloblastomas) common in children and others (e.g., glioblastomas) in adults aged 50–70. Risk factors include genetic syndromes (e.g., neurofibromatosis) and radiation exposure, though most cases are sporadic [8]. Symptoms, such as headaches, seizures, cognitive changes, or motor deficits, vary by tumor location and size [9].

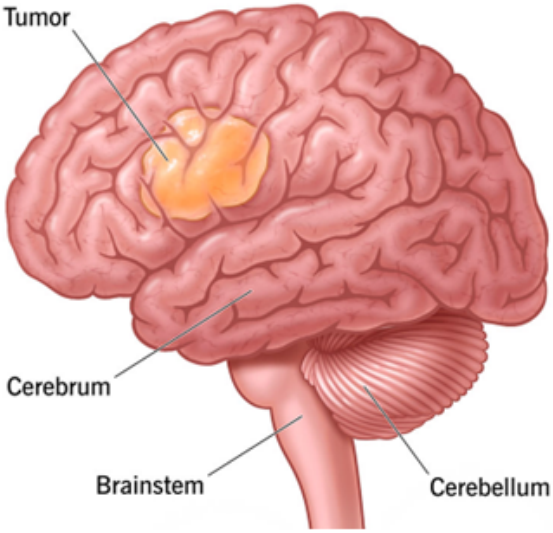


Figure I.8: Brain Tumors.

Diagnosis relies on MRI or Computed Tomography (CT, an X-ray-based imaging method) scans, often supplemented by biopsy for detailed analysis [10]. Treatment typically involves surgery, radiation, and chemotherapy, but challenges like tumor heterogeneity and the blood-brain barrier limit effectiveness [11]. For example, glioblastomas, a highly aggressive tumor type, have a median survival of 12–18 months despite intensive therapy [12]. Advances in molecular profiling, guided by the World Health Organization, are improving personalized treatment approaches [13]. This chapter explores primary brain tumors, their features, diagnosis, and treatment, providing a foundation for understanding these complex neoplasms [7].

I.3.1 Types of Brain Tumors

Brain tumors are diverse, with over 150 identified types, categorized based on their cellular origin (glial or non-glial) and malignancy status [13]. Below, we outline the major types of benign and malignant primary brain tumors.

I.3.1.1 Benign Brain Tumors

Benign brain tumors are typically slow-growing and noncancerous, though their location can complicate treatment. They are characterized by slow growth and a lower likelihood of metastasis compared to malignant tumors. However, their location within the brain or spinal cord can lead to significant neurological complications, making diagnosis and treatment challenging. These tumors vary widely in their origin, behavior, and impact on surrounding structures. Below is a detailed overview of key types of benign brain tumors, their characteristics, and associated clinical considerations [7]:

- **Chordomas**: These rare, slow-growing tumors typically arise at the skull base or lower spine, particularly in the clivus or sacrum. Originating from remnants of the notochord, chordomas are generally benign but can be locally invasive, complicating surgical resection due to their proximity to critical structures like the brainstem or spinal cord. Symptoms may include headaches, vision changes, or lower back pain, depending on the tumor's location. Their rarity and challenging location often necessitate advanced imaging and multidisciplinary treatment approaches, including surgery and proton beam therapy [14].
- **Craniopharyngiomas**: These tumors develop near the pituitary gland, often involving the hypothalamus or optic chiasm, which complicates complete surgical removal. Predominantly benign, craniopharyngiomas are more common in children and older adults, presenting with symptoms such as hormonal imbalances, vision loss, or increased intracranial pressure. Their deep brain location and adherence to critical structures often result in partial resection, requiring adjuvant therapies like radiation. Long-term management is complex due to potential endocrine and neurological deficits [7].
- **Gangliocytomas, Gangliomas, and Anaplastic Gangliogliomas**: These rare tumors originate from neuronal or mixed neuronal-glial cells and are typically benign, though anaplastic gangliogliomas may exhibit more aggressive behavior. Gangliocytomas and gangliomas are most common in children and young adults, often presenting with seizures due to their cortical location. Complete surgical resection is the primary treatment, but the rarity of these tumors and their potential for malignant transformation (in anaplastic cases) necessitate careful histopathological evaluation and long-term follow-up [13].

- **Glomus Jugulare**: These tumors, also known as paragangliomas, arise from paraganglion cells near the jugular vein at the skull base. Typically benign, they are slow-growing but can cause significant symptoms, including hearing loss, tinnitus, or cranial nerve deficits, due to their proximity to critical neurovascular structures. Treatment is challenging and may involve surgery, radiation, or observation for asymptomatic cases, depending on tumor size and patient symptoms. Their vascular nature requires preoperative embolization in some cases to reduce surgical risks [7].

- **Meningiomas**: Representing the most common primary brain tumors, meningiomas originate in the meninges, the protective layers surrounding the brain and spinal cord. Approximately 90% of meningiomas are benign, growing slowly and often remaining asymptomatic for years. However, their location (e.g., along the falx cerebri or convexity) can cause symptoms like headaches, seizures, or focal neurological deficits. Rare malignant or atypical meningiomas may exhibit more aggressive behavior. Treatment options include observation for small, asymptomatic tumors, surgical resection, or stereotactic radiosurgery for inoperable cases. The high prevalence and generally favorable prognosis underscore the importance of early detection and tailored management [7].

- **Pineocytomas**: These slow-growing tumors arise in the pineal gland, which regulates melatonin production and circadian rhythms. Pineocytomas are rare, typically affecting young adults, and may cause symptoms such as hydrocephalus (due to compression of the cerebral aqueduct), headaches, or Parinaud's syndrome (impaired upward gaze). Surgical resection is the primary treatment, but the deep location of the pineal gland poses technical challenges. Stereotactic radiosurgery is an alternative for small or residual tumors. Their benign nature generally confers a good prognosis with appropriate management [7].

- **Pituitary Adenomas**: These tumors are tumors of the pituitary gland, a key regulator of endocrine function. Most are benign and slow-growing, classified as microadenomas ($\leq$ 10 mm) or macroadenomas ($\geq$ 10 mm). They may be functional, producing hormone-related disorders such as hyperprolactinemia, Cushing's disease, or acromegaly, or nonfunctional, often presenting with mass effects like optic chiasm compression and vision loss. Treatment varies by type and includes medical therapy (e.g., dopamine agonists for prolactinomas), surgery, or radiotherapy for refractory cases. Because of their impact on hormonal balance, long-term endocrine follow-up is usually required [7].

- **Schwannomas**: Schwannomas are benign peripheral nerve sheath tumors arising from Schwann cells, which form the myelin sheath of peripheral nerves. The most common subtype, vestibular schwannoma (acoustic neuroma), involves the eighth cranial nerve

and often causes hearing loss, tinnitus, and balance problems. These slow-growing tumors are managed with observation, surgery, or stereotactic radiosurgery depending on size and symptoms. Although outcomes are generally favorable, preservation of cranial nerve function, particularly hearing and facial nerve integrity, remains a major surgical concern [7].

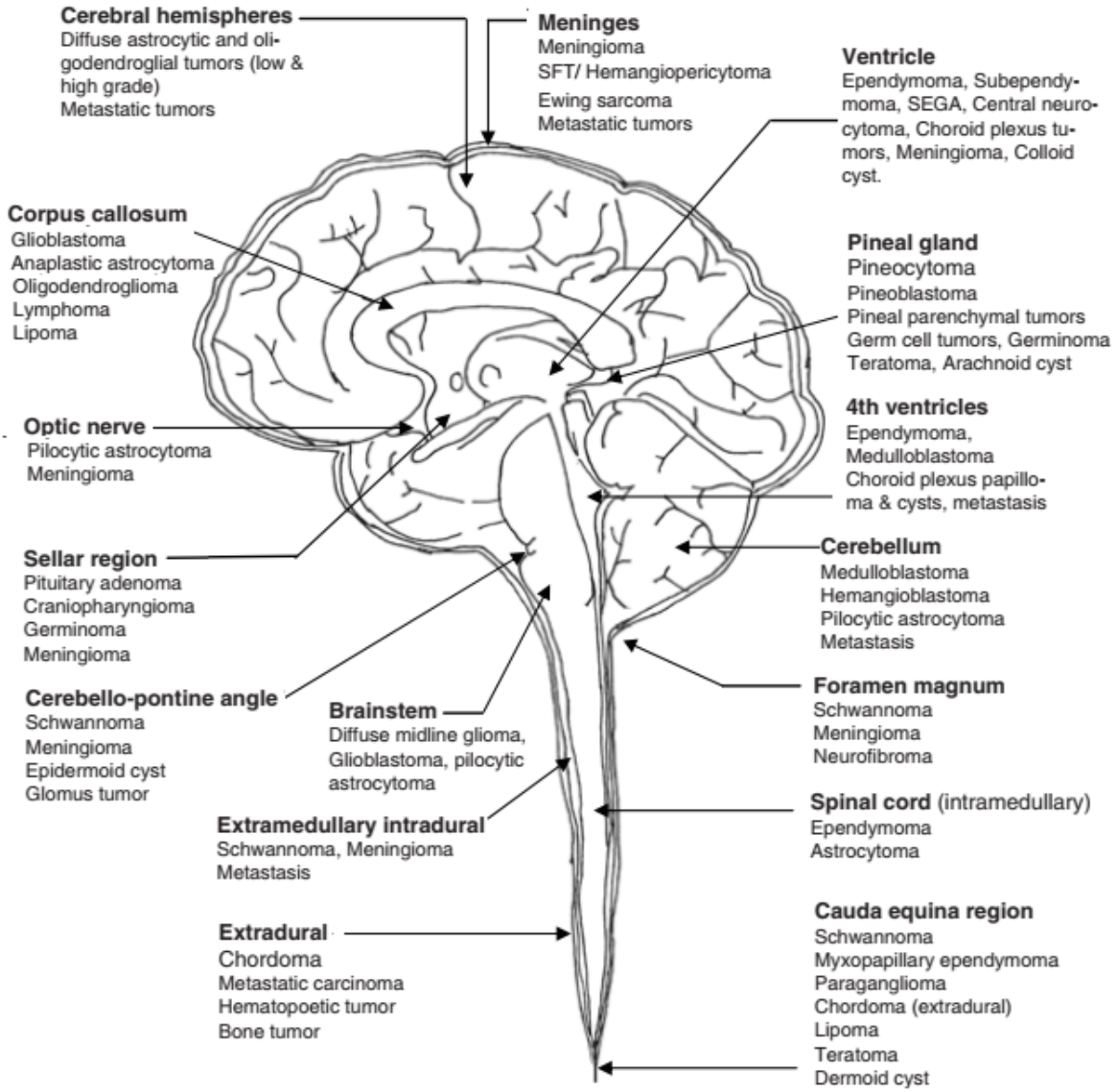


Figure I.9: Anatomical Distribution of Major Brain Tumors [2]

This section offers a detailed overview of benign brain tumors, underscoring their clinical relevance, diagnostic complexity, and therapeutic implications. By examining the distinct biological and anatomical features of each tumor type, it highlights how these differences influence clinical presentation, treatment planning, and long-term outcomes. The discussion emphasizes the necessity of individualized management strategies in neuro-oncology, where patient-specific factors and tumor characteristics must guide decision-making to optimize

prognosis and preserve neurological function [7].

I.3.1.2 Malignant Brain Tumors

Malignant brain tumors represent a significant subset of primary brain tumors, with gliomas constituting approximately 78% of all cancerous primary brain tumors [6]. These tumors originate from various types of glial cells, which support and protect neurons in the central nervous system. Gliomas are particularly challenging due to their infiltrative nature, rapid growth, and resistance to conventional treatments [11].

Below is an overview of the key types of gliomas, each characterized by distinct cellular origins, growth patterns, and clinical implications:

- **Astrocytomas**: These are the most prevalent type of glioma, originating from astrocytes, the star-shaped glial cells that provide both structural and metabolic support to neurons. Primarily found in the cerebral hemispheres, these tumors exhibit a wide spectrum of malignancy, from low-grade variants such as pilocytic astrocytomas to highly aggressive high-grade forms. Clinical manifestations are largely determined by tumor location and size, commonly including seizures, headaches, and various neurological deficits [13].

- **Glioblastomas**: As the most aggressive subtype of astrocytomas, glioblastomas are characterized by rapid growth, extensive infiltration, and poor prognosis. They often exhibit necrosis and microvascular proliferation, making them resistant to therapy. Glioblastomas primarily affect adults and are associated with a median survival of approximately 12–18 months despite multimodal treatment involving surgery, radiation, and chemotherapy [12].

- **Ependymomas**: Originating from ependymal cells lining the ventricles of the brain and the central canal of the spinal cord, ependymomas are more common in children and young adults. These tumors can obstruct cerebrospinal fluid flow, leading to hydrocephalus and increased intracranial pressure. Their proximity to critical brain structures complicates surgical resection [7].

- **Oligodendrogliomas**: These rare tumors arise from oligodendrocytes, the myelin-producing cells that insulate neuronal axons. Often found in the cerebral hemispheres, oligodendrogliomas tend to grow more slowly than other gliomas and are frequently associated with specific genetic markers, such as 1p/19q co-deletion, which can predict better response to treatment. They are more common in adults and may present with seizures [13].

- **Medulloblastomas**: Predominantly affecting children, medulloblastomas are fast-growing tumors typically located in the cerebellum at the base of the skull. They are the most common malignant brain tumor in pediatric populations and often spread through the cerebrospinal fluid to other parts of the brain or spinal cord. Advances in molecular subtyping have identified distinct subgroups, which influence prognosis and treatment strategies [13].

The management of gliomas remains a significant clinical challenge due to their heterogeneity, aggressive behavior, and limited response to therapies [11]. Ongoing research into molecular profiling, targeted therapies, and immunotherapies holds promise for improving outcomes in patients with these devastating tumors. The WHO classification (2021) categorizes brain tumors based on molecular and histological features:

- Gliomas (astrocytomas, oligodendrogliomas, glioblastomas).
- Neuronal and mixed neuronal-glial tumors (gangliogliomas, dysembryoplastic neuroepithelial tumors).
- Meningiomas (benign, atypical, malignant).
- Embryonal tumors (medulloblastomas, atypical teratoid/rhabdoid tumors) .

Recent genomic studies highlight key mutations driving tumorigenesis: IDH1/2 mutations (Isocitrate Dehydrogenase, lower-grade gliomas), MGMT promoter methylation (predicts chemosensitivity in glioblastoma), and BRAF V600E mutation (gene frequently observed in pilocytic astrocytomas and gangliogliomas) [13].

I.3.2 Diagnosis

Diagnosing brain tumors involves a multi-step process, usually triggered by symptoms like headaches, seizures, or neurological deficits. A neurological exam evaluates functions such as balance, cognition, vision, and reflexes to detect abnormalities [10]. Based on the findings, imaging and other tests confirm the tumor's presence, type, location, and malignancy. Early, accurate diagnosis is vital, especially for aggressive tumors like glioblastomas, as delays can worsen outcomes [11]. Key diagnostic tests include:

- **Brain MRI or CT Scan**: MRI, an imaging technique using magnetic fields to visualize detailed brain structures, is the gold standard for identifying tumor size, location, and characteristics, often enhanced with a contrast agent (e.g., gadolinium) to highlight abnormal tissue. CT, an X-ray-based imaging method, is a faster alternative used when MRI is contraindicated (e.g., for patients with pacemakers) or unavailable, though it provides less detail. Both techniques can detect mass effect, edema, or CSF obstruction [15].

- **Biopsy**: A sample of tumor tissue is obtained and examined microscopically to determine the tumor's type (e.g., glioma, meningioma) and malignancy grade (e.g., WHO Grade I–IV). Biopsies may be performed during surgical resection or via stereotactic biopsy, a minimally invasive procedure using 3D imaging to target hard-to-reach tumors, reducing risks to eloquent brain areas like those controlling speech [14].

- **Spinal Tap (Lumbar Puncture)**: This procedure analyzes CSF for cancer cells, proteins, or other markers, particularly when meningeal involvement or tumor spread (e.g., leptomeningeal disease) is suspected. It is less common but valuable for tumors like medulloblastomas, which may metastasize through CSF pathways [7].

- **Specialized Tests**: Blood or CSF tests for tumor markers (e.g., alpha-fetoprotein for germ cell tumors) or gene abnormalities (e.g., IDH1/2 mutations, 1p/19q co-deletion) aid in diagnosis and prognosis. Advanced techniques, such as liquid biopsy, analyze circulating tumor DNA in blood or CSF, offering a non-invasive method to detect molecular signatures, especially for recurrent tumors [13].

Accurate diagnosis combines imaging and tissue analysis. While MRI/CT localizes tumors, biopsy confirms histology and molecular features (e.g., IDH status for treatment planning). Challenges include differentiating tumors from mimics (e.g., abscesses) and sampling deep lesions. Emerging techniques like positron emission tomography, magnetic resonance spectroscopy, and AI-enhanced imaging improve detection and characterization [16]. Molecular profiling (e.g., MGMT methylation) now guides WHO classification and therapy selection [13]. The diagnostic strategy balances precision with minimal invasiveness, with ongoing advances enhancing detection and personalization [10].

I.3.3 Management and Treatment

Treatment strategies for brain tumors depend on tumor type, size, location, and patient health [14]. Common approaches include:

- **Surgery**: This involves opening the skull to remove the tumor, aiming to maximize resection while preserving critical brain functions, such as motor skills or speech. Awake craniotomy, where the patient remains conscious during parts of the surgery, may be used to map and avoid eloquent brain areas, reducing neurological deficits [14].

- **Radiation Therapy**: High-dose X-rays or other radiation types are used to destroy or shrink tumors by damaging cancer cell DNA. In pediatric patients, radiation is approached cautiously due to risks of developmental impairments, such as cognitive delays, necessitating alternative strategies like proton therapy for precision [7].

- **Radiosurgery**: This non-invasive technique delivers highly focused radiation (e.g., gamma rays or proton beams) to the tumor, sparing healthy tissue. It is often used for small or inoperable tumors, such as metastases or acoustic neuromas, with systems like Gamma Knife or CyberKnife [7].
- **Brachytherapy**: Radioactive implants are placed directly near or within the tumor to deliver localized radiation, minimizing damage to surrounding brain tissue. This is less common but effective for specific tumors, such as recurrent gliomas [7].
- **Chemotherapy**: Drugs, administered orally or intravenously, target rapidly dividing cancer cells, often used post-surgery to eliminate residual tumor cells. Agents like temozolomide are standard for glioblastomas, though the blood-brain barrier limits drug delivery, reducing efficacy [17].
- **Immunotherapy**: This enhances the body's immune system to recognize and attack cancer cells, using approaches like checkpoint inhibitors or vaccines. It is an emerging option for brain tumors, with ongoing trials for glioblastomas showing promise but limited widespread use [11].
- **Targeted Therapy**: Drugs like bevacizumab target specific molecular features of cancer cells (e.g., vascular endothelial growth factor in glioblastomas), minimizing harm to healthy tissue. These therapies rely on molecular profiling to identify actionable mutations, such as IDH1/2 in low-grade gliomas [13].
- **Watchful Waiting**: Small, asymptomatic tumors, such as low-grade meningiomas, may be monitored with regular imaging (e.g., MRI) rather than immediate intervention, especially in elderly patients or those with significant health risks [7].

Emerging approaches are addressing these challenges. Tumor-treating fields, a non-invasive therapy using electric fields to disrupt cancer cell division, have shown efficacy in extending survival for glioblastoma patients [12]. Advances in molecular profiling, such as identifying MGMT promoter methylation, guide personalized treatments, improving outcomes for specific tumor subtypes [13]. Clinical trials exploring combination therapies, such as immunotherapy with targeted agents, are expanding treatment options [11].

I.3.4 Epidemiology

Primary brain tumors are relatively rare, with an incidence of approximately 5 per 100,000 people annually in the United States; global estimates range from 4 to 30 per 100,000, depending on diagnostic capabilities and reporting systems [6]. Among children under 15, around 4,100 are diagnosed yearly with brain or central nervous system tumors, the second most common pediatric cancer after leukemia.

Incidence varies by sex and tumor type: males have a slightly higher overall incidence, whereas meningiomas (typically benign) are more common in females, possibly due to hormonal influences. Glioblastomas, the most aggressive brain tumors, are increasing in prevalence, particularly among adults over 60, reflecting both improved detection and demographic aging [8]. Racial and ethnic disparities also exist; for example, higher rates of certain gliomas are reported in white populations, while meningiomas are more frequently diagnosed in black populations. Socioeconomic factors, such as access to neuroimaging and specialized care, also influence reported incidence, as early detection is more feasible in developed healthcare systems [6].

I.4 Planes for Brain Imaging

In neuroimaging, brain structures are commonly visualized using three orthogonal anatomical planes, each offering a distinct perspective for clinical diagnosis and research analysis [18, 19]. These planes serve as a standardized spatial framework, facilitating consistent image acquisition, interpretation, and comparison across modalities and subjects. They are defined as follows:

- **Axial plane** (transverse plane): Acquired as horizontal slices, oriented parallel to the anterior commissure-posterior commissure line, this view separates the superior (upper) and inferior (lower) portions of the brain. It is widely used to evaluate ventricular size, cortical thickness, and supratentorial lesions.
- **Sagittal plane**: Obtained as vertical slices oriented parallel to the midline, dividing the brain into left and right hemispheres. This plane is optimal for visualizing the corpus callosum, brainstem, and midline structures, and for assessing symmetry.

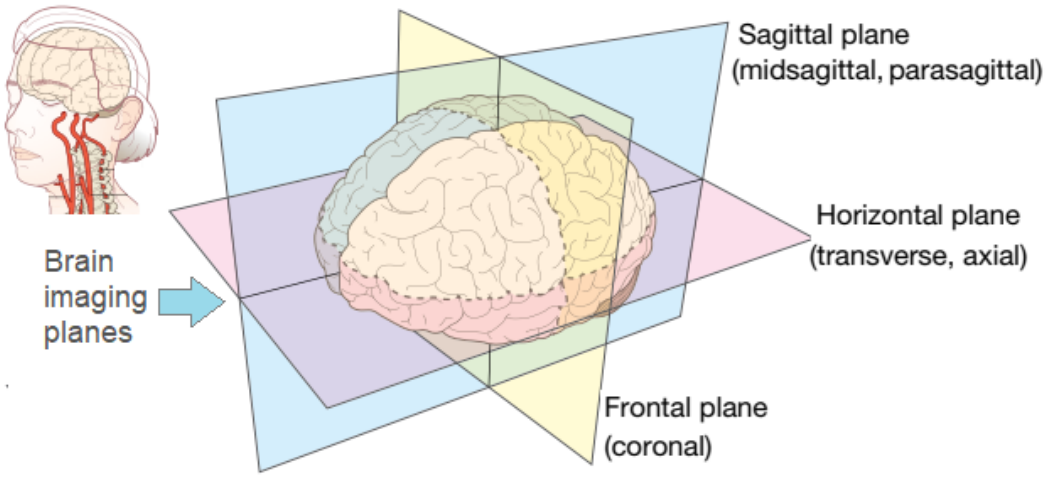
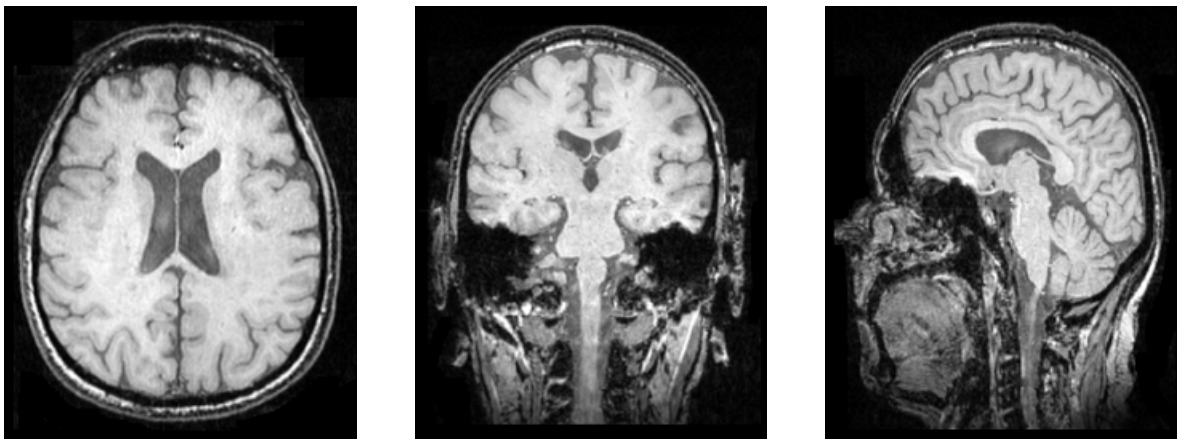


Figure I.10: Standard Anatomical Planes for Brain MRI [3].

- **Coronal plane**: Acquired as vertical slices perpendicular to the sagittal plane, dividing the brain into anterior (front) and posterior (back) sections. It is particularly valuable for examining hippocampal morphology, pituitary lesions, and the extent of cortical or subcortical abnormalities.

  In magnetic resonance imaging (MRI), these planes can be acquired directly during scanning or reconstructed from three-dimensional datasets. The selection of the imaging plane is dictated by clinical indications, lesion location, and the need for multiplanar reformatting in post-processing. This standardized approach allows radiologists and neuroscientists to correlate findings with anatomical landmarks and facilitates integration with stereotactic and functional imaging data [15, 18].



(a) Axial (b) Sagittal (c) Coronal

Figure I.11: Standard 2D Slice Projection Used in MRI.

I.5 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a non-invasive diagnostic technique that utilizes the magnetic properties of atomic nuclei, primarily hydrogen protons, to produce detailed cross-sectional and three-dimensional images of biological tissues. Based on the principles of nuclear magnetic resonance (NMR), MRI provides exceptional soft-tissue contrast, enabling visualization of anatomical structures and physiological processes without the risks of ionizing radiation. Since its development in the early 1970s, credited to Paul Lauterbur for introducing spatial encoding and Raymond Damadian for exploring its diagnostic applications, MRI has transformed clinical diagnostics, neuroscience research, and oncological imaging [20, 21]. Continuous advancements in pulse sequence design, hardware, and image reconstruction algorithms have enhanced its versatility, establishing MRI as a cornerstone of modern medical imaging.

Figure I.12: MRI Scanner, Note the Coil Over the Patient's Head[4].

I.5.1 Fundamental Principles

MRI is based on the principles of NMR, where atomic nuclei with an odd number of protons and/or neutrons act as tiny magnetic dipoles. In clinical MRI, hydrogen nuclei (^{1}H) are favored due to their high abundance in tissues and strong NMR signal. The procedure begins by aligning these nuclear spins in a strong static magnetic field (B_0). A radiofrequency (RF) pulse is then applied to perturb the alignment, tipping the net magnetization away from equilibrium. As the spins relax back, they emit signals that are detected by receiver coils. These signals contain spatial and chemical information that can be reconstructed into detailed anatomical images.

I.5.1.1 Static Magnetic Field and Proton Alignment Mechanism

In the absence of an external magnetic field, nuclear spins are randomly oriented. When a strong static magnetic field (B_0) is applied, the spins align either parallel (low-energy state) or antiparallel (high-energy state) to B_0. Due to a slight excess of spins in the parallel orientation, a net longitudinal magnetization vector M_0 is produced, which serves as the origin of the measurable MRI signal [20]. The magnitude of this equilibrium magnetization can be expressed as:

$$M_0 = \frac{N\gamma^2\hbar^2 I(I+1)B_0}{3k_B T} \tag{I.1}$$

where γ is the gyromagnetic ratio of the nucleus, N the spin density, $\hbar$ the reduced Planck constant, and I the nuclear spin quantum number. The Boltzmann constant, k_B, and the absolute temperature, T, determine the population distribution of spin states. Together, these parameters define the net magnetization of a sample in a magnetic field, which forms the

basis of NMR and MRI signals.

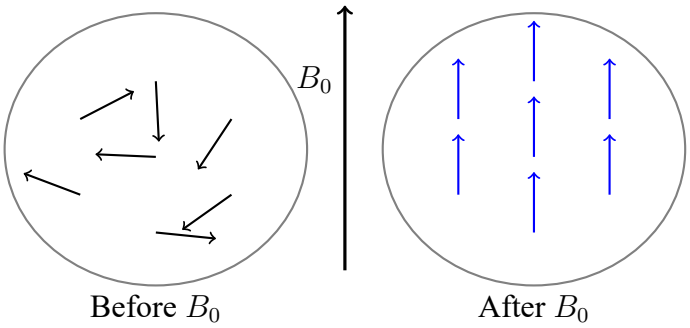


Figure I.13: Random Spin Orientations Become Aligned in Presence of B_0.

I.5.1.2 Larmor Precession and Resonance Frequency

When placed in a strong external magnetic field B_0, nuclear spins exhibit a precessional motion around the field axis at the Larmor frequency, given by

$$\omega_0 = \gamma B_0 \tag{I.2}$$

where ω_0 is the angular precession frequency in radians per second, γ is the gyromagnetic ratio, and B_0 is the static magnetic field strength measured in Tesla. In this environment, protons precess with their rotational axes preferentially aligned parallel to the direction of B_0. The longitudinal component of each spin vector, along the z-axis, remains constant in magnitude and orientation, representing the net magnetization along the static field. Meanwhile, the transverse components, oriented along the x- and y-axes, oscillate sinusoidally over time, tracing a circular path around the z-axis:

$$M_x(t) = M_\perp \cos(\omega_0 t + \phi), \quad M_y(t) = M_\perp \sin(\omega_0 t + \phi) \tag{I.3}$$

I.5.1.3 Radiofrequency Excitation and Flip Angle Dynamics

An radio frequency pulse oscillating at the Larmor frequency introduces a secondary magnetic field (B_1) perpendicular to B_0, tipping the net magnetization away from its equilibrium position. The *flip angle* α determined by the amplitude and duration of the RF pulse controls the extent of magnetization rotation:

$$\alpha = \gamma B_1 \tau \tag{I.4}$$

where B_1 represents the amplitude of the applied RF field, and τ denotes the duration of the RF pulse. A 90° pulse rotates the net magnetization completely into the transverse plane,

thereby maximizing the detectable NMR signal. In contrast, a $180°$ pulse inverts the longitudinal magnetization, a principle that underlies key MRI techniques such as inversion recovery and spin-echo sequences [21].

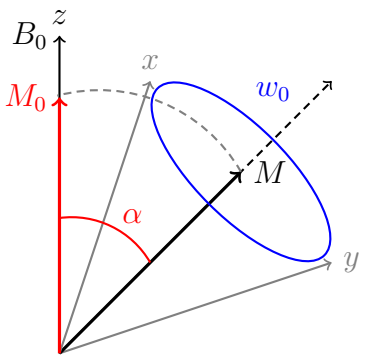


Figure I.14: Larmor Precession & Flip Angle.

I.5.1.4 Signal Generation and Image Reconstruction

After the RF pulse is switched off, nuclear spins gradually return to their equilibrium state, a process during which they emit a time-dependent RF signal known as the Free Induction Decay (FID). The recovery of the longitudinal magnetization along the z-axis and the decay of the transverse magnetization in the xy-plane both follow exponential laws, expressed as:

$$M_z(t) = M_0 \left(1 - e^{-t/T_1}\right) \tag{I.5}$$

$$M_{xy}(t) = M_{xy}(0), e^{-t/T_2} \tag{I.6}$$

Here, T1 represents the longitudinal (spin-lattice) relaxation time, characterizing the recovery of magnetization along the main magnetic field, while T2 is the transverse (spin-spin) relaxation time, describing the decay of magnetization perpendicular to the field.

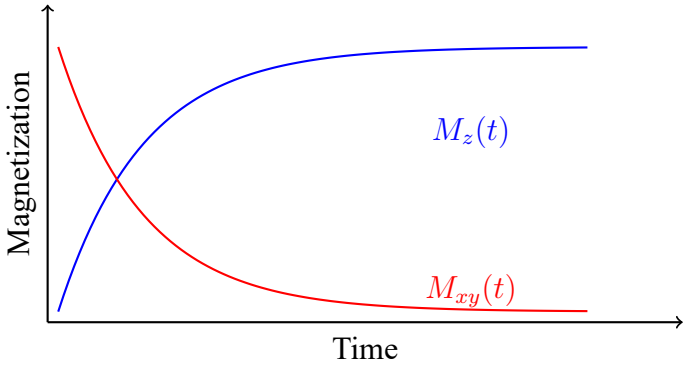


Figure I.15: T1 Recovery (Blue) and T2 Decay (Red) Processes.

In magnetic resonance imaging, the free induction decay signal generated by precessing

nuclear spins is detected by receiver coils, digitized, and mathematically transformed into spatially resolved images using the Fourier transform. To achieve spatial localization, magnetic field gradients are applied in three orthogonal directions: the slice-selection gradient restricts excitation to a specific tissue slice, the frequency-encoding gradient distinguishes signal along one in-plane axis based on frequency differences, and the phase-encoding gradient introduces controlled phase shifts to encode spatial information along the perpendicular axis. Together, these gradients allow reconstruction of high-resolution images from the raw FID signal. To improve efficiency, modern parallel imaging techniques such as sensitivity encoding (SENSE) and Generalized autocalibrating partially parallel acquisitions (GRAPPA) exploit the spatial sensitivity profiles of multiple receiver coils to accelerate acquisition, thereby shortening scan times and reducing motion-related artifacts and geometric distortions [22].

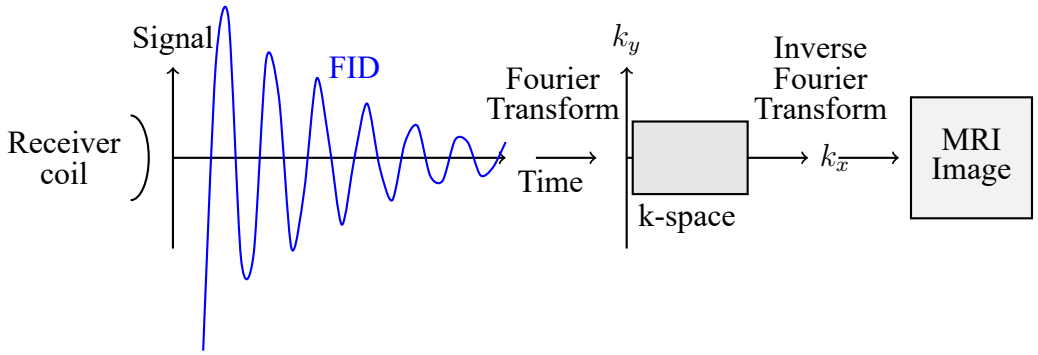


Figure I.16: Representation of Image Reconstruction.

I.5.2 Physical Principles of MRI

I.5.2.1 Nuclear Magnetic Resonance Fundamentals

NMR arises from the quantum mechanical property of nuclear spin, which allows certain nuclei to occupy discrete energy levels when placed in an external magnetic field B_0. When subjected to a radiofrequency excitation at the Larmor frequency (Equation I.2), spins can transition between these quantized states, absorbing and subsequently re-emitting energy. In MRI, the principle of NMR is harnessed for spatially resolved imaging by superimposing controlled magnetic field gradients (G_x, G_y, G_z):

$$\omega(r) = \gamma \left[B_0 + G_x x + G_y y + G_z z\right] \tag{I.7}$$

These gradients induce predictable variations in the Larmor frequency as a function of position, thereby encoding spatial information into the measured signal. The data acquired in this process populate a frequency-phase domain known as *k-space*, from which the spatial image is reconstructed via inverse Fourier transformation [4].

I.5.2.2 Magnetic Moment and Magnetization Dynamics

Each hydrogen nucleus (proton) possesses an intrinsic spin angular momentum, giving rise to a magnetic moment μ. In the presence of B_0, these moments precess about the field axis at the Larmor frequency (Equation I.2). The macroscopic behavior of the bulk magnetization vector M, representing the sum of all individual moments, is governed by the Bloch equations:

$$\frac{dM}{dt} = \gamma M \times B - \frac{M_x\hat{i} + M_y\hat{j}}{T_2} - \frac{(M_z - M_0)\hat{k}}{T_1} \tag{I.8}$$

Here, the first term represents precession around B, the second term represents transverse magnetization decay (T2 relaxation), and the third term represents longitudinal recovery toward equilibrium (T1 relaxation). This formulation encapsulates the fundamental temporal evolution of magnetization in the presence of both static and time-varying fields [23].

I.5.2.3 Relaxation Mechanisms

Following RF excitation, the net magnetization disturbed from equilibrium undergoes relaxation through two simultaneous but distinct mechanisms:

Longitudinal (T1) Relaxation: Also termed spin-lattice relaxation, T1 refers to the recovery of the longitudinal magnetization component (M_z) toward its equilibrium value M_0 (Equation I.5). This process occurs as spins transfer excess energy to the surrounding molecular lattice, which acts as a thermal reservoir. The rate of recovery varies across tissues: those with short T1 values regain equilibrium more rapidly, resulting in higher signal intensity on T1-weighted images. In contrast, tissues with longer T1, such as cerebrospinal fluid (CSF, $\sim$2000 ms at 1.5 T), appear darker due to slower recovery [20].

Transverse (T2) Relaxation: Also known as spin-spin relaxation, T2 describes the progressive loss of phase coherence among spins precessing in the transverse plane, leading to an exponential decay of M_{xy} (Equation I.6). Unlike T1, this process does not involve energy exchange with the lattice but instead arises from local magnetic interactions between neighboring spins. Importantly, T2 is always less than or equal to T1, with typical values of $\sim$80 ms for fat and $\sim$2000 ms for CSF at 1.5 T. Shorter T2 values cause faster signal decay, which strongly influences the brightness of tissues in T2-weighted imaging [20].

Free Induction Decay (FID): Immediately following the RF pulse, the coherent precession of transverse magnetization induces an oscillating voltage in the receiver coil, known as the FID signal (Equation I.9). This signal contains both amplitude and phase informa-

tion about the spin system. In practice, the FID is often sampled using gradient echo or spin echo sequences to mitigate the effects of magnetic field inhomogeneities and optimize image quality (Equation I.10) [21].

$$S(t) \propto e^{-t/T_2^*} \tag{I.9}$$

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_{\text{inhom}}} \tag{I.10}$$

where T_2^* incorporates both spin-spin relaxation and magnetic field inhomogeneities.

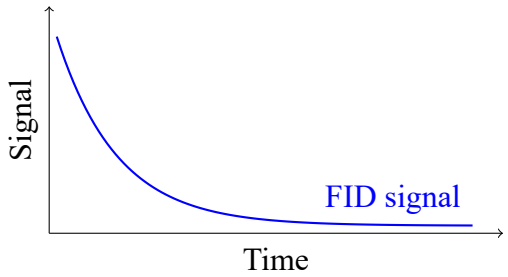


Figure I.17: Free Induction Decay Following RF Excitation.

I.5.3 Image Contrast and Weighted Imaging Techniques

MRI image contrast arises from the complex interplay between tissue-specific properties and pulse sequence parameters. The intrinsic factors that govern signal behavior include the longitudinal relaxation time (T_1), the transverse relaxation time (T_2), and the proton density (ρ), each contributing differently to tissue signal characteristics. These intrinsic properties are modulated by extrinsic imaging parameters such as the repetition time (TR), echo time (TE), and flip angle (α), which can be adjusted to emphasize particular contrast mechanisms. By carefully selecting these parameters, MRI can generate images weighted toward T1, T2, or proton density, thereby enhancing diagnostic specificity. In the case of spin-echo sequences, the signal intensity (S) can be modeled mathematically as:

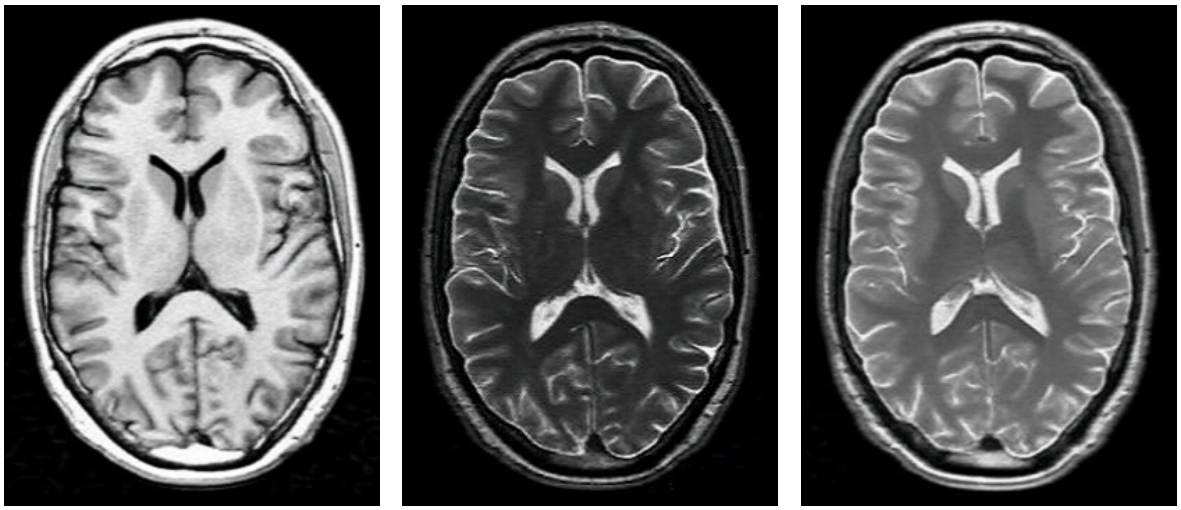
$$S = \rho \left(1 - e^{-\frac{TR}{T_1}}\right) e^{-\frac{TE}{T_2}} \frac{\sin\alpha}{1 - \cos\alpha\, e^{-\frac{TR}{T_1}}} \tag{I.11}$$

This equation illustrates how variations in TR, TE, and α modulate the relative contributions of T1, T2, and ρ to the image contrast.

T1-Weighted Imaging: T1-weighted (T1W) sequences are designed with short repetition times (TR, 400–600 ms) and short echo times (TE, 10–20 ms) to accentuate differences in longitudinal relaxation. In these images, tissues with short T1 values, such as fat, appear hyperintense (bright), whereas tissues with long T1 values, such as cerebrospinal fluid, appear hypointense (dark). T1W imaging offers superior anatomical detail and is particularly

valuable for post-contrast studies, as gadolinium-based contrast agents reduce T1 relaxation times and selectively enhance the visibility of pathological lesions [24].

T2-Weighted Imaging: T2-weighted (T2W) sequences employ long TR (2000–3000 ms) and long TE (80–120 ms) to accentuate differences in transverse relaxation. Fluids, which have long T2 values, appear hyperintense, while tissues with shorter T2, such as white matter, appear darker. T2W images are sensitive to pathological changes, particularly those associated with increased water content, such as edema, inflammation, demyelination, and many tumors [22].



(a) T1W (b) T2W (c) PDW

Figure I.18: Image Contrast and Weighted Imaging Techniques.

Proton Density-Weighted Imaging: Proton density-weighted (PDW) sequences are designed to minimize T1 and T2 contrast by employing a long repetition time (TR > 2000 ms) and a short echo time (TE < 30 ms). Under these conditions, the signal intensity primarily reflects the density of mobile hydrogen protons (ρ) within the tissue. PDW imaging is particularly useful for detecting subtle structural differences or tissue changes in which contrast based on relaxation times alone is limited, providing complementary information to T1- and T2-weighted sequences [20].

Advanced Contrast Methods: Beyond conventional weightings, several advanced techniques extend the diagnostic capabilities of MRI:

- *Diffusion-Weighted Imaging (DWI):* Sensitive to the microscopic motion of water molecules, enabling the detection of acute ischemia and characterization of tissue microstructure.
- *Perfusion MRI:* Measures cerebral blood volume and flow, useful in tumor grading and assessing ischemic penumbra.

- *Magnetization Transfer Imaging:* Enhances contrast between macromolecule-rich tissues and free water, aiding in white matter disease evaluation.
- *Functional MRI (fMRI):* Detects blood-oxygen-level-dependent changes to map neural activity.

These methods provide functional and microstructural information that complements conventional anatomical imaging [21].

I.5.4 Spatial Encoding Techniques

MRI depends on spatial encoding to precisely determine the anatomical origin of the detected signals. This localization is accomplished through the controlled application of magnetic field gradients in conjunction with RF pulses. By varying the local magnetic field strength, these gradients induce predictable changes in the Larmor frequency of nuclear spins, allowing their spatial position to be inferred. Spatial encoding in MRI is achieved through three key steps: slice selection, which isolates a specific tissue slice; frequency encoding, which differentiates positions along one in-plane axis; and phase encoding, which resolves spatial information along the orthogonal in-plane axis [20].

Slice Selection: Slice selection determines the location and thickness of the imaged cross-section.

- An RF pulse is applied while a magnetic field gradient is oriented *perpendicular* to the desired imaging plane (e.g., G_z for an axial slice).
- The gradient induces a *linear variation of the Larmor frequency* along its axis, such that spins at different positions precess at different frequencies.
- By selecting an RF pulse with a specific *frequency bandwidth*, only spins within the corresponding magnetic field range are excited.
- The *slice thickness* Δz is determined by the RF bandwidth (Δf) and the gradient strength (G) as follows:

$$\Delta z = \frac{\Delta f}{\gamma G} \tag{I.12}$$

  where γ is the gyromagnetic ratio.

Frequency Encoding (Readout Gradient): Frequency encoding, also known as the readout gradient, is applied during signal acquisition to localize the signal along one in-plane axis (typically the x-axis).

- A constant gradient (G_x) is applied during readout, causing spins at different x positions to precess at different frequencies.

- The Fourier transform of the acquired signal separates these frequency components, mapping them to specific spatial positions.
- This step directly determines the spatial resolution in the frequency-encoded direction.

Phase Encoding: Phase encoding is used to localize signal along the remaining in-plane axis (typically the y-axis).

- A gradient (G_y) is applied for a short duration prior to signal acquisition.
- This gradient imparts a position-dependent phase shift to the spins, proportional to their y-coordinate.
- The phase encoding gradient is incrementally varied across successive acquisitions, building a complete dataset in k-space.
- After reconstruction, these phase differences translate into spatial localization in the phase-encoded direction.

Combined Encoding Process: By sequentially applying slice selection, phase encoding, and frequency encoding, the MRI system achieves full three-dimensional spatial localization. This coordinated process is fundamental to all MRI pulse sequences, enabling the generation of high-resolution anatomical images with precise spatial correspondence.

I.5.5 Characteristics of MRI Images

Magnetic resonance imaging is distinguished by its ability to generate images with exceptional anatomical detail and high intrinsic tissue contrast. Key attributes that set MRI apart from other imaging modalities include superior spatial resolution, enabling precise delineation of small structures; excellent contrast resolution, which differentiates tissues based on their relaxation properties; multiplanar imaging capability, allowing acquisition in any anatomical plane without repositioning the patient; and the exploitation of specific tissue signal characteristics, such as T1, T2, and proton density, to highlight pathological or functional differences [23].

High Spatial Resolution: MRI can routinely achieve in-plane spatial resolutions between 0.5 mm and 1 mm, depending on acquisition parameters and hardware capabilities. This allows for the visualization of small anatomical structures such as cranial nerves, cortical layers, and fine vascular details. The achievable resolution is determined by:

- **Matrix size:** Higher matrix dimensions yield finer pixel resolution.
- **Field of view:** A smaller field of view increases resolution for a given matrix.

- **Slice thickness:** Thin slices reduce partial volume effects but lower signal-to-noise ratio.

Superior Contrast Resolution: One of MRI's defining strengths is its contrast resolution, with the ability to detect tissue signal differences as small as approximately 0.1%. This is largely due to the dependence of image contrast on intrinsic tissue properties such as T1, T2, and proton density, which can be selectively emphasized using tailored pulse sequences. Such versatility allows differentiation between subtle pathological changes, including early demyelination, edema, or small ischemic lesions.

Multiplanar Capability: MRI data acquisition and reconstruction allow for imaging in any arbitrary plane (axial, sagittal, coronal, or oblique) without physically repositioning the patient. This capability is particularly useful for:

- Visualizing complex anatomical relationships (e.g., in the skull base).
- Planning surgical or interventional approaches.
- Correlating findings across different imaging planes.

Isotropic 3D acquisitions further enable high-quality multiplanar reconstructions with no loss of detail.

Signal Characteristics of Bone Cortical bone produces negligible signal in conventional MRI sequences because of its low proton density and extremely short T2 relaxation time. As a result, bone appears as a signal void (black) on most pulse sequences. This property is advantageous for visualizing bone-soft tissue interfaces but limits MRI's utility in evaluating cortical bone microstructure compared to CT imaging [23].

I.5.6 Common MRI Artifacts and Mitigation

MRI image quality can be degraded by various artifacts that arise from physical, physiological, and technical factors. Understanding these artifacts and their mitigation strategies is essential for accurate diagnosis and research imaging.

Noise and Signal-to-Noise Ratio (SNR): Random noise in MRI originates from multiple sources, including thermal noise from the patient, the radiofrequency (RF) coil, and the electronic system. SNR improves with:

- *Voxel volume:* Larger voxels increase the amount of signal collected, enhancing SNR. However, this comes at the expense of spatial resolution.

- *Number of signal averages:* Acquiring multiple excitations and averaging the results reduces random noise but increases total scan time.
- *Magnetic field strength:* Higher field strengths provide higher SNR, though they may also exacerbate other artifacts.

Mitigation strategies: Optimize voxel size for the clinical question, use phased-array coils to improve sensitivity, and apply parallel imaging methods to maintain SNR while reducing scan time [25].

Partial Volume Effects: Partial volume effects occur when a single voxel contains multiple tissue types, causing averaging of their signals. This leads to blurred boundaries between anatomical structures, especially in regions with fine detail or complex morphology. Mitigation strategies:

- Acquire images with smaller voxel sizes (higher spatial resolution).
- Use thin slices in the slice-select direction to reduce through-plane averaging.
- Apply isotropic 3D acquisitions when possible, allowing reformatting in multiple planes without quality loss [24].

Motion Artifacts: Motion from voluntary (patient movement) or involuntary (respiration, cardiac pulsation, swallowing) sources produces blurring, ghosting, or misregistration of structures. Even subtle motion can severely affect quantitative imaging and functional MRI. Mitigation strategies:

- Use physiological gating or triggering (e.g., Electrocardiography gating for cardiac MRI, respiratory gating for thoracic/abdominal imaging).
- Employ faster sequences such as single-shot echo-planar imaging or turbo spin echo to shorten acquisition time.
- Apply motion correction algorithms during image reconstruction.
- Provide patient comfort aids (e.g., cushions, verbal reassurance) to minimize voluntary movement [22].

I.5.7 Advantages and Limitations

Advantages: MRI offers several distinct advantages that have made it a cornerstone of modern diagnostic imaging [23]:

- *Exceptional soft-tissue contrast resolution:* Allows visualization of subtle differences in tissue composition, particularly valuable in neuroimaging, musculoskeletal imaging, and oncological assessment.

- *Multiplanar imaging capability:* Enables acquisition of images in axial, sagittal, coronal, and oblique planes without physically repositioning the patient, improving anatomical localization.
- *No ionizing radiation:* Safer for repeated follow-up scans, pediatric patients, and pregnant individuals (after the first trimester), making it suitable for longitudinal studies.
- *Functional and quantitative imaging extensions:* Techniques such as fMRI, DWI, diffusion tensor imaging (DTI), and perfusion imaging expand MRI's scope to physiological and metabolic assessment.

Limitations: MRI also presents several limitations that must be considered [4]:

- *High cost:* Installation and operational expenses limit access, particularly in resource-constrained settings.
- *Long acquisition times:* Scans ranging from several minutes to over an hour increase susceptibility to motion artifacts and limit suitability for unstable or uncooperative patients.
- *Motion sensitivity:* Even minor patient movements can significantly degrade image quality; although motion correction methods exist, they are not always fully effective.
- *Patient discomfort:* The confined bore can cause claustrophobia in 5–10% of patients, and loud gradient noise requires adequate hearing protection.
- *Contraindications:* Certain metallic implants (e.g., pacemakers, aneurysm clips, cochlear implants) may pose safety risks; thorough screening is essential.
- *Limited visualization of calcifications and bone:* Reduced sensitivity compared to CT, with additional susceptibility artifacts near metal.

I.5.8 Clinical Applications

MRI has evolved into one of the most versatile modalities in modern diagnostic radiology, offering detailed visualization of anatomical structures and providing functional and metabolic information without the use of ionizing radiation. Its applications span virtually all organ systems, with particularly significant contributions in the following domains:

- **Neuroimaging:** MRI is the gold standard for central nervous system imaging due to its unparalleled soft-tissue contrast and ability to image in multiple planes without repositioning the patient. In tumor detection, contrast-enhanced T1-weighted imaging delineates lesion borders, while T2-weighted and FLAIR sequences reveal associated edema and infiltration. In multiple sclerosis, MRI is indispensable for identifying demyelinating plaques, monitoring disease progression, and evaluating therapeutic efficacy. DWI enables rapid diagnosis of acute ischemic stroke within minutes of onset, allowing for timely intervention. Advanced techniques such as DTI provide white matter tractogra-

phy for pre-surgical planning, while fMRI maps cortical activation in tasks related to motor, sensory, and language functions.

- **Musculoskeletal Imaging:** High-resolution MRI enables detailed visualization of cartilage morphology, ligament integrity, tendon pathology, and bone marrow changes, making it indispensable for sports medicine and orthopedic diagnostics. In the knee, for instance, proton density-weighted and T2-weighted sequences detect meniscal tears, anterior cruciate ligament injuries, and early cartilage degeneration. In the spine, MRI is the preferred modality for assessing intervertebral disc herniation, spinal cord compression, and degenerative changes, often eliminating the need for myelography.

- **Oncology:** Beyond structural assessment, MRI offers multiparametric evaluation for tumor characterization, staging, and treatment monitoring. Dynamic contrast-enhanced MRI assesses tumor vascularity and permeability, which are critical for evaluating angiogenesis in malignancies. In the prostate, multiparametric MRI has become the reference standard for cancer detection and risk stratification. Whole-body MRI, particularly with diffusion-weighted sequences, is increasingly used for staging metastatic disease and monitoring therapy response, especially in pediatric oncology where minimizing radiation exposure is crucial.

- **Cardiac Imaging:** Cardiac MRI provides comprehensive assessment of myocardial structure, function, perfusion, and viability. Cine MRI sequences allow precise quantification of ventricular volumes, ejection fraction, and wall motion abnormalities. Late gadolinium enhancement imaging identifies areas of myocardial fibrosis or infarction, aiding in differentiating ischemic from non-ischemic cardiomyopathies. Stress perfusion MRI evaluates myocardial ischemia under pharmacologic stress, complementing invasive coronary angiography. Additionally, magnetic resonance angiography provides high-resolution, contrast-enhanced imaging of coronary and great vessels without ionizing radiation, facilitating assessment of congenital heart disease and vascular anomalies [21].

I.6 Conclusion

Brain tumors present a complex clinical challenge due to their diverse types, variable clinical presentations, and potential for significant neurological impact. Advances in diagnostic imaging, surgical techniques, and targeted therapies have improved outcomes, particularly for benign tumors [14]. However, malignant tumors like glioblastomas remain difficult to treat, underscoring the need for continued research into their etiology and novel therapeutic approaches [11]. Understanding the epidemiology, symptoms, and management options is critical for optimizing patient care and guiding future investigations [7].

CHAPTER

2 Supervised Machine Learning Algorithms

Introduction

Machine Learning (ML) is a fundamental component of artificial intelligence, allowing computers to analyze data, detect intricate patterns, and make predictions or decisions. Unlike traditional programming, ML algorithms learn and adapt directly from data, making them highly versatile across domains. These algorithms are generally classified into four paradigms, each addressing different tasks and challenges. Supervised Learning relies on labeled datasets to train predictive models, while Unsupervised Learning explores unlabeled data to discover hidden patterns and structures. Semi-Supervised Learning combines labeled and unlabeled data, enhancing performance when annotations are scarce. Reinforcement Learning, in contrast, teaches agents through interaction with an environment, optimizing strategies based on rewards and penalties. Together, these approaches form the backbone of intelligent systems applied in healthcare, finance, robotics, and autonomous technologies.

This chapter systematically introduce supervised machine learning algorithms, emphasizing supervised learning, where labeled training data is utilized to predict either discrete outputs (classification) or continuous outputs (regression). To foster a deep understanding and encourage innovation, we derive key classification algorithms from first principles, including logistic regression, naïve Bayes, Support Vector Machines (SVMs), decision trees, and neural networks, elucidating their mathematical foundations and design trade-offs. The chapter first focuses on classification algorithms, deriving the following classic methods: Perceptron, SVMs, logistic regression, naïve Bayes, and decision trees. Subsequently, we explore four notable regression algorithms: simple linear regression, multiple linear regression, Lasso regression, and Ridge regression, providing a comprehensive analysis of their principles.

II.1 Classification

In supervised learning, we work with a dataset $D = (x_1, y_1), \ldots, (x_n, y_n)$, where x represents input features and $y \in \{1, \ldots, K\}$ denotes discrete labels. Classification algorithms aim to learn a mapping from inputs x to labels y, with binary classification occurring when $K = 2$ and multiclass classification when $K > 2$. A classifier $h \in H$, where H is a set of models, maps a d-dimensional feature vector $\varphi(x)$ to a k-dimensional label, such that $y = h(\varphi(x))$. The objective is to select a model that generalizes effectively to unseen data through a process known as model selection, which balances model complexity to prevent underfitting (models too simple to capture patterns) or overfitting (models too complex, fitting noise in the data). For models $H = \{1, 2, 3\}$, both training and test losses decrease, indicating underfitting due to insufficient model capacity. Conversely, for models $H = \{6, 7, 8\}$, the training loss decreases while the test loss increases, suggesting overfitting caused by excessive model complexity.

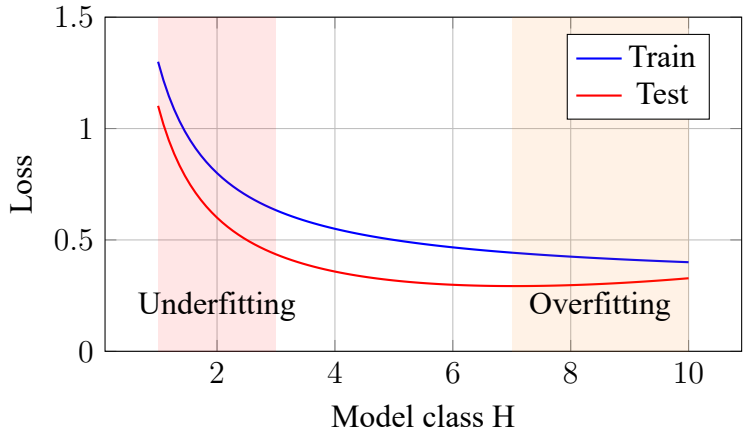


Figure II.1: Model Complexity Trade-Off Between Underfitting and Overfitting.

II.1.1 Perceptron

The perceptron is a fundamental linear classifier for binary classification [26]. Given an input feature vector $x \in \mathbb{R}^d$, it predicts a label $y \in \{+1, -1\}$ using the decision function:

$$h(x; \theta) = \text{sign}(\theta^\top x + \theta_0) = \begin{cases} +1 & \text{if } \theta^\top x + \theta_0 \geq 0, \\ -1 & \text{otherwise} \end{cases} \tag{II.1}$$

where: $\theta \in \mathbb{R}^d$ is the weight vector, $\theta_0 \in \mathbb{R}$ is the bias term, and the equation $\theta^\top x + \theta_0 = 0$ defines a $(d-1)$-dimensional hyperplane in the feature space. Geometrically, θ is the normal vector to the decision hyperplane. All points x lying on the same side as θ are classified as

$+1$, while points on the opposite side are classified as -1. The perceptron's classification performance is quantified by its *error rate*, defined as the fraction of misclassified samples:

$$\mathcal{E}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[y_i \neq h(x_i;\theta)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[y_i(\theta^\top x_i + \theta_0) \leq 0] \tag{II.2}$$

where: $\mathbb{I}[\cdot]$ is the indicator function (1 if the condition is true, 0 otherwise), A misclassification occurs when $y_i(\theta^\top x_i + \theta_0) \leq 0$, and n is the total number of samples. This error measure counts cases where the predicted label $\text{sign}(\theta^\top x_i + \theta_0)$ disagrees with the true label $y_i \in \{-1, +1\}$.

The perceptron is a mistake-driven learning algorithm that iteratively updates its parameters when misclassifications occur. Assuming linearly separable data, it initializes with $\theta = \vec{0}$ and $\theta_0 = 0$, then applies the following update rule for each misclassified example (x_i, y_i):

$$\begin{cases} \theta^{(k+1)} = \theta^{(k)} + \eta y_i x_i, \\ \theta_0^{(k+1)} = \theta_0^{(k)} + \eta y_i, \end{cases} \tag{II.3}$$

where: $\eta = \frac{1}{k+1}$ is the adaptive learning rate, k counts the number of parameter updates.

Algorithm 1 Perceptron Algorithm

function FIT(X, y)
    $k \leftarrow 1$
    for epoch $= 1$ **to** num_epochs **do**
        for $i = 1$ **to** N **do**
            if $y_i(\theta \cdot x_i + \theta_0) \leq 0$ **then**
                $\eta \leftarrow \frac{1}{k+1}$; $k \leftarrow k + 1$ ▷ learning rate update
                $\theta \leftarrow \theta + \eta y_i x_i$; $\theta_0 \leftarrow \theta_0 + \eta y_i$ ▷ update weights
            end if
        end for
    end for
    return θ, θ_0
end function
function PREDICT(X)
    $\hat{y} \leftarrow \text{sign}(\theta \cdot X + \theta_0)$
    return $\hat{y}$
end function

The algorithm cycles through the dataset for a fixed number of epochs, and data is typically shuffled between epochs to prevent order bias. For linearly separable data, the perceptron is guaranteed to converge to a perfect classifier after a finite number of iterations. For test data X_{test}, predictions are made via:

$$\hat{y} = \text{sign}(\theta^\top X_{\text{test}} + \theta_0) \tag{II.4}$$

The perceptron algorithm is concisely captured in the provided pseudo-code, detailing a binary linear classifier with two core functions: fit and predict [27]. Below is a streamlined explanation of the algorithm and its components, tailored to the description provided.

II.1.2 Support Vector Machines

Support Vector Machines (SVMs), introduced in 1992 by Boser, Guyon, and Vapnik and formally named in 1995 by Cortes and Vapnik [28], are supervised learning models grounded in statistical learning theory and the principle of structural risk minimization. They construct an optimal separating hyperplane by maximizing the geometric margin, defined as the distance between the decision boundary and the nearest samples, or support vectors, thereby ensuring robustness and generalization. For non-separable data, SVMs employ a soft-margin formulation with slack variables and a regularization parameter C, while the kernel trick extends their capability to nonlinear problems through functions such as polynomial and radial basis function (RBF) kernels. Owing to this balance of theoretical rigor and practical adaptability, SVMs remain widely applied in fields like bioinformatics, computer vision, and text mining.

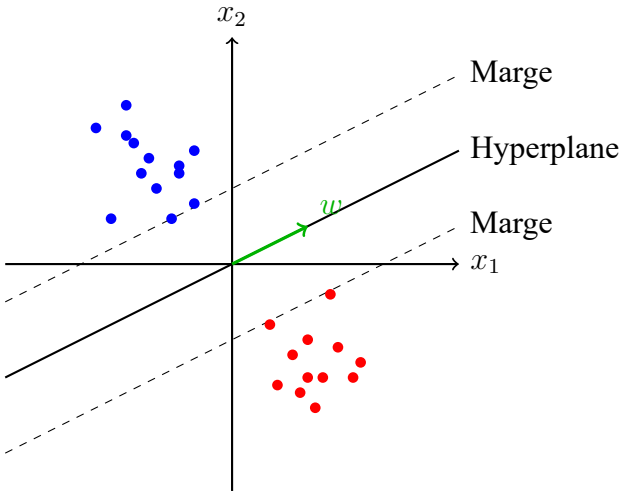


Figure II.2: Max Margin Solution of SVM Classifier.

For a binary classification problem with dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, the optimal hyperplane $\theta \cdot x + \theta_0 = 0$ is found by solving:

$$\min_{\theta,\theta_0} \frac{1}{2}\|\theta\|^2 \quad \text{subject to} \quad y_i(\theta \cdot x_i + \theta_0) \geq 1 \quad \forall i \tag{II.5}$$

The support vectors are the data points that lie on the margin boundaries and satisfy the condition $y_i\,(\theta \cdot x_i + \theta_0) = 1$, where x_i denotes the feature vector, $y_i \in \{-1, +1\}$ is the class label, θ is the weight vector, and θ_0 is the bias term.

For non-linearly separable data, SVMs employ kernel functions $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ that implicitly map data to higher dimensions (linear: $x_i \cdot x_j$, polynomial: $(x_i \cdot x_j + c)^d$, and exponential: $\exp(-\gamma \|x_i - x_j\|^2)$). For noisy data, the soft-margin formulation introduces slack variables ξ_i:

$$\min_{\theta, \theta_0, \xi} \frac{1}{2}\|\theta\|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(\theta \cdot x_i + \theta_0) \geq 1 - \xi_i, \xi_i \geq 0 \tag{II.6}$$

where the parameter $C > 0$ balances margin maximization and tolerance to misclassification. SVMs aim to maximize the margin, the minimum distance from the hyperplane to the nearest data points, ensuring robust generalization. For computational efficiency and to enable kernelization, the problem is expressed in its dual form using nonnegative Lagrange multipliers α_i, leading to the Lagrangian:

$$L(\theta, \theta_0; \alpha) = \frac{1}{2}\|\theta\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\theta \cdot x_i + \theta_0) - 1]. \tag{II.7}$$

Taking gradients with respect to θ and θ_0 yields $\theta = \sum_{i=1}^{n} \alpha_i y_i x_i$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. Substituting these into the Lagrangian gives the quadratic programming problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j [x_i^T x_j], \tag{II.8}$$

subject to $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. In this dual formulation, the data points appear only through their inner products $x_i^T x_j$, which simplifies computation and allows the use of kernel functions for nonlinear classification. This results in a quadratic programming problem of dimension n that can be solved using standard methods. Once the optimal solution $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)$ is obtained, the weight vector of the maximum-margin hyperplane is given by:

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i \tag{II.9}$$

The support vectors are the elements x_i of the training sample for which the coefficients α_i^* are non-zero. The associated decision function is thus:

$$f(x) = \sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^* \tag{II.10}$$

The decision function is given by $\text{sign}(f(x_i))$. In the dual formulation, the inner product $x_i^T x_j$ measures similarity between data points and can be replaced by a kernel function $K(x_i, x_j)$, allowing SVMs to handle nonlinear data by implicitly mapping it into a higher-

dimensional feature space. Kernels are symmetric ($K(x_i, x_j) = K(x_j, x_i)$) and positive semi-definite, with common examples including the radial basis function (RBF) kernel:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right). \tag{II.11}$$

This enables SVMs to find optimal separating hyperplanes in complex datasets without explicitly computing the transformed features, as referenced in prior discussions on non-linear SVM variants.

In most problems, the linearity assumption is too restrictive, and the optimal separator must take a more complex form. The kernel method is an effective way to address this problem. A nonlinear transformation Φ is applied to the input data vectors, yielding their redescription in the new feature space F. The separating hyperplane obtained in F is called the generalized optimal hyperplane. To optimize the separator, the solution takes the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i^* y_i \Phi(x_i) \cdot \Phi(x) + b^* \tag{II.12}$$

This problem and its solution depend only on the dot product $\Phi(x_i) \cdot \Phi(x_j)$. Rather than choosing the nonlinear transformation $\Phi : X \to F$, we can directly choose a function $k : X \times X \to \mathbb{R}$ called a *kernel function*.

Algorithm 2 SVM Algorithm

function FIT(X, y, kernel K)
    Solve quadratic programming problem:
        $\max_\alpha \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$
        subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$
    Identify support vectors: $\mathcal{S} = \{i : \alpha_i > 0\}$
    Compute bias: $b = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(y_s - \sum_{i=1}^{N} \alpha_i y_i) K(x_i, x_s)\right)$
    return Support vectors $\{\alpha_i > 0\}$, weights $\alpha \in \mathbb{R}^N$, bias b
end function
function PREDICT(X_{test}, support vectors $\{(x_s, y_s, \alpha_s)\}_{s \in \mathcal{S}}$, b, kernel K)
    Compute decision function: $f(x) = \sum_{s \in \mathcal{S}} \alpha_s y_s K(x_s, x) + b$
    $\hat{y} = \text{sign}(f(x))$
    return Predicted class $\hat{y} \in \{-1, 1\}$
end function

Since SVMs are binary classifiers, solving a multi-class problem involves transforming it into a combination of binary problems. All proofs for the results in this section can be found in [29]. Let $Y = \{C_1, \ldots, C_k, \ldots, C_Q\}$ be a set of Q classes. A multi-class SVM implements vector functions $h = (h_k)_{1 \leq k \leq Q}$, where the component functions $h_k(\cdot) = \langle w_k \cdot \Phi(\cdot) \rangle + b_k$ are constructed similarly to binary SVM decision functions. An individual is assigned to the class corresponding to the highest output. The model parameters are determined by solving

the following optimization problem:

$$\min_{w_k,b_k,\xi_i} \frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2 + C\sum_{i=1}^{m}\xi_i \tag{II.13}$$

subject to: $\langle w_{y_i}\cdot\Phi(x_i)\rangle + b_{y_i} \geq \langle w_k\cdot\Phi(x_i)\rangle + b_k + 2 - \xi_i \quad \forall i, k \neq y_i.$

The formulation of this optimization problem implicitly includes, for simplicity, pseudo-variables $\alpha_{iy_i}, (1 \leq i \leq m)$ all equal to 0. The expression for w_k in terms of α_{ik} and the training data is:

$$w_k = \sum_{i=1}^{m}\alpha_{ik}\Phi(x_i) \tag{II.14}$$

The b_k values are derived at the optimum from the Kuhn-Tucker conditions.

II.1.3 Logistic Regression

Logistic regression is a widely used classification algorithm that models the probability of class membership in a probabilistic framework [30]. It estimates $p(C_k \mid x)$, the probability that an input x belongs to class C_k. In binary classification with classes C_1 and C_2, Bayes' theorem expresses the posterior probability in terms of the class-conditional likelihood and prior.

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)} = \frac{1}{1 + \frac{p(x|C_2)p(C_2)}{p(x|C_1)p(C_1)}} \tag{II.15}$$

where $p(C_k)$ represents prior class probabilities. The model applies a logistic (sigmoid) function to a linear combination of features, mapping real-valued inputs to $[0, 1]$. Parameters correspond to log-odds, making them interpretable. This framework also extends to multi-class classification via the softmax function and supports regularization to prevent overfitting.

$$p(C_1|x) = \frac{1}{1+\exp(-a)} = \sigma(a) \text{ where } a = \ln\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} \tag{II.16}$$

For multiclass classification ($K > 2$), the generalization leads to the softmax function:

$$p(C_k|x) = \frac{\exp(a_k)}{\sum_{i=1}^{K}\exp(a_i)} \text{ where } a_k = \ln p(x|C_k)p(C_k) \tag{II.17}$$

The softmax function transforms a vector of K real values into a probability distribution that sums to unity. For binary logistic regression parameterized by θ, the model specification is: $p(y|x,\theta) = \text{Bernoulli}(y|\sigma(\theta^T x))$. The joint distribution for data points (x_i, y_i) is:

$$p(x_i, y_i|\theta) = p(y_i|x_i,\theta)p(x_i|\theta) = \text{Bernoulli}(y|\sigma(\theta^T x))p(x_i|\theta) \tag{II.18}$$

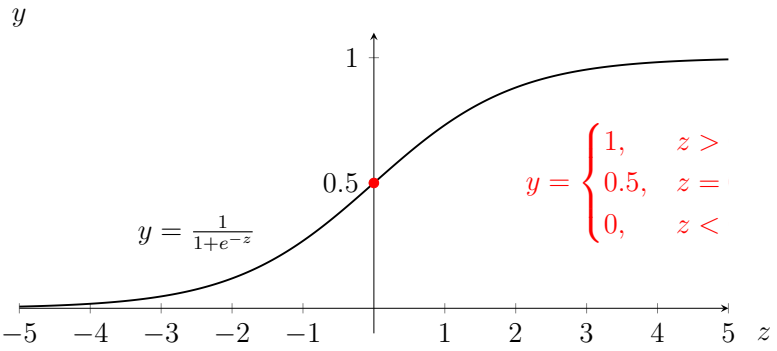


Figure II.3: The Sigmoid Function and Logistic Function.

Since we typically do not model the data distribution $p(x_i|\theta) = p(x_i)$, the log-likelihood becomes:

$$\begin{aligned} \log p(D|\theta) &= \sum_{i=1}^{n} \log p(x_i, y_i|\theta) \\ &= \sum_{i=1}^{n} \left[y_i \log \sigma(\theta^T x_i) + (1 - y_i) \log(1 - \sigma(\theta^T x_i)) \right] \end{aligned} \tag{II.19}$$

Parameter estimation is performed by maximizing the log-likelihood, or equivalently, by minimizing the negative log-likelihood (NLL):

$$\min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \text{NLL}(\theta) = \max_{\theta} \log p(D|\theta) \tag{II.20}$$

We employ stochastic gradient descent for optimization: $\theta_{k+1} = \theta_k - \eta_k g_k$ where g_k is the gradient and η_k is the learning rate. Convergence is guaranteed by the Robbins-Monro conditions:

$$\sum_{k=1}^{\infty} \eta_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \eta_k^2 < \infty \tag{II.21}$$

These conditions are satisfied by the learning rate schedule: $\eta_k = (\tau_0 + k)^{-\kappa}$ where $\tau_0 \geq 0$ and $\kappa \in (0.5, 1]$. The gradient of the log-likelihood is derived as:

$$\frac{d}{d\theta} \log p(D|\theta) = \sum_{i=1}^{n} [y_i - \mu_i] \, x_i = -X^T(\mu - y) \text{ where } \mu_i = \sigma(\theta^T x_i) \tag{II.22}$$

This derivation uses the property $\frac{d}{dx}\sigma(x) = (1 - \sigma(x))\sigma(x)$. To prevent overfitting, we add L2 regularization:

$$\mathcal{L}_{\text{reg}}(\theta) = \text{NLL}(\theta) + \lambda \theta^T \theta \tag{II.23}$$

The regularized gradient becomes:

$$g_k = X^T(\mu - y) + 2\lambda\theta \tag{II.24}$$

Algorithm 3 Logistic Regression Algorithm

function LR_OBJECTIVE(θ, X, y, λ)
  $\mu_i \leftarrow \sigma(\theta^\top X_i)$ ▷ Sigmoid function
  cost $\leftarrow -\sum_{i=1}^{n} [y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)] + \lambda\theta^\top\theta$
  grad $\leftarrow X^\top(\mu - y) + 2\lambda\theta$ ▷ Compute gradient
  return cost, grad
end function
function FIT(X, y)
  $\eta_i \leftarrow (\tau + i)^{-\kappa}$ ▷ Learning rate schedule
  for $i = 1$ **to** num_iter **do**
    cost, grad $\leftarrow$ lr_objective(θ, X, y, λ)
    $\theta \leftarrow \theta - \eta_i \cdot$ grad ▷ Update parameters
  end for
  return θ
end function
function PREDICT(X, θ)
  $\hat{y} \leftarrow \sigma(\theta^\top X)$ ▷ Prediction probability
  return $\hat{y}$
end function

The logistic regression algorithm comprises three primary functions: lr_objective, fit, and predict. Within the lr_objective function, the regularized objective function and its gradient are computed as outlined in the text. The fit function then initializes the learning rate and iteratively updates the theta parameters by moving in the direction opposite to the gradient. Finally, the predict function uses the trained model to make binary label predictions on test data. In the subsequent code listing, a synthetic Gaussian mixture dataset is utilized to train the logistic regression model [27].

II.1.4 Naive Bayes

The Naive Bayes algorithm is a probabilistic classifier based on Bayes' theorem. It relies on the strong assumption that features are conditionally independent given the class label. Despite this simplification, Naive Bayes is widely used due to its efficiency and strong empirical performance in diverse domains [31]. Under the conditional independence assumption, the class conditional density factorizes as:

$$p(x_i \mid y = c, \theta) = \prod_{j=1}^{D} p(x_{ij} \mid y = c, \theta_{jc}) \tag{II.25}$$

The Figure II.4 illustrates the probabilistic graphical model for Naive Bayes: (left) training documents $(x_{i,j}, y_i)$ with parameters (θ, π), (right) test document $(x_{t,j}, y_t)$ generated from the same parameters.

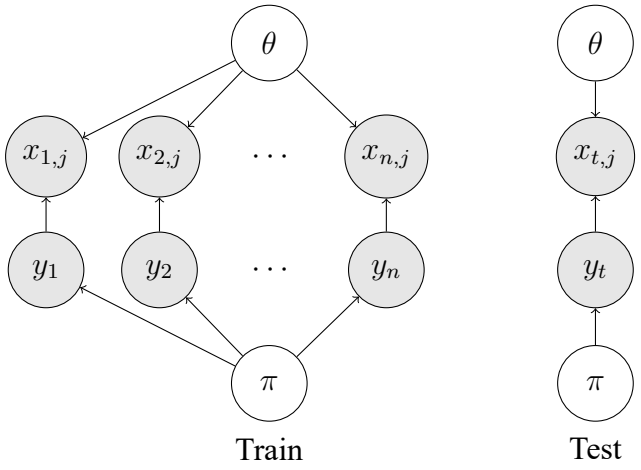


Figure II.4: Probabilistic Graphical Model for Naive Bayes

For document classification, we often use Bernoulli Naive Bayes. Let: x_{ij}: Bernoulli random variable indicating presence (1) or absence (0) of word $j \in \{1, \ldots, D\}$ in document $i \in \{1, \ldots, N\}$, θ_{jc}: parameter for word j given class label $y = c \in \{1, \ldots, C\}$, and π: prior distribution over class labels. The number of learnable parameters is: $|\theta| + |\pi| = O(DC) + O(C) = O(DC)$ which makes the model parameter-efficient and resistant to overfitting. The class conditional density is given by:

$$p(x|y = c, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{D} p(x_{ij}|y = c, \theta_{jc}) = \prod_{i=1}^{n} \prod_{j=1}^{D} \text{Bernoulli}(x_{ij}|\theta_{jc}) \tag{II.26}$$

The joint distribution for a single document i is:

$$p(x_i, y_i|\theta) = p(y_i|\pi) \prod_{j=1}^{D} p(x_{ij}|y_i, \theta) \qquad = \prod_{c=1}^{C} \pi_c^{\mathbb{1}[y_i=c]} \prod_{j=1}^{D} \prod_{c=1}^{C} p(x_{ij}|\theta_{jc})^{\mathbb{1}[y_i=c]}$$

The log-likelihood objective function becomes:

$$\log p(D|\theta) = \sum_{i=1}^{n} \log p(x_i, y_i|\theta) \qquad = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{D} \sum_{c=1}^{C} \sum_{i:y_i=c} \log p(x_{ij}|\theta_{jc})$$

subject to the constraint $\sum_c \pi_c = 1$. We solve this constrained optimization problem

using the Lagrangian:

$$\mathcal{L}(\theta, \lambda) = \log p(D|\theta) + \lambda \left(1 - \sum_c \pi_c\right) \tag{II.27}$$

Optimizing with Lagrangian multipliers gives:

$$\begin{cases} \hat{\pi}_c = \frac{N_c}{N_{\text{tot}}} \\ \hat{\theta}_{jc} = \frac{N_{jc}}{N_c} \end{cases} \quad \text{where} \quad \begin{cases} N_{jc} = \sum_{i:y_i=c} x_{ij} \\ N_c = \sum_i \mathbb{1}[y_i = c] \end{cases} \tag{II.28}$$

To incorporate prior knowledge and smooth parameter estimates, we can use conjugate priors:

$$\begin{cases} p(\pi|D) = \text{Dirichlet}(N_1 + \alpha_1, \ldots, N_C + \alpha_C) \\ p(\theta_{jc}|D) = \text{Beta}((N_c - N_{jc}) + \beta_0, N_{jc} + \beta_1) \end{cases} \tag{II.29}$$

The conjugate prior formulation enables closed-form posterior updates and computational

Algorithm 4 Naive Bayes Algorithm

function FIT(X, y)
  Initialize $N_c \leftarrow 0$; $N_{jc} \leftarrow 0$ for all j, c
  for $i \leftarrow 1$ **to** n **do**
    $c \leftarrow y_i$; $N_c \leftarrow N_c + 1$ ▷ class label for i^{th} example
    for $j \leftarrow 1$ **to** D **do**
      if $x_{ij} = 1$ **then** $N_{jc} \leftarrow N_{jc} + 1$
      end if
    end for
  end for
  $\hat{\pi}_c \leftarrow \frac{N_c}{n}$; $\hat{\theta}_{jc} \leftarrow \frac{N_{jc}}{N_c}$
  return $\hat{\pi}_c$; $\hat{\theta}_{jc}$
end function
function PREDICT(X)
  for $c \leftarrow 1$ **to** C **do**
    $\log p[c] \leftarrow \log \pi_c$
    for $j \leftarrow 1$ **to** D **do**
      if $x_j = 1$ **then** $\log p[c] \leftarrow \log p[c] + \log \hat{\theta}_{jc}$
      else $\log p[c] \leftarrow \log p[c] + \log(1 - \hat{\theta}_{jc})$
      end if
    end for
  end for
  $c^* \leftarrow \arg\max_c \log p[c]$
  return c^*
end function

efficiency. For test-time prediction, we compute the posterior class probability:

$$p(y = c|x_i, D) \propto p(y = c|D) \prod_{j=1}^{D} p(x_{ij}|y = c, D) = \hat{\pi}_c \prod_{j=1}^{D} \text{Bernoulli}(x_{ij}|\hat{\theta}_{jc})$$

Taking logarithms:

$$\log p(y = c|x_i, D) \propto \log \hat{\pi}_c + \sum_{j=1}^{D} \left[\mathbb{1}[x_{ij} = 1] \log \hat{\theta}_{jc} + \mathbb{1}[x_{ij} = 0] \log(1 - \hat{\theta}_{jc}) \right]$$

This formulation provides an efficient mechanism for document classification using the Naive Bayes framework. The Naive Bayes algorithm is summarized in the pseudo-code shown in Algorithm 4 [27].

II.1.5 Decision Trees

Classification and Regression Trees (CART) represent a family of algorithms that partition the input space into axis-parallel regions, where each leaf node corresponds to a distinct region [32]. These models offer both interpretability and insights into feature importance through their hierarchical structure. The CART algorithm employs a greedy, recursive partitioning approach, as the optimal space partitioning problem is NP-complete. During training, the objective is to construct a binary tree that optimizes a chosen objective function while balancing underfitting and overfitting. The key decision involves selecting the optimal feature and threshold for data splitting.

Given an input data matrix $X \in \mathbb{R}^{n \times d}$ with n observations and d features, we seek the optimal feature $j \in \{1, \ldots, d\}$ and threshold $t \in \tau_j$ that minimizes the cost function:

$$(j^*, t^*) = \arg \min_{j \in \{1,\ldots,d\}} \min_{t \in \tau_j} \text{cost}(\{x_i, y_i : x_{ij} \leq t\}) + \text{cost}(\{x_i, y_i : x_{ij} > t\}) \tag{II.30}$$

where τ_j represents the set of candidate thresholds for feature j, typically constructed as midpoints between adjacent data values. The Gini index measures node impurity and is defined as:

$$G = \sum_{k=1}^{K} \pi_k (1 - \pi_k) = 1 - \sum_{k=1}^{K} \pi_k^2 \tag{II.31}$$

where π_k represents the proportion of points in region D belonging to class k:

$$\pi_k = \frac{1}{|D|} \sum_{i \in D} \mathbb{I}[y_i = k] \tag{II.32}$$

The Gini index corresponds to the expected error rate and reaches its minimum (zero) for pure nodes. An alternative impurity measure is entropy, defined as:

$$H(\pi) = -\sum_{k=1}^{K} \pi_k \log \pi_k \tag{II.33}$$

Entropy quantifies uncertainty and similarly attains its minimum (zero) for pure nodes. Consider one-dimensional data $X = [1.5, 1.7, 2.3, 2.7, 2.7]$ with class labels $y = [1, 1, 2, 2, 3]$. For threshold $t = 2.0$ (midpoint between 1.7 and 2.3), the Gini impurity is computed as:

$$G = \frac{2}{5} G_{\text{left}} + \frac{3}{5} G_{\text{right}} = \frac{2}{5} \times 0 + \frac{3}{5} \times \left(1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right) = 0.27$$

Here, $G_{\text{left}} = 0$ (pure node with class labels $[1, 1]$) and $G_{\text{right}} = 1 - \frac{4}{9} - \frac{1}{9} = \frac{4}{9}$ (mixed node with classes $[2, 2, 3]$).

The CART splitting procedure can be summarized in Algorithm 5 [27]:

Algorithm 5 CART Algorithm

procedure FINDBESTSPLIT(X, y)
  for each feature $j = 1$ to d **do**
    Sort $X[\cdot, j]$ and generate thresholds τ_j
    for each threshold $t \in \tau_j$ **do**
      $m \leftarrow |X|$ ▷ size of current node
      $m_k^{\text{left}} \leftarrow$ class counts for $\{x_i : x_{ij} \leq t\}$
      $m_k^{\text{right}} \leftarrow$ class counts for $\{x_i : x_{ij} > t\}$
      $i \leftarrow |\{x_i : x_{ij} \leq t\}|$
      $G_i^{\text{left}} \leftarrow 1 - \sum_{k=1}^{K} \left(\frac{m_k^{\text{left}}}{i}\right)^2$; $G_i^{\text{right}} \leftarrow 1 - \sum_{k=1}^{K} \left(\frac{m_k^{\text{right}}}{m-i}\right)^2$;
      $G_i \leftarrow \frac{i}{m} G_i^{\text{left}} + \frac{m-i}{m} G_i^{\text{right}}$
      impurity_reduction $\leftarrow$ current_impurity $- G_i$
    end for
  end for
  return $(j^*, t^*) \leftarrow \arg\min_{j,t} G_i$
end procedure
procedure BUILDTREE($X, y, depth$)
  if stopping criteria met **then** **return** leaf node
  else
    $(j^*, t^*) \leftarrow$ FINDBESTSPLIT(X, y)
    $X_{\text{left}}, y_{\text{left}} \leftarrow \{x_i : x_{ij^*} \leq t^*\}$
    $X_{\text{right}}, y_{\text{right}} \leftarrow \{x_i : x_{ij^*} > t^*\}$
    left_child $\leftarrow$ BUILDTREE$(X_{\text{left}}, y_{\text{left}}, depth + 1)$
    right_child $\leftarrow$ BUILDTREE$(X_{\text{right}}, y_{\text{right}}, depth + 1)$
    return internal node with (j^*, t^*), left_child, right_child
  end if
end procedure

The CART algorithm minimizes cost functions like the Gini index by iterating through features and thresholds. For feature j, we sort the training data $X[\cdot, j]$ in $\mathcal{O}(n \log n)$ time and consider midpoints between adjacent values. Gini index for a node is computed as:

$$G = 1 - \sum_{k=1}^{K} \left(\frac{m_k}{m} \right)^2 \tag{II.34}$$

where m is the node size and m_k is the number of points belonging to class k. For each threshold τ_i, we compute:

$$\begin{cases} G_i & = \frac{i}{m} G_i^{\text{left}} + \frac{m-i}{m} G_i^{\text{right}} \\ G_i^{\text{left}} & = 1 - \sum_{k=1}^{K} \left(\frac{m_k^{\text{left}}}{i} \right)^2 \\ G_i^{\text{right}} & = 1 - \sum_{k=1}^{K} \left(\frac{m_k^{\text{right}}}{m-i} \right)^2 \end{cases} \tag{II.35}$$

After finding the optimal feature and threshold, nodes are split recursively until maximum depth is reached. During prediction, test data is traversed from root to leaf to obtain class labels.The tree grows recursively by calling BUILDTREE until reaching maximum depth. At each step, the algorithm computes the majority class and Gini index, determines the best split, and initializes child nodes with the partitioned data.

II.1.6 Ensemble Methods

Ensemble learning methods combine the predictions of multiple base learners to achieve better generalization performance than any individual model alone [33]. The core idea is that while a single classifier may suffer from high variance, bias, or instability, aggregating several models can reduce these weaknesses, leading to improved accuracy and robustness. Ensemble methods leverage the diversity of base learners to capture different aspects of the data, often outperforming even the best individual model.

Formally, consider M base classifiers $h_1, h_2, \ldots, h_M$, each producing a prediction for an input x. An ensemble combines these predictions into a final decision:

$$H(x) = \text{Combine}\left(h_1(x), h_2(x), \ldots, h_M(x)\right). \tag{II.36}$$

The combination rule varies by method, with common approaches including majority voting, weighted voting, or averaging. Ensemble methods are particularly effective when base learners are diverse (i.e., make uncorrelated errors) and when the aggregation mechanism effectively balances their strengths.

The primary ensemble techniques include *bagging*, *boosting*, *stacking*, and *voting*, each addressing different aspects of model performance:

II.1.6.1 Bagging (Bootstrap Aggregating)

Bagging is an ensemble technique that primarily reduces variance by training multiple base learners on different bootstrap samples of the training data and aggregating their predictions [34]. Each base learner is trained independently on a random subset sampled with replacement. For classification tasks, the final prediction is determined by majority voting, whereas for regression tasks it is obtained by averaging the predictions. Formally, the bagged model is defined as:

$$H(x) = \begin{cases} \text{mode}\{h_1(x), \ldots, h_M(x)\} & \text{for classification,} \\ \frac{1}{M} \sum_{m=1}^{M} h_m(x) & \text{for regression,} \end{cases} \tag{II.37}$$

Algorithm 6 Bagging Algorithm

function FIT(X, y, base_learner h, number of bootstraps M)
  Initialize ensemble $\mathcal{H} = \emptyset$
  for $m = 1, 2, \ldots, M$ **do**
    Sample bootstrap dataset D_m of size N from (X, y) ▷ sample with replacement
    Train base learner $h_m = h(D_m)$
    Add h_m to ensemble $\mathcal{H}$
  end for
  return ensemble $\mathcal{H} = \{h_1, \ldots, h_M\}$
end function
function PREDICT($\mathcal{H}, X$, mode)
  for each sample $x_i \in X$ **do**
    Get predictions $\hat{y}_{i,m} = h_m(x_i)$ for all $m = 1, \ldots, M$
    $\hat{y}_i = \begin{cases} \text{mode}\{\hat{y}_{i,1}, \ldots, \hat{y}_{i,M}\} & \text{mode = “classification”} \\ \frac{1}{M} \sum_{m=1}^{M} \hat{y}_{i,m} & \text{mode = “regression”} \end{cases}$
  end for
  return predictions $\hat{y}$
end function

Bagging is especially effective with high-variance, low-bias models such as decision trees. A widely used extension is *Random Forests*, which increase model diversity further by randomly selecting subsets of features at each split in the trees, thereby reducing correlation among the base learners and improving overall performance.

Random Forest Random Forests extend bagging by incorporating randomness in both data sampling and feature selection, enhancing model diversity and performance [35]. This ensemble method excels in classification and regression tasks due to its accuracy, robustness to noise, and interpretability, making it a popular choice for tabular data.The approach involves constructing multiple decision trees, each one trained on a unique subset of the data

and features, with final predictions aggregated to produce robust outputs. Key components include:

- *Bootstrap sampling*: Each tree is trained on a random sample of the training data drawn with replacement. This ensures diversity among trees and reduces variance in the ensemble. Different trees see different subsets of the data, making the model more robust.
- *Random feature selection*: At each split, only a random subset of features is considered. This reduces correlation between trees, prevents dominance of strong predictors, and improves generalization. It allows the ensemble to capture diverse patterns in the data.
- *Aggregation*: Predictions from all trees are combined to produce the final output. For classification, majority voting is used; for regression, predictions are averaged. Aggregation stabilizes predictions and reduces overall error.
- *Error estimation*: Unused (out-of-bag) samples for each tree provide an unbiased estimate of generalization error. This eliminates the need for a separate validation set. It also allows monitoring model performance during training.

Mathematically, the random forest predictor for an input x is defined as:

$$T(x) = \begin{cases} \text{mode}\{T_1(x), \ldots, T_M(x)\} & \text{for classification,} \\ \frac{1}{M}\sum_{m=1}^{M} T_m(x) & \text{for regression,} \end{cases} \tag{II.38}$$

where h_m represents the m-th decision tree trained on a bootstrap sample and a random subset of features, and M is the total number of trees.

For classification, the probability estimate for class k can be expressed as:

$$P(y = k|x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}(T_m(x) = k), \tag{II.39}$$

where $\mathbb{I}$ is the indicator function, and the final class prediction is:

$$\hat{y} = \arg\max_k P(y = k|x). \tag{II.40}$$

The out-of-bag (OOB) error, used to estimate generalization performance, is computed as:

$$\text{OOB Error} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \neq \hat{y}_i^{\text{OOB}}), \tag{II.41}$$

where $\hat{y}_i^{\text{OOB}}$ is the majority vote (classification) or average (regression) of predictions from trees where the i-th sample was not included in the bootstrap sample.

Feature importance can be quantified using the mean decrease in impurity (e.g., Gini

impurity for classification). For a feature j, the importance is:

$$\text{Imp}_j = \frac{1}{M} \sum_{m=1}^{M} \sum_{t \in T_m} \Delta i(t, j), \tag{II.42}$$

where $\Delta i(t, j)$ is the decrease in impurity at node t when feature j is used for splitting, and T_m is the set of nodes in tree m. The training and prediction process can be formalized as Algorithm 7.

Algorithm 7 Random Forest Algorithm

function FIT(Training data (X, y), number of trees M, number of features d_{try})
    Initialize forest $\mathcal{F} = \emptyset$
    for $m = 1, 2, \ldots, M$ **do**
        Sample bootstrap dataset D_m of size N from (X, y) ▷ sample with replacement
        Grow an unpruned decision tree T_m on D_m:
        for each node in the tree **do**
            Randomly select d_{try} features from all p features
            Choose the best split among these d_{try} features
            Split the node into two child nodes
        end for
        Add tree T_m to the forest $\mathcal{F}$
    end for
    return forest $\mathcal{F} = \{T_1, T_2, \ldots, T_M\}$
end function
function PREDICT($\mathcal{F}, X$, mode)
    for each sample $x_i \in X$ **do**
        Obtain predictions $\hat{y}_i^{(m)} = T_m(x_i)$ for all $m = 1, \ldots, M$
        $\hat{y}_i = \begin{cases} \text{mode}\{\hat{y}_{i,1}, \ldots, \hat{y}_{i,M}\} & \text{mode = “classification”} \\ \frac{1}{M} \sum_{m=1}^{M} \hat{y}_{i,m} & \text{mode = “regression”} \end{cases}$
    end for
    return aggregated predictions $\hat{y}$
end function

Random Forests effectively balance bias and variance, making them highly competitive for predictive tasks. They also provide valuable interpretability through feature importance scores, derived from the frequency and impact of features in tree splits. These scores help identify the most influential variables, aiding in model understanding and feature selection. Additionally, their robustness to overfitting and ability to handle high-dimensional data make Random Forests a versatile and powerful tool in machine learning.

II.1.6.2 Stacking

Exploits model diversity by training a meta-learner to combine the predictions of multiple base learners. Stacking is an ensemble learning technique that leverages the diversity of

multiple base learners by introducing a *meta-learner* to optimally combine their outputs. Unlike bagging or boosting, which primarily focus on variance reduction and bias correction respectively, stacking emphasizes learning how to integrate heterogeneous models in order to achieve better generalization performance. The key idea is to use the predictions of base learners as inputs for training a higher-level model. The final stacked model makes predictions according to:

$$H(x) = g(h_1(x), h_2(x), \ldots, h_M(x)), \tag{II.43}$$

where $g(\cdot)$ denotes the meta-learner's combination function.

Algorithm 8 Stacking Algorithm

function FIT(X, y, base_learners $h_1, \ldots, h_M$, meta-learner $\mathcal{H}$)
    Split (X, y) into K folds
    Initialize meta-feature matrix $Z \in \mathbb{R}^{N \times M}$
    for each base learner h_m **do**
        for $k = 1, 2, \ldots, K$ **do**
            Train $h_m^{(k)}$ on (X_{-D_k}, y_{-D_k})
            Predict $\hat{z}_{i,m} = h_m^{(k)}(x_i)$ for $i \in D_k$
        end for
    end for
    Train meta-learner $\mathcal{H}$ on (Z, y)
    return $\{h_1, \ldots, h_M\}, \mathcal{H}$
end function
function PREDICT($\{h_1, \ldots, h_M\}, \mathcal{H}, X$)
    Compute base predictions $\hat{z}_m = h_m(X)$
    Form meta-feature vector $z = (\hat{z}_{h_1}, \ldots, \hat{z}_{h_M})$
    $\hat{y} = \mathcal{H}(z)$
    return $\hat{y}$
end function

Stacking often achieves superior predictive accuracy, particularly when base models capture complementary aspects of the data distribution. However, it is computationally expensive, sensitive to overfitting if not carefully regularized, and requires robust validation strategies to prevent information leakage.

II.1.6.3 Voting

Combines predictions from multiple models using simple rules, such as majority voting for classification or averaging for regression. Voting is one of the simplest ensemble methods, as it combines predictions from multiple base learners without the need for a meta-learner. The principle is to aggregate individual predictions either by majority decision (*hard voting*) or by averaging probabilities or outputs (*soft voting*). This makes voting computationally inexpensive and straightforward to implement, though generally less flexible than stacking.

The ensemble prediction is given by:

$$H(x) = \begin{cases} \text{mode}\big(h_1(x), h_2(x), \ldots, h_M(x)\big) & \text{mode = “hard voting”} \\ \frac{1}{M}\sum_{m=1}^{M} p_m(x) & \text{mode = “soft voting”} \end{cases} \tag{II.44}$$

where $p_m(x)$ denotes the probability output of classifier h_m.

Voting is particularly effective when the base learners are diverse and individually accurate. Although it lacks the adaptability of stacking, it remains a robust and easy-to-use ensemble approach for both classification and regression tasks.

Algorithm 9 Voting Algorithm

function FIT(X, y, base_learners $h_1, \ldots, h_M$, mode)
    Train each base learner h_m independently on (X, y)
    return $\{h_1, \ldots, h_M\}$
end function
function PREDICT($\{ h_1, \ldots, h_M\}$, X, mode)
    if mode = “hard” **then**
        $\hat{y}_m = h_m(X)$ for $m = 1, \ldots, M$
        $\hat{y} = \text{mode}(\hat{y}_1, \ldots, \hat{y}_M)$
    else if mode = “soft” **then**
        Obtain probability outputs $p_m(X)$ from each h_m
        $\hat{y} = \frac{1}{M}\sum_{m=1}^{M} p_m(X)$
    end if
    return $\hat{y}$
end function

II.1.6.4 Boosting

Reduces *bias* by sequentially training models, with each new model focusing on correcting the errors of the previous ones. AdaBoost, Gradient Boosting, and modern variants like XGBoost are widely used. Boosting focuses on reducing bias by training models sequentially, where each new learner emphasizes the mistakes of its predecessors. Unlike bagging, which trains models independently, boosting adapts iteratively to the weaknesses of prior learners.

1. **AdaBoost (Adaptive Boosting)**, proposed by Freund and Schapire [36], is one of the most influential boosting algorithms. It assigns weights to training instances, increasing the influence of misclassified samples, and combines the learners using weighted voting. The final AdaBoost classifier makes predictions according to

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right), \tag{II.45}$$

where α_t is the weight assigned to weak learner h_t based on its accuracy.

Algorithm 10 AdaBoost Algorithm

function FIT(X, y, WEAKLEARN, number of rounds T)
  Initialize weights $w_{1,i} = \frac{1}{N}$ for $i = 1, \ldots, N$
  for $t = 1, 2, \ldots, T$ **do**
    Compute distribution $p_t = \frac{w_t}{\sum_{i=1}^{N} w_{t,i}}$
    Train weak hypothesis $h_t = \text{WEAKLEARN}(X, y, p_t)$
    Compute error $\varepsilon_t = \sum_{i=1}^{N} p_{t,i} \, |h_t(x_i) - y_i|$
    Compute weight update factor $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\varepsilon_t}{\varepsilon_t} \right)$
    Update sample weights: $w_{t+1,i} = w_{t,i} \cdot \exp(-\alpha_t \, y_i \, h_t(x_i))$
  end for
  Normalize final weights: $w_{T+1,i} \leftarrow \frac{w_{T+1,i}}{\sum_i w_{T+1,i}}$
  return $\{h_1, \ldots, h_T\}, \{\alpha_1, \ldots, \alpha_T\}$
end function
function PREDICT($\{h_t, \alpha_t\}_{t=1}^{T}, X$)
  Compute weighted sum: $s(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$
  $\hat{y} = \text{sign}(s(x))$
  return $\hat{y}$
end function

2. **Gradient Boosting** extends AdaBoost by formulating boosting as an optimization problem over a differentiable loss function $L(y, F(x))$. Instead of reweighting instances, it fits new learners to the residual errors (i.e., negative gradients of the loss) made by the current model. Each iteration incrementally improves predictions by moving in the direction of steepest descent in function space. The final model is expressed as:

$$F_T(x) = F_0(x) + \sum_{t=1}^{T} \gamma_t h_t(x), \tag{II.46}$$

where each γ_t is a step size (learning rate) that controls the contribution of the weak learner h_t. Gradient Boosting is highly flexible, supporting various differentiable loss functions for regression, classification, and ranking tasks. Modern implementations such as *XGBoost* [37], *LightGBM* [38], and *CatBoost* [39] enhance scalability, regularization, and handling of categorical features, making Gradient Boosting one of the most widely used ensemble methods in practice.

Practical considerations: Implementing ensemble methods requires careful consideration of several factors:

- **Base learner choice**: Weak learners are typically used in boosting, while bagging benefits from stronger, less correlated learners like deep decision trees.

- **Computational cost**: Boosting and stacking are computationally intensive due to sequential training and meta-learner optimization, respectively. Bagging is parallelizable, making it more scalable.
- **Overfitting**: Boosting methods, especially with many iterations, can overfit noisy data. Regularization techniques and early stopping are critical.
- **Diversity**: Ensemble performance depends on the diversity of base learners. Random Forests achieve this through feature subsampling, while stacking relies on using heterogeneous models (e.g., combining SVMs, decision trees, and neural networks).

Algorithm 11 Gradient Boosting Algorithm

function FIT(X, y, loss function L, base learner h, number of rounds T, learning rate η)
  Initialize model: $F_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c)$
  for $t = 1, 2, \ldots, T$ **do**
    Compute residuals: $r_{i,t} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{t-1}}$
    Fit weak learner h_t on data $\{(x_i, r_{i,t})\}$
    Find:$\gamma_t = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, F_{t-1}(x_i) + \gamma h_t(x_i))$
    Update model: $F_t(x) = F_{t-1}(x) + \eta.\gamma_t h_t(x)$
  end for
  return $\{h_1, \ldots, h_T\}, \{\gamma_1, \ldots, \gamma_T\}, F_T$
end function
function PREDICT(F_T, X)
  return $F_T(x)$
end function

Recent developments in ensemble methods (as deep ensembles [40]) have significantly advanced their performance and applicability and provide optimized implementations with features like handling missing values, categorical variables, and GPU acceleration.

Light Gradient Boosting Machines (LightGBM) LightGBM is an advanced gradient boosting framework designed for large-scale, high-dimensional datasets, delivering significant improvements in speed and memory efficiency compared to traditional boosting methods [38]. Widely favored in industry and machine learning competitions, particularly for sparse and large datasets, it maintains high predictive accuracy while leveraging innovative techniques to enhance training efficiency.

The framework builds an ensemble of decision trees sequentially, with each tree correcting the errors of its predecessors by minimizing a specified loss function. LightGBM's key innovations include histogram-based splitting, which bins continuous features into discrete intervals to reduce memory usage and speed up split finding; leaf-wise tree growth, which selects the leaf with the maximum loss reduction at each step, resulting in deeper, specialized trees; Gradient-based One-Side Sampling (GOSS), which prioritizes samples with large

gradients while subsampling those with small gradients; and Exclusive Feature Bundling (EFB), which combines mutually exclusive sparse features to reduce dimensionality. Additionally, LightGBM supports GPU acceleration and parallelization, enabling efficient training on large-scale datasets.

The objective function at iteration t is defined as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \tag{II.47}$$

where $l(y_i, \hat{y}_i)$ is the loss function (e.g., mean squared error for regression, log loss for classification), $\hat{y}_i^{(t-1)}$ is the prediction from the previous iteration, $f_t(x_i)$ is the output of the new tree, and $\Omega(f_t)$ is a regularization term, typically:

$$\Omega(f_t) = \gamma L + \frac{1}{2}\lambda \sum_{j=1}^{L} w_j^2, \tag{II.48}$$

where L is the number of leaves, w_j is the weight of leaf j, γ penalizes the number of leaves, and λ controls the complexity of leaf weights.

In histogram-based splitting, continuous feature values are binned into K discrete bins, reducing the number of split points to evaluate. The gain for a split on feature j at threshold s is computed as:

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{II.49}$$

where I_L and I_R are the sets of samples in the left and right child nodes, I is the parent node, $g_i = \frac{\partial l(y_i,\hat{y}_i)}{\partial \hat{y}_i}$ is the first-order gradient (**gradient**), $h_i = \frac{\partial^2 l(y_i,\hat{y}_i)}{\left(\partial \hat{y}_i\right)^2}$ is the second-order gradient (**Hessian**), and γ and λ are regularization parameters. The training procedure is formalized as as Algorithm 12.

For GOSS, the sample set is split into a subset A with large gradients (top a%) and a subset B with small gradients, from which a fraction b% is randomly sampled. The gradient for samples in B is scaled by $\frac{1-a}{b}$, and the split gain is adjusted accordingly:

$$\text{Gain}_{\text{GOSS}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in A_L} g_i + \frac{1-a}{b}\sum_{i\in B_L} g_i\right)^2}{\sum_{i\in A_L} h_i + \frac{1-a}{b}\sum_{i\in B_L} h_i + \lambda} + (\text{right node}) - (\text{parent node})\right] - \gamma \tag{II.50}$$

where A_L and B_L are samples in the left child from sets A and B.

LightGBM's optimizations make it exceptionally fast and scalable, often outperforming traditional gradient boosting methods like XGBoost in both speed and accuracy. Its ability to

handle large, sparse datasets, combined with robust predictive performance and support for feature importance analysis, makes it a powerful tool for a wide range of machine learning tasks.

Algorithm 12 LightGBM Algorithm

function FIT(X, y, loss function L, number of rounds T, learning rate η)
    Bin continuous features into histograms.
    Initialize predictions: $\hat{y}_i^{(0)} = 0$ or a constant.
    for $t = 1, 2, \ldots, T$ **do**
        Compute gradients g_i and Hessians h_i.
        Apply GOSS to select samples with large gradients Equation (II.50).
        Apply Exclusive Feature Bundling (EFB) to bundle mutually exclusive features.
        Find optimal splits using histogram-based search and Equation (II.49).
        Grow tree leaf-wise $T_m(x)$, selecting the leaf with maximum loss reduction, subject to depth/leaf constraints.
        Update predictions: $f_m(x) = f_{m-1}(x) + \eta\, T_m(x)$.
    end for
    return $\{f_1, \ldots, f_T\}$
end function
function PREDICT($\{f_1, \ldots, f_T\}, X$)
    $\hat{y} = \hat{y}^{(0)} + \eta \sum_{t=1}^{T} f_t(X)$.
    return $\hat{y}$
end function

Extreme Gradient Boosting (XGBoost) XGBoost is a highly efficient and scalable implementation of gradient boosting, optimized for large-scale datasets and widely adopted in both academic research and industry [37]. It enhances classical gradient boosting through innovations that improve accuracy, prevent overfitting, and accelerate training, making it a benchmark for structured data tasks in machine learning competitions.

The method constructs an ensemble of regression trees sequentially, with each tree correcting the errors of its predecessors by minimizing a regularized loss function. Key features include L1 and L2 regularization to control model complexity, second-order optimization for precise updates, native handling of missing values and sparse data, and efficient split finding via weighted quantile sketches. Additionally, XGBoost leverages parallelization and GPU acceleration to scale effectively to large datasets.

The objective function at iteration t is defined as Equation (II.47), and $\Omega(f_t)$ is the regularization term:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} |w_j|, \tag{II.51}$$

where T is the number of leaves, w_j is the weight of leaf j, γ penalizes the number of leaves, λ controls L2 regularization, and α controls L1 regularization.

To optimize the objective, XGBoost uses a second-order Taylor approximation of the loss:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{N} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t), \tag{II.52}$$

where g_i is the gradient and h_i is the Hessian (Equation (II.49)). For a tree with leaf set I_j (samples mapped to leaf j), the optimal leaf weight w_j^* is derived by minimizing the objective:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i + \alpha \cdot \text{sign}\left(\sum_{i \in I_j} g_i\right)}{\sum_{i \in I_j} h_i + \lambda}, \tag{II.53}$$

where the L1 term introduces a soft thresholding effect. The corresponding objective value for the tree is:

$$\mathcal{L}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{II.54}$$

The gain for a split on feature k at threshold s as Equation (II.49). The training procedure is formalized as Algorithm 13:

Algorithm 13 XGBoost Algorithm

function FIT(X, y, loss function L, number of rounds T, learning rate η)
    Initialize predictions: $\hat{y}_i^{(0)} = 0$ or a constant.
    for $t = 1, 2, \ldots, T$ **do**
        Compute gradients g_i and Hessians h_i (Equation (II.49)).
        Use weighted quantile sketch to find candidate split points.
        For each node, evaluate splits using Equation (II.49) and select the best.
        Grow tree level-wise, assigning samples to leaves based on splits.
        Compute optimal leaf weights using Equation (II.53).
        Update predictions: $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$.
    end for
    return $\{f_1, \ldots, f_T\}$
end function
function PREDICT($\{f_1, \ldots, f_T\}, X$)
    $\hat{y} = \hat{y}^{(0)} + \eta \sum_{t=1}^{T} f_t(X)$.
    return $\hat{y}$
end function

XGBoost's sparsity-aware split finding enables it to handle missing values by learning the optimal default direction for each feature. For a feature with missing values, the algorithm evaluates whether to assign missing samples to the left or right child to maximize the gain. Additionally, the weighted quantile sketch approximates split points efficiently for large datasets, ensuring scalability. Feature importance is derived from the frequency and gain of splits involving each feature, aiding interpretability.

XGBoost's combination of regularization, second-order optimization, and computational

efficiency makes it exceptionally powerful for structured data tasks. Its flexibility in supporting various loss functions and its ability to scale via parallelization and Graphics Processing Unit (GPU) acceleration have established it as a leading tool in ML applications.

Categorical Ggradient Boostings (CatBoost) CatBoost is a gradient boosting framework optimized for datasets with categorical features, designed to reduce overfitting and handle complex data efficiently [39]. It stands out in applications like recommendation systems, search ranking, and tabular business data, often outperforming other boosting methods when categorical feature engineering is challenging.

The algorithm builds an ensemble of decision trees sequentially, with each tree correcting the errors of its predecessors by minimizing a regularized loss function. Key innovations include ordered boosting to prevent target leakage, efficient categorical encoding using ordered target statistics, and symmetric (oblivious) trees that apply the same split at each level for faster training and inference. CatBoost also excels with small or noisy datasets, supports multi-class classification natively, and leverages GPU acceleration for scalability.

The objective function at iteration t is defined as Equation (II.47), and $\Omega(f_t)$ is the regularization term:

$$\Omega(f_t) = \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{II.55}$$

where T is the number of leaves, w_j is the weight of leaf j, and λ controls L2 regularization.

CatBoost uses a second-order approximation to optimize the objective:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{N} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t), \tag{II.56}$$

where g_i is the gradient and h_i is the Hessian (Equation (II.49)). For a tree with leaf set I_j (samples mapped to leaf j), the optimal leaf weight is:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \tag{II.57}$$

minimizing the approximated objective. The corresponding objective value is:

$$\mathcal{L}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda}. \tag{II.58}$$

The gain for a split on feature k at threshold s as Equation (II.49). Categorical features are encoded using ordered target statistics to avoid target leakage. For a categorical feature

x_k with value c, the encoded value for sample i is computed as:

$$\text{TS}_{i,k}(c) = \frac{\sum_{j \in P_i, x_{j,k}=c} y_j + \alpha \cdot \mu}{\sum_{j \in P_i, x_{j,k}=c} 1 + \alpha}, \tag{II.59}$$

where P_i is the set of samples before i in a random permutation, μ is a prior (e.g., global mean of y), and α is a smoothing parameter. The training procedure is formalized as Algorithm 14.

Algorithm 14 CatBoost Algorithm

function FIT(X, y, loss function L, number of rounds T, learning rate η)
    Generate random permutations of training data (X, y).
    Initialize predictions: $\hat{y}_i^{(0)} = 0$ or a constant.
    for $t = 1, 2, \ldots, T$ **do**
        Compute gradients g_i and Hessians h_i (Equation (II.49)).
        Encode categorical features using ordered target statistics Equation (II.59).
        Find optimal splits for symmetric trees using Equation (II.49).
        Grow symmetric tree f_t, applying the same split at each level.
        Compute optimal leaf weights using Equation (II.57).
        Update predictions: $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$.
    end for
    return $\{f_1, \ldots, f_T\}$
end function
function PREDICT($\{f_1, \ldots, f_T\}, X$)
    $\hat{y} = \hat{y}^{(0)} + \eta \sum_{t=1}^{T} f_t(X)$.
    return $\hat{y}$
end function

CatBoost's ordered boosting mitigates target leakage by computing gradients and target statistics using different permutations of the data, ensuring unbiased estimates. Symmetric trees reduce model complexity and speed up inference, as each level uses a single split condition. The algorithm's robustness to noisy data, efficient handling of categorical features, and GPU support make it highly effective for real-world applications with diverse data types. Feature importance, derived from split gains, enhances interpretability, further solidifying CatBoost's utility in structured data tasks.

II.2 Regression

Regression analysis is a fundamental tool in statistics and machine learning for quantifying the relationship between a dependent variable and one or more predictors. While simple and multiple linear regression provide unbiased parameter estimates under classical assumptions, they are often sensitive to multicollinearity and can exhibit high variance in practical settings.

To address these limitations, regularization techniques such as ridge and lasso regression have been developed. Ridge regression mitigates instability by applying an ℓ_2 penalty that shrinks coefficient estimates toward zero, reducing variance without eliminating predictors, whereas lasso regression employs an ℓ_1 penalty to induce sparsity, enabling automatic variable selection and enhancing model interpretability. These methods, by controlling the bias–variance trade-off, provide a robust foundation for modern regression modeling, improving both predictive performance and the interpretability of complex models.

II.2.1 Simple Linear Regression

Simple Linear Regression constitutes the most elementary form of regression analysis, modeling the relationship between a single explanatory variable and a response variable [41, 42]. Formally, the model is expressed as:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{II.60}$$

Here, y_i represents the dependent variable, x_i the independent variable, β_0 the intercept, β_1 the slope coefficient, and ϵ_i an error term with zero mean, $\mathbb{E}[\epsilon_i] = 0$, and constant variance, $\text{Var}(\epsilon_i) = \sigma^2$. Parameter estimation is performed using the Ordinary Least Squares (OLS) method, which determines β_0 and β_1 by minimizing the Residual Sum of Squares (RSS) [41]:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{II.61}$$

Differentiating with respect to β_0 and β_1 yields the closed-form solutions:

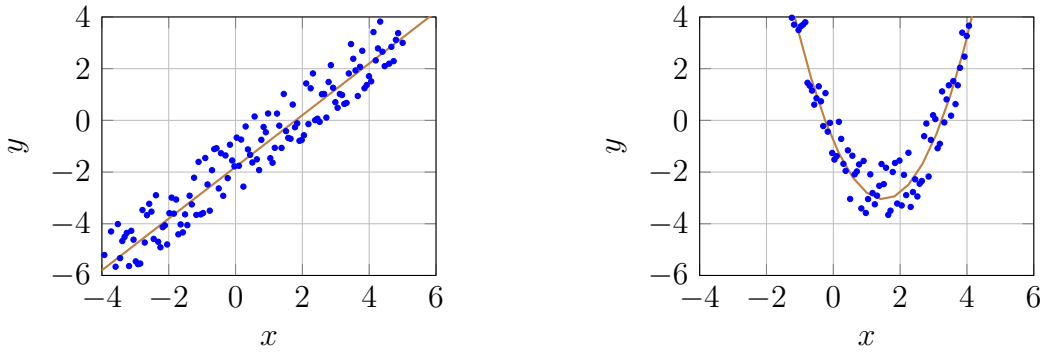
$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases} \tag{II.62}$$

The estimators are unbiased, $\mathbb{E}[\hat{\beta}_j] = \beta_j \quad (j = 0, 1)$, and their variances are:

$$\begin{cases} \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \end{cases} \tag{II.63}$$

Theorem 1 (Gauss-Markov Theorem) *Under the assumptions of linearity, independence, homoscedasticity, and no perfect multicollinearity, the OLS estimators are the best linear unbiased estimators of the regression coefficients[43].*

While powerful for one-dimensional analysis, simple regression is limited in scope, as



(a) Linear (b) Polynomial

Figure II.5: Bidimensional Models.

real-world data typically depends on multiple explanatory variables. The gradients in Algorithm 15 arise directly from differentiating the loss function. The objective is to minimize the residual sum of squares Equation (II.61). For the intercept, differentiating yields:

$$\begin{cases} \text{grad}_{\beta_0} = \nabla_{\beta_0} L = \frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) \\ \text{grad}_{\beta_1} = \nabla_{\beta_1} L = \frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) x_i \end{cases} \tag{II.64}$$

These expressions show that gradient descent updates are derived rigorously from calculus applied to the least-squares criterion.

Algorithm 15 Simple Linear Regression Algorithm

function FIT(X, y)
    Initialize β_0, β_1
    for $t = 1, 2, \ldots, T$ **do**
        $\hat{y} = \beta_0 + \beta_1 X$
        $\text{grad}_{\beta_0} = -2 \sum (y - \hat{y})$
        $\text{grad}_{\beta_1} = -2 \sum (y - \hat{y}) X$
        $\beta_0 \leftarrow \beta_0 - \eta \, \text{grad}_{\beta_0}$
        $\beta_1 \leftarrow \beta_1 - \eta \, \text{grad}_{\beta_1}$
    end for
    return (β_0, β_1)
end function
function PREDICT($w = (\beta_0, \beta_1)$, X)
    $\hat{y} = \beta_0 + \beta_1 X$
    return $\hat{y}$
end function

II.2.2 Multiple Linear Regression

Multiple Linear Regression generalizes simple linear regression by incorporating multiple predictors [41, 42]. The model is specified as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{II.65}$$

or in matrix notation:

$$y = X\beta + \epsilon \tag{II.66}$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times(p+1)}$ is the design matrix, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$ is the parameter vector, and $\epsilon \sim N(0, \sigma^2 I)$ is the noise vector with $\mathbb{E}[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2 I$. The ordinary least squares estimator is:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{II.67}$$

provided $X^T X$ is invertible. The variance of the estimator is :

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \tag{II.68}$$

The optimization problem minimizes the squared error loss:

$$L(\beta) = (y - X\beta)^T (y - X\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \tag{II.69}$$

Differentiating with respect to β yields the gradient:

$$\nabla_\beta L(\beta) = -2X^T (y - X\beta) \tag{II.70}$$

This gradient represents the direction of steepest increase in the loss function; moving in the opposite direction reduces the loss. For a learning rate $\eta > 0$, the iterative update rule for gradient descent is:

$$\beta^{(t+1)} = \beta^{(t)} - \eta \nabla_\beta L(\beta^{(t)}) = \beta^{(t)} + 2\eta X^T (y - X\beta^{(t)}) \tag{II.71}$$

A key challenge in multiple regression is multicollinearity, where high correlation among predictors causes $X^T X$ to become nearly singular, leading to inflated variances of $\hat{\beta}$ and unstable inference[41]. The mean squared error (MSE) of predictions can be decomposed as:

$$\text{MSE}(\hat{y}) = \text{Bias}^2(\hat{y}) + \text{Var}(\hat{y}) + \sigma^2 \tag{II.72}$$

This decomposition highlights the bias-variance trade-off inherent in regression analysis.

Algorithm 16 Multiple Linear Regression Algorithm

function FIT(X, y)
    Initialize β
    for $t = 1, 2, \ldots, T$ **do**
        $\hat{y} = X\beta$
        grad $= -2X^T(y - \hat{y})$
        $\beta \leftarrow \beta - \eta$ grad
    end for
    return β
end function
function PREDICT(β, X)
    $\hat{y} = X\beta$
    return $\hat{y}$
end function

Since MLR assumes no regularization, the estimators are unbiased but may exhibit high variance when multicollinearity is present.

II.2.3 Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) regression extends linear regression by introducing an ℓ_1 penalty term on the coefficients (sum of the absolute values), promoting sparsity where some coefficients are shrunk exactly to zero, thus performing variable selection [44, 45]. The optimization problem is defined as:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{II.73}$$

where $\sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 norm. The tuning parameter $\lambda \geq 0$ controls the strength of regularization: larger λ encourages more coefficients to be exactly zero. Equivalently, the lasso problem can be expressed as a constrained optimization:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{II.74}$$

where t is related to λ.

The objective function can be written compactly as:

$$L(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{II.75}$$

The non-differentiability of the ℓ_1 penalty at zero prevents direct use of gradient descent.

Instead, coordinate descent is employed, optimizing one coefficient β_j at a time while keeping others fixed, cycling through coordinates until convergence. For each coefficient β_j, the partial residual is defined as:

$$\begin{cases} r_j = y - \sum_{k \neq j} x_k \beta_k \\ z_j = \sum_{i=1}^{n} x_{ij}^2 \\ p_j = \sum_{i=1}^{n} x_{ij} r_{j,i} \end{cases} \tag{II.76}$$

where x_k denotes the k-th predictor column of X. The update rule for β_j uses the soft-thresholding operator [46]:

$$\beta_j \leftarrow \frac{1}{z_j} S(p_j, \lambda/2) \tag{II.77}$$

where $S(p_j, \lambda/2) = \text{sign}(p_j) \cdot \max(|p_j| - \lambda/2, 0)$. This approach is efficient due to the closed-form updates, and the ℓ_1 penalty naturally enforces sparsity.

Once convergence is reached, predictions are made as:

$$\hat{y} = X\beta \tag{II.78}$$

The non-differentiability of the absolute value introduces computational challenges but yields sparsity. Algorithms like coordinate descent or Least Angle Regression (LARS) are commonly used to compute lasso solutions. Under appropriate regularity conditions, lasso achieves consistency in both estimation and variable selection when λ is chosen properly.

Algorithm 17 Lasso Regression Algorithm

function FIT(X, y)
    Initialize $\beta = 0$
    repeat
        for $j = 1, 2, \ldots, p$ **do**
            $r_j = y - \sum_{k \neq j} x_k \beta_k$
            $z_j = \sum_{i=1}^{n} x_{ij}^2$
            $p_j = \sum_{i=1}^{n} x_{ij} r_{j,i}$
            $\beta_j \leftarrow \frac{\text{sign}(p_j)\max(|p_j| - \lambda/2, 0)}{z_j}$
        end for
    until convergence
    return β
end function
function PREDICT(β, X)
    $\hat{y} = X\beta$
    return $\hat{y}$
end function

II.2.4 Ridge Regression

Ridge Regression introduces an L_2 penalty on the regression coefficients to stabilize estimates in the presence of multicollinearity [47, 45]. The optimization problem is:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{II.79}$$

In closed form:

$$\hat{\beta}^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y \tag{II.80}$$

Unlike lasso, ridge does not set coefficients to zero but shrinks them continuously towards the origin. Its effect can be studied via eigen decomposition: if $X^T X = V D V^T$, then:

$$\hat{\beta}^{\text{ridge}} = V(D + \lambda I)^{-1} V^T X^T y \tag{II.81}$$

which highlights that ridge penalizes directions associated with small eigenvalues, thus reducing instability. Ridge regression trades a small bias for a substantial reduction in variance, often decreasing the overall prediction error:

$$\text{MSE}(\hat{\beta}^{\text{ridge}}) = \text{Bias}^2(\hat{\beta}^{\text{ridge}}) + \text{Var}(\hat{\beta}^{\text{ridge}}) \tag{II.82}$$

Ridge regression, also known as Tikhonov regularization, is an extension of ordinary least squares regression that introduces an ℓ_2 penalty on the coefficients (Euclidean norm). This penalty shrinks the regression coefficients, reducing variance and improving generalization, particularly when multicollinearity is present. The optimization problem is defined as:

$$L(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{II.83}$$

where $\sum_{j=1}^{p} \beta_j^2$ is the squared ℓ_2 norm of the coefficients. The parameter $\lambda \geq 0$ controls the strength of the regularization:

- $\lambda = 0$: Ridge regression reduces to ordinary least squares.
- $\lambda > 0$: Larger values shrink the coefficients but never make them exactly zero.

Taking the gradient with respect to β:

$$\nabla_\beta L(\beta) = -2X^T(y - X\beta) + 2\lambda\beta \tag{II.84}$$

This consists of two terms: $-2X^T(y - X\beta)$, the ordinary least squares gradient, and $2\lambda\beta$,

Algorithm 18 Ridge Regression Algorithm

```
function FIT(X, y)
    Initialize β
    for t = 1, 2, ..., T do
        ŷ = Xβ
        grad = −2X^T(y − ŷ) + 2λβ
        β ← β − η grad
    end for
    return β
end function
function PREDICT(β, X)
    ŷ = Xβ
    return ŷ
end function
```

the ridge penalty term. The iterative update rule becomes:

$$\beta \leftarrow \beta - \eta \left(-2X^T(y - X\beta) + 2\lambda\beta\right) \tag{II.85}$$

where η is the learning rate. Once the final vector β is obtained, predictions for new data are made as:

$$\hat{y} = X\beta \tag{II.86}$$

Ridge regression is especially useful when predictors are highly correlated, since the ℓ_2 regularization stabilizes the estimation and improves generalization.

II.2.5 Comparative Summary

The following table summarizes the essential characteristics of the four regression methods discussed:

Table II.1: Comparative Summary of Regression Methods.

Method	Penalty	Variable Selection	Bias	Variance Control
Simple Regression	None	No	Unbiased	No
Multiple Regression	None	No	Unbiased	Sensitive to collinearity
Lasso Regression	ℓ_1 norm	Yes	Biased	Strong (sparse solutions)
Ridge Regression	ℓ_2 norm	No	Biased	Strong (shrinks all coefficients)

Simple and multiple regression provide unbiased estimators [41, 42], as they do not impose penalties on the regression coefficients. While this property ensures unbiasedness, it also comes with drawbacks. In multiple regression, particularly when predictors are highly

correlated, the absence of regularization leads to high variance in the coefficient estimates. This sensitivity to multicollinearity reduces the stability and interpretability of the model, often limiting its predictive performance in real-world applications.

To mitigate these challenges, regularized regression methods such as ridge [47] and lasso [44] were introduced. Ridge regression employs an ℓ_2 norm penalty, shrinking all coefficients toward zero but never forcing any exactly to zero. This approach stabilizes the estimates, reduces the effect of multicollinearity, and generally improves the robustness of predictions [45]. By distributing shrinkage across all predictors, ridge regression ensures that correlated variables are handled more effectively than in standard multiple regression.

Lasso regression, by contrast, applies an ℓ_1 norm penalty, which not only shrinks coefficient estimates but also sets some of them exactly to zero. This property enables automatic variable selection, yielding sparse solutions that are especially useful in high-dimensional data scenarios. Both ridge and lasso introduce a degree of bias through regularization, but the resulting reduction in variance often leads to superior predictive accuracy. Ultimately, the choice between these methods depends on the problem context: ridge is generally preferred when multicollinearity is the primary issue, whereas lasso is particularly advantageous for identifying the most relevant predictors in complex datasets.

Conclusion

In conclusion, understanding the supervised learning paradigm establishes the foundation for exploring the broader spectrum of machine learning algorithms. This chapter expanded upon that basis by analyzing the objectives, mechanisms, strengths, and limitations of each method. Organizing algorithms by category clarified their interrelationships and underscored their distinctive roles, while the inclusion of applications and pseudocode bridged theoretical concepts with practical implementation. Collectively, these elements provide the reader with the knowledge to select, adapt, and apply supervised learning techniques effectively in real-world AI systems.

CHAPTER 3

State of the Art on MGMT Promoter Methylation

III.1 Introduction

O^6-methylguanine-DNA methyltransferase (MGMT) promoter methylation represents a critical molecular biomarker in gliomas, particularly glioblastoma. Hypermethylation of the MGMT promoter leads to transcriptional silencing of the gene, reducing the cell's DNA repair capacity and thereby increasing sensitivity to alkylating chemotherapeutic agents such as temozolomide (TMZ) [48, 49]. Clinically, patients harboring a methylated MGMT promoter generally demonstrate prolonged progression-free survival and overall survival compared to those with unmethylated promoters [48].

Conventional assessment relies on tissue sampling followed by pyrosequencing, methylation-specific Polymerase Chain Reaction (PCR), or next-generation sequencing [50, 51]. These methods, while accurate, are hindered by cost, complexity, and intratumoral heterogeneity. In parallel, radiomics quantitative extraction of features from medical imaging combined with machine learning or deep learning can detect subtle imaging signatures of molecular alterations [5, 52]. AI models trained on multiparametric MRI or hybrid modalities are increasingly achieving clinically relevant accuracy [53, 54]. This review integrates molecular knowledge with advances in non-invasive prediction, synthesizing recent evidence to guide translation into clinical practice.

III.2 Molecular and Clinical Significance

MGMT is a DNA repair enzyme that removes alkyl adducts from the O^6 position of guanine, reversing mutagenic lesions. This repair activity counteracts the effects of alkylating

chemotherapeutic agents, such as temozolomide. Consequently, MGMT expression influences tumor cell sensitivity to therapy and impacts patient outcomes [48].

III.2.1 Molecular Basis

The MGMT promoter contains Cytosine-Phosphate-Guanine dinucleotide (CpG) islands whose methylation disrupts transcription factor binding and reduces chromatin accessibility. Hypermethylation results in decreased mRNA and protein expression, leaving tumor cells vulnerable to TMZ. Approximately 35–45% of primary glioblastomas and up to 70% of lower-grade IDH-mutant gliomas harbor MGMT promoter methylation [48, 49]. The alteration is generally stable over the disease course and often aligns with a proneural gene expression profile [55].

III.2.2 Clinical Relevance

MGMT promoter methylation status functions as both:

- **Predictive biomarker** : serves as a molecular indicator of the tumor's likelihood to benefit from temozolomide (TMZ) and related alkylating agents. High MGMT promoter methylation silences the repair enzyme, reducing its ability to reverse TMZ-induced DNA damage and thereby enhancing chemosensitivity [48, 49].

- **Prognostic marker** : independently correlates with more favorable clinical outcomes regardless of therapy. MGMT promoter methylation is linked to longer progression-free and overall survival in both lower-grade gliomas and glioblastoma [48, 49], reflecting reduced DNA repair capacity and slower tumor evolution.

Meta-analyses report median overall survival of 21–24 months for methylated versus 12–15 months for unmethylated glioblastoma under standard chemoradiation [48, 49]. In elderly or frail patients, MGMT methylation status strongly influences treatment decisions between TMZ monotherapy and radiotherapy [56].

III.2.3 Clinical Integration

Testing for MGMT promoter methylation is recommended in major neuro-oncology guidelines for treatment stratification [57]. However, assay heterogeneity and variable cutoffs complicate comparisons. Consensus protocols, proficiency testing, and quality assurance are essential for widespread adoption [58].

III.3 Testing Methodologies

Accurate assessment of MGMT promoter methylation is essential for both research and patient management. Current approaches rely primarily on analysis of tumor tissue obtained by biopsy or resection [51].

III.3.1 Tissue-Based Molecular Diagnostics

Accurate determination of MGMT promoter methylation traditionally relies on analysis of tumor tissue obtained by surgical resection or stereotactic biopsy. Multiple laboratory techniques are available, each with distinct analytical characteristics, sensitivity, and clinical practicality [50, 51].

- **Methylation-Specific PCR (MS-PCR)** : Following sodium bisulfite conversion of unmethylated cytosines to uracil, primer pairs specific to either methylated or unmethylated DNA are used to amplify the promoter region. MS-PCR is cost-effective, requires limited DNA, and is easily implemented in routine laboratories. However, it provides only a qualitative "methylated/ unmethylated" result and is highly dependent on assay design and DNA quality[50].
- **Quantitative or Real-Time MS-PCR (qMS-PCR)** : Builds on MS-PCR by incorporating fluorescent probes or intercalating dyes to measure amplification in real time. Quantitative signal output enables determination of relative methylation levels against a reference gene, supporting threshold-based interpretation. This method improves reproducibility but still interrogates a limited set of CpG sites [51].
- **Pyrosequencing** : Considered a gold standard in many institutions, pyrosequencing provides single-nucleotide resolution and returns methylation percentages at multiple CpG dinucleotides within the promoter. Commonly applied cutoffs for calling a tumor "methylated" range between 10–21%, depending on the assay. Its quantitative nature facilitates harmonization, though instrument cost and technical expertise are required [51].
- **High-Resolution Melting (HRM) Analysis** : Detects subtle differences in DNA melting curves after bisulfite conversion, distinguishing methylated from unmethylated templates without the need for labeled probes. HRM is rapid and cost-efficient, but its accuracy can be affected by DNA quality and requires careful calibration [51].
- **Next-Generation Sequencing (NGS) Approaches** : Targeted bisulfite sequencing or hybrid-capture panels enable quantitative assessment of hundreds of CpG sites simultaneously and can be multiplexed with other biomarkers (IDH mutation, TERT promoter,

copy-number changes). NGS provides unparalleled depth and scalability but is more expensive and requires robust bioinformatic support [51].

Each method balances trade-offs between cost, throughput, sensitivity, and quantitative accuracy. Choice of platform is often dictated by local resources, regulatory requirements, and the level of precision needed for clinical decision-making.

III.3.2 Assay Variability and Cutoff Definition

Despite widespread clinical use, significant inter-study and inter-laboratory variability persists in MGMT promoter testing. Differences in technical workflows, CpG site selection, and interpretation criteria make direct comparison of results challenging. Major sources of inconsistency include:

- **Diverse CpG Loci Targeted** : Commercial and in-house assays probe different regions of the MGMT promoter, and methylation density may vary across loci. Results from assays interrogating distinct CpG sites are not always interchangeable [48, 59].
- **Non-Uniform Methylation Thresholds** : Published studies report a wide spectrum of cutoffs for designating a tumor "methylated" (e.g., $> 6\%$, $> 9\%$, $> 21\%$ by pyrosequencing). Lack of consensus hampers pooled analyses and clinical guideline alignment [48, 60].
- **Pre-Analytical and Laboratory Differences** : Variability in DNA extraction methods, bisulfite conversion efficiency, amplicon design, and reporting practices can introduce bias. Even subtle workflow differences may alter quantitative output and case classification [58].

The absence of harmonized standards leads to discordant results, limiting cross-trial comparability and diminishing the statistical power of meta-analyses. International efforts are ongoing to address these challenges, including *ring trials*, proficiency testing programs, and consensus statements aimed at defining robust cutoff criteria [58, 59]. Harmonization of analytical protocols, reference materials, and quality-control metrics will be critical for global adoption and for ensuring MGMT methylation retains predictive value across clinical settings.

III.3.3 Non-Invasive and Emerging Approaches

Given the morbidity and sampling bias associated with repeat surgical biopsies, there is strong motivation to develop assays that capture MGMT methylation status without direct tumor tissue. Several complementary strategies are under active investigation:

- **Liquid Biopsy** : Cell-free tumor DNA and circulating tumor DNA shed into cerebrospinal fluid or peripheral blood can be interrogated for MGMT promoter methylation using sensitive bisulfite or next-generation sequencing methods [61, 62].

- **Radiogenomics** : Quantitative imaging signatures derived from MRI or hybrid modalities (PET/MRI) are linked to molecular alterations through machine learning or deep learning models [5, 53, 52].

- **Epigenome-Wide Profiling** : Genome-scale methylation arrays and sequencing platforms allow simultaneous interrogation of thousands of CpG sites, supporting discovery of integrated biomarker panels that combine MGMT status with other prognostic or predictive loci [63].

These non-invasive and high-throughput strategies are promising but remain largely research tools. Robust analytical validation, large-scale prospective studies, and demonstration of clinical utility are required before they can supplant or complement tissue-based testing in routine neuro-oncology practice.

III.4 Methodological Landscape for Imaging-Based Prediction

Prediction of MGMT promoter methylation through imaging relies on radiomics and artificial intelligence (AI) pipelines that transform routine scans into quantitative biomarkers [52, 5]. Although specific implementations differ across studies, most share a series of core steps:

1. **Image Acquisition and Preprocessing** : Standardized acquisition of multiparametric MRI (T1W, T1CE, T2W, FLAIR) or hybrid PET/MRI, followed by intensity normalization, bias-field correction, co-registration, and, when appropriate, harmonization across scanners to reduce inter-institutional variability [52].

2. **Region-of-Interest (ROI) Segmentation** : Delineation of tumor subregions (enhancing core, necrotic zone, edema) performed manually by experts, semi-automatically using threshold or region-growing algorithms, or fully automated via deep learning (U-Net, nnU-Net) to ensure reproducible volumetric labels [5].

3. **Feature Extraction** : Derivation of quantitative descriptors from the ROI, including first-order statistics (intensity histograms), shape and morphological metrics, higher-order texture features (GLCM, GLRLM), frequency or wavelet-transformed features, and increasingly, deep features from convolutional or transformer architectures [52, 5].

4. **Model Training and Validation** : Classical machine learning (support vector machines, random forests, logistic regression), ensemble approaches (gradient boosting, LightGBM), or deep learning (3D CNNs, transformers) are trained to classify methylation status. Cross-validation, independent test sets, and external validation cohorts are essential for mitigating overfitting and establishing generalizability [52, 5].

Predictive performance is commonly summarized using metrics such as accuracy, sensitivity, specificity, and the area under the receiver-operating-characteristic curve (ROC-AUC). Incorporation of time-to-event analyses (progression-free and overall survival) further enhances clinical interpretability, demonstrating whether imaging-based classifiers add prognostic value beyond standard molecular markers. Rigorous methodological reporting and open-source code sharing remain critical to accelerate reproducibility and clinical translation [52, 5].

III.5 Notable Studies

Machine learning (ML) has emerged as a powerful tool for predicting MGMT promoter methylation in gliomas, complementing conventional tissue-based diagnostics. Unlike classical radiomics, which relies on manual feature extraction and statistical modeling, ML approaches leverage automated learning algorithms to identify complex patterns in multi-modal imaging data. Here, we summarize representative ML studies, highlighting methodological strategies, datasets, and predictive performance.

III.5.1 Classical Machine Learning on Radiomic Features

Early machine learning studies leveraged handcrafted radiomic features extracted from conventional MRI sequences, including T1-weighted, T2-weighted, FLAIR, and contrast-enhanced scans. These radiomic features quantify tumor shape, texture, intensity, and heterogeneity, providing a non-invasive window into underlying molecular alterations such as MGMT promoter methylation.

- **Sasaki et al.** extracted 489 texture features from multiparametric MRI and applied Lasso regression. The model achieved an overall accuracy of 67%, demonstrating proof-of-concept for radiomics-driven MGMT assessment. This study highlighted that even conventional MRI contains subtle, quantifiable patterns linked to epigenetic status [64].

- **Hajianfar et al.** focused on peri-tumoral edema, reasoning that the tumor microenvironment may harbor predictive epigenetic signals. Decision tree models trained

on edema-region radiomic features reached an AUC of 0.78, suggesting that features beyond the tumor core (such as infiltrative edema) carry important information for MGMT methylation prediction [65].

These studies established the foundation for classical ML approaches, demonstrating that radiomic features derived from standard imaging could serve as surrogates for molecular profiling. Despite moderate predictive performance, they provided valuable insights into which imaging-derived characteristics are most informative and paved the way for later multi-sequence and multi-modal models.

III.5.2 Multiparametric MRI and Hybrid Approaches

Building on classical radiomics, later studies expanded feature spaces by integrating multiple MRI sequences and introducing more sophisticated feature selection strategies. Multiparametric MRI captures complementary aspects of tumor biology, including structural heterogeneity, edema, and vascularity, which can improve prediction of MGMT promoter methylation.

- **Tasci et al.** employed a hybrid feature weighting and selection strategy across T1-weighted, T2-weighted, and FLAIR MRI, achieving an accuracy of 81.6%. This study demonstrated that multi-sequence input provides a richer representation of tumor characteristics, enabling improved discrimination compared with single-sequence models [66].

- **Crisi et al.** focused on dynamic susceptibility contrast (DSC) perfusion MRI, extracting hemodynamic and vascular features. Their model reached an AUC of 0.84, highlighting the relevance of vascular correlates in MGMT methylation prediction and emphasizing the importance of functional imaging beyond anatomical scans [67].

- **Korfiatis et al.** evaluated T2-weighted texture features using support vector machines and random forest classifiers, obtaining an AUC of 0.85 on an independent cohort. This study reinforced that careful feature engineering combined with robust machine learning algorithms can yield high predictive performance, particularly when validated on external data [68].

Collectively, these multiparametric and hybrid approaches illustrated that combining structural and functional MRI sequences, along with systematic feature selection, enhances the ability to capture subtle imaging biomarkers of MGMT promoter methylation. They laid the groundwork for the integration of multi-modal imaging and the transition toward deep learning-based pipelines.

III.5.3 PET and Modality Fusion

Integration of positron emission tomography with MRI offers complementary metabolic and structural information, which can enhance prediction of MGMT promoter methylation beyond anatomical imaging alone. PET provides insights into tumor metabolism, proliferation, and heterogeneity that are not captured by MRI sequences, enabling machine learning models to exploit multi-dimensional features for improved accuracy.

- **Qian et al.** incorporated ^{18}F-fluorodopa PET uptake parameters into a random forest classifier alongside MRI-derived features. This multi-modal approach achieved $80\,\% \pm 10$ accuracy, demonstrating that metabolic information captured by PET can highlight regions of epigenetic significance and improve predictive performance compared with MRI alone [53].
- **Jiang et al.** (2024) performed feature fusion by combining T1W, T1CE, and FLAIR MRI sequences for lower-grade gliomas. Using a machine learning model trained on the fused feature set, the study achieved an AUC of 0.761. The fusion of multiple MRI sequences captures both tumor core characteristics and peritumoral microenvironment, illustrating the benefit of integrating complementary imaging modalities for methylation prediction [54].

Overall, PET/MRI fusion enables a richer representation of glioma biology, combining structural, functional, and metabolic signatures. Such approaches are particularly valuable for detecting spatial heterogeneity in tumors, where methylation status may vary across different regions. Moving forward, integrating PET with advanced MRI sequences (e.g., perfusion or diffusion imaging) and deep learning models may further enhance prediction accuracy and clinical applicability.

III.5.4 Advanced Machine Learning and Deep Architectures

Recent studies increasingly leverage ensemble methods and deep learning architectures to automate feature extraction and enhance prediction accuracy for MGMT promoter methylation.

- **Han et al.** developed a radiomics workflow originally designed for 1p/19q co-deletion that was adapted to MGMT prediction. The model achieved AUCs of 0.887 on training and 0.760 on validation datasets, demonstrating the flexibility and transferability of robust radiomics pipelines across molecular biomarkers [69].
- **Karabacak et al.** analyzed glioma cases from the national cancer database using LightGBM and random forest classifiers. Their models yielded AUC values between

0.813 and 0.896. The study also provided a SHAP-enabled web calculator, highlighting the importance of model interpretability and practical clinical deployment [70].

- **Do et al.** employed a genetic algorithm for feature selection in low-grade gliomas, achieving 75% accuracy. This approach illustrates the feasibility of heuristic and optimization-based feature selection in small datasets where conventional approaches may underperform [71].

- **Yu et al.** introduced a transformer-based deep learning model that integrates intratumoral and peri-tumoral features, achieving an AUC of 0.923. This approach represents a state-of-the-art benchmark, demonstrating the potential of attention-based architectures in radiogenomics for capturing spatial and contextual tumor information [72].

- **Boulkhiout et al.** presented a comprehensive machine learning pipeline for non-invasive MGMT prediction. The workflow included: 1) feature extraction from multimodal MRI, 2) feature selection using LightGBM and CatBoost, and 3) classification via an ensemble of four ML models (XGBoost, CatBoost, LightGBM, RandomForest) combined through a voting mechanism. Using the BraTS 2021 dataset, the method achieved 92.86% accuracy and 96.84% AUC [5].

The bibliometric analysis of MGMT promoter methylation research (2000–2025) shows growth from 119 publications in 2000 to a peak of 3,630 in 2021, followed by a slight decline (Figure III.1). The post-2010 surge reflects its key role in glioma research for diagnosis, prognosis, and therapy. The recent decrease suggests either stabilization or a shift to specialized studies, yet MGMT methylation remains central in neuro-oncology.

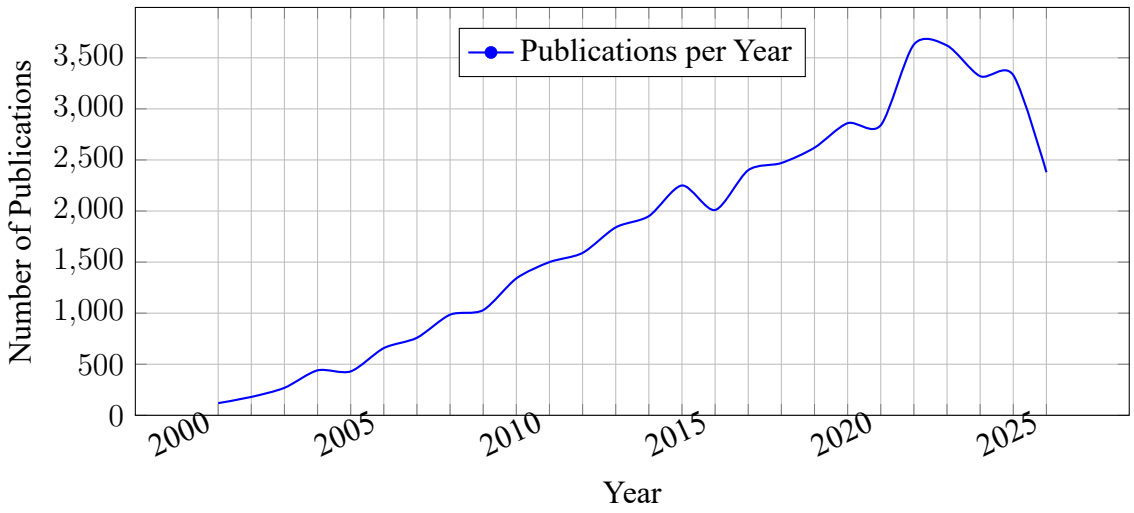


Figure III.1: Growth of MGMT Promoter Methylation Publications (2000–2025).

III.5.5 Representative AI Studies

The following table summarizes key artificial intelligence and machine learning studies that have evaluated MGMT promoter methylation prediction in gliomas. It reports the year of publication, main algorithmic approach, imaging modalities analyzed, reported performance (AUC or accuracy), and the dataset or cohort on which the model was tested.

Table III.1: Performance of AI Models for MGMT Methylation Prediction.

Study	Year	Approach	Modalities	AUC/ Acc	Dataset
Faghani et al.[73]	2023	3D DenseNet121	T2	0.65	BraTS2021
Mun et al.[74]	2021	EfficientNet	Multimodal	0.70	BraTS2021
Calabrese et al.[75]	2022	CNN+RF +SMOTE	Radiomics	0.77	BraTS2021 + Private
Ibtehaz et al.[76]	2024	Deep fusion	Multimodal	0.75	BraTS2021 + Lumiere
Sasaki et al.[64]	2019	LASSO regression on 489 texture features	Multi-sequence MRI	67% Acc	Institutional
Hajianfar et al.[65]	2020	Decision tree on edema-region features	T2/FLAIR	0.78	Local cohort
Tasci et al.[66]	2020	Hybrid feature weighting	T1w/T2w/ FLAIR	81.6% Acc	Multi-center
Crisi et al.[67]	2020	Radiomics on DSC perfusion	DSC-MRI	0.84	Single center
Korfiatis et al.[68]	2020	SVM/RF on T2 textures	T2	0.85	Institutional
Qian et al.[53]	2021	Random forest	F-DOPA PET	80% $\pm$ 10	Institutional
Jiang et al.[54]	2024	Feature fusion	T1/T1CE/ FLAIR	0.761	Multi-center
Han et al.[69]	2022	Radiomics workflow for 1p/19q / MGMT	Multimodal MRI	0.760 (val)	Multi-institutional
Karabacak et al.[70]	2023	LightGBM+RF, SHAP web app	Multi-sequence MRI	0.813–0.896	National cohort
Do et al.[71]	2022	Genetic algorithm feature selection	T2/FLAIR	75% Acc	Institutional
Yu et al.[72]	2024	Transformer-based DL model	Multimodal MRI	0.923	Multi-center
Proposed [5]	2025	Ensemble learning on 15 radiomics features	T1w/T1CE/ T2w/FLAIR	0.9286 Acc and 0.9684 AUC	BraTS2021 559 patients

III.6 Discussion

Table III.1 shows a clear trajectory of improvement in AI-driven MGMT methylation prediction. Early radiomics approaches as Sasaki et al.[64], and Hajianfar et al.[65] achieved only 65–78% accuracy, largely constrained by small institutional datasets and hand-crafted texture features. Hybrid feature selection and multi-sequence inputs as Tasci et al.[66], and Korfiatis et al.[68] improved performance to 81–85% by capturing more robust tumor morphology and intensity patterns.

The incorporation of perfusion MRI and amino acid PET (Crisi et al.[67], and Qian et al.[53]) added functional and metabolic signatures, raising AUCs and underscoring the value of non-anatomical imaging. Feature-fusion studies as Jiang et al.[54] demonstrated the feasibility of harmonizing T1, T1CE, and FLAIR features across centers, although AUCs remain moderate (0.76). More recently, deep learning and transformer-based pipelines (Faghani et al.[73], Mun et al.[74]) have leveraged large-scale MRI data to learn complex imaging phenotypes. Recent deep learning models, particularly transformer-based architectures, have set new benchmarks with AUCs up to 0.923 Yu et al.[72], while ensemble methods achieved 92.86% accuracy (Boulkhiout et al.[5]). Karabacak et al.[70] illustrated how SHAP-based interpretability and web-based deployment can bridge research models to clinical workflows.

Radiogenomic studies on MGMT promoter methylation have shown that multi-sequence MRI remains the primary input for most predictive models. Incorporating genuinely multimodal datasets, such as MRI combined with PET or perfusion imaging, consistently enhances performance by capturing complementary biological information beyond anatomy alone. Model effectiveness is highly influenced by dataset size and harmonization; multicenter studies, including those by Tasci, Jiang, and Yu, demonstrate greater generalizability than single-center reports [66, 54]. Deep learning architectures (3D CNNs, attention models, and transformers) tend to outperform classical machine learning methods when trained on sufficiently large and diverse datasets, while modality fusion across T1w, T2w, FLAIR, perfusion, and PET sequences enables simultaneous characterization of tumor structure and microenvironment.

Interpretability tools such as SHAP and saliency maps are increasingly recognized as essential for clinician trust, allowing radiomic features to be linked meaningfully to histopathology. Integrating survival endpoints further ensures that imaging biomarkers align with clinically relevant outcomes [70]. Despite these advances, several challenges remain. Dataset heterogeneity, due to variations in imaging protocols, scanner vendors, and methylation cutoffs, continues to limit reproducibility and complicate meta-analyses. Class imbalance in public datasets may reduce model robustness and artificially inflate performance metrics, while limited external validation (many models are tested solely on internal data) can overestimate

predictive accuracy. Ultimately, translating these models into clinical practice will require prospective, multi-center studies to confirm reproducibility, demonstrate cost-effectiveness, and satisfy regulatory requirements.

Overall, the field is shifting from proof-of-concept radiomics toward clinically actionable AI pipelines. Combining structural, perfusion, diffusion, and metabolic modalities with harmonized methylation assays will likely yield further gains in accuracy and reliability. Future research should prioritize standardization of imaging acquisition, integration of multi-omics and liquid biopsy markers, and the development of open-source, well-annotated datasets to accelerate benchmarking, transparency, and translation into routine care.

III.7 Conclusion

MGMT promoter methylation remains a cornerstone biomarker in gliomas. Tissue-based testing is the clinical gold standard, but radiomics and AI hold substantial promise for non-invasive evaluation. Harmonization of protocols, external validation, and integration with clinical workflows will determine whether AI-driven radiogenomics achieves routine adoption in precision neuro-oncology.

CHAPTER

4 Proposed Approaches for MGMT Detection

IV.1 Introduction

This thesis introduces a novel machine learning framework for non-invasive prediction of MGMT methylation status in glioblastoma patients, utilizing radiomic features extracted from multi-modal MRI scans in the Brain Tumor Segmentation 2021 dataset [77]. This dataset provides annotated tumor segmentation masks and MGMT methylation labels, enabling targeted analysis of tumor-specific regions. Our methodology integrates advanced techniques: radiomic feature extraction to capture tumor characteristics, feature selection using LightGBM and CatBoost to identify highly predictive features, and a classification pipeline combining four ML models (XGBoost, CatBoost, LightGBM, and RandomForest) through voting and stacking mechanisms to maximize prediction accuracy. By leveraging tumor-specific data and robust ML algorithms, our framework offers a scalable, non-invasive tool for precision oncology, enhancing clinical decision-making in GBM management.

The importance of this work stems from its potential to revolutionize GBM diagnostics and treatment. By predicting MGMT methylation status non-invasively, our approach minimizes the need for invasive surgical procedures and supports rapid, personalized treatment planning. The use of multi-modal MRI data ensures comprehensive tumor characterization, while the ensemble ML strategy boosts predictive accuracy by harnessing the strengths of multiple models. The standardized, well-annotated BraTS2021 dataset ensures the reliability and reproducibility of our results. Ultimately, this study advances radiomics and precision medicine, paving the way for improved patient outcomes through accurate, timely, and non-invasive diagnostic tools for GBM.

IV.2 Materials and Methods

This study develops a machine learning framework to predict MGMT promoter methylation status in glioma patients using multi-sequence MRI scans from the BraTS2021 dataset. The dataset includes 577 patients, each with 3D MRI scans across four sequences (T1, T1CE, T2, and FLAIR) and corresponding tumor segmentation masks. Predicting MGMT methylation status is critical, as it influences treatment response to temozolomide, a standard chemotherapy drug for gliomas. By employing radiomics and advanced ML techniques, this research aims to build a robust, accurate predictive model to support personalized treatment planning, addressing challenges related to glioma heterogeneity and MRI data variability.

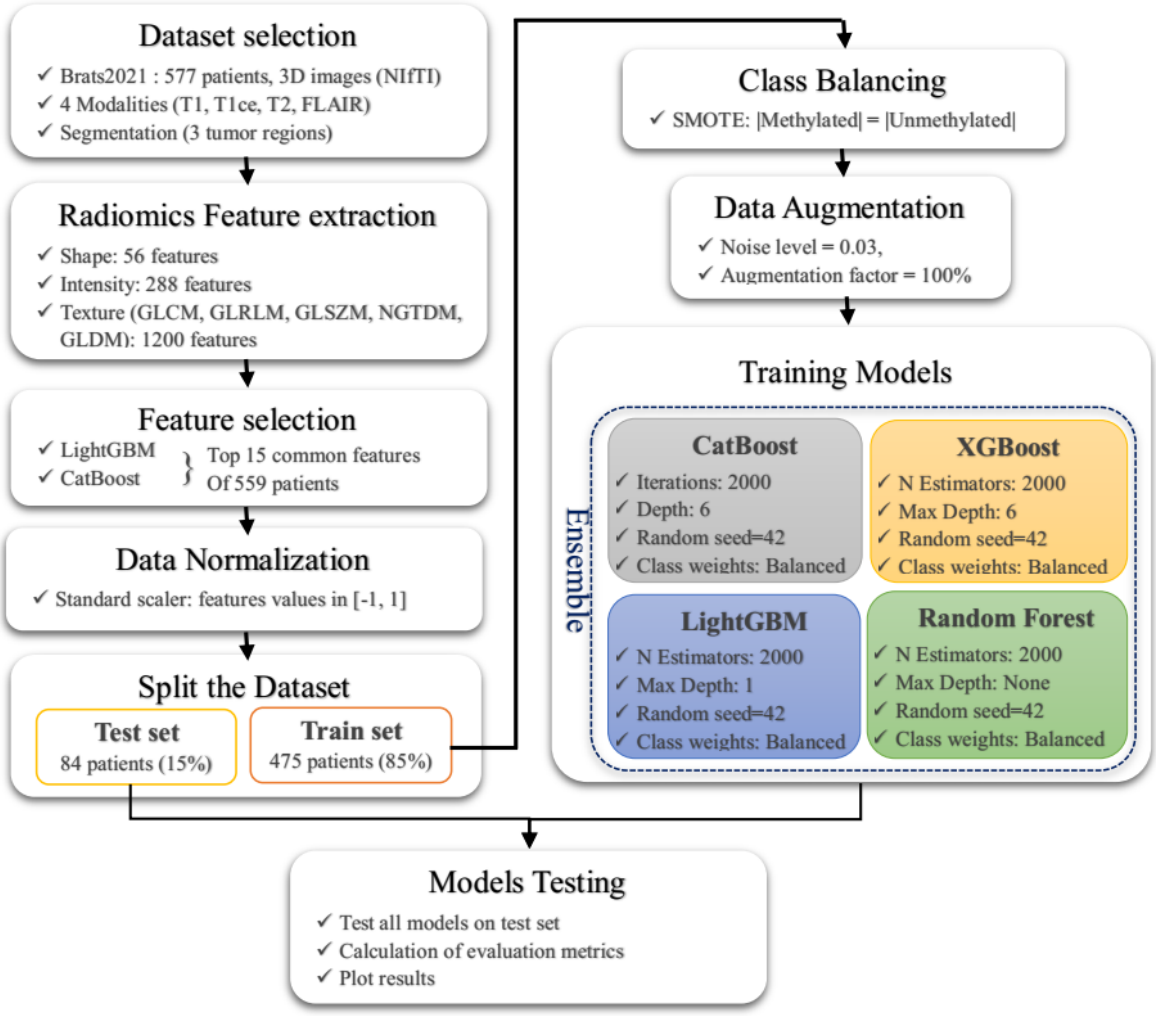


Figure IV.1: The Modeling Pipeline.

The proposed ML pipeline begins with data collection and preprocessing of the BraTS 2021 dataset, where 559 patients' data are retained after removing incomplete or corrupted records. Radiomics features are extracted from the segmented tumor regions across all four MRI sequences, capturing shape, intensity, and texture characteristics. Feature selection is

performed using LightGBM and CatBoost, identifying 15 common predictive features to reduce dimensionality. The data is normalized using a StandardScaler to a range of [-1, 1] and split into training (475 patients $\sim$ 85%) and test (84 patients $\sim$ 15%) sets. To address class imbalance between methylated and unmethylated MGMT statuses, Synthetic Minority Oversampling Technique (SMOTE) is applied to the training set. Additionally, data augmentation introduces Gaussian noise (noise level of 0.03, 100% augmentation factor) to enhance model robustness by increasing the diversity of the training data.

The pipeline culminates in the training of an ensemble model comprising CatBoost, XGBoost, LightGBM, and Random Forest, optimized through cross-validation for superior performance. The ensemble combines predictions to improve accuracy and generalization. The test set is used to evaluate the model, with performance assessed through metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Results are visualized using ROC curves, confusion matrices, feature importance plots, and predicted probability distributions, generated with tools like Matplotlib and Seaborn. This comprehensive pipeline aims to deliver a reliable and interpretable model for predicting MGMT methylation status, potentially aiding clinical decision-making and improving patient outcomes in glioma treatment.

IV.2.1 Dataset Selection

The Brain Tumor Segmentation 2021 dataset is a comprehensive resource designed to advance research in glioblastoma analysis, providing both tumor segmentation maps and MGMT promoter methylation status labels for a cohort of patients. The dataset supports two primary tasks: Task 1 focuses on tumor segmentation for 1,251 patients, while Task 2 addresses the prediction of MGMT methylation status for 585 labeled patients and 87 unlabeled ones. Our analysis concentrates on the 577 patients common to both tasks, comprising 276 with unmethylated MGMT promoters and 301 with methylated promoters. Each patient's data includes high-resolution isotropic voxel images ($240 \times 240 \times 155$) across four MRI modalities (T1-weighted, contrast-enhanced T1, T2-weighted, and fluid-attenuated inversion recovery) alongside segmented masks that delineate tumor subregions, including necrosis (label 1), enhancing tumor (label 2), and edema (label 4). These rich, multi-modal data enable precise characterization of tumor heterogeneity and support the development of robust machine learning models for non-invasive diagnostic applications.

The BraTS2021 dataset's structured annotations and multi-modal imaging provide a unique opportunity to integrate tumor segmentation with molecular biomarker prediction, specifically MGMT methylation status, which is critical for tailoring GBM treatment strategies. By leveraging the 577 labeled patients' data, our study utilizes the detailed segmentation masks to extract radiomic features from the specified tumor subregions across all four MRI modalities. These features capture the complex morphological and textural characteristics of the

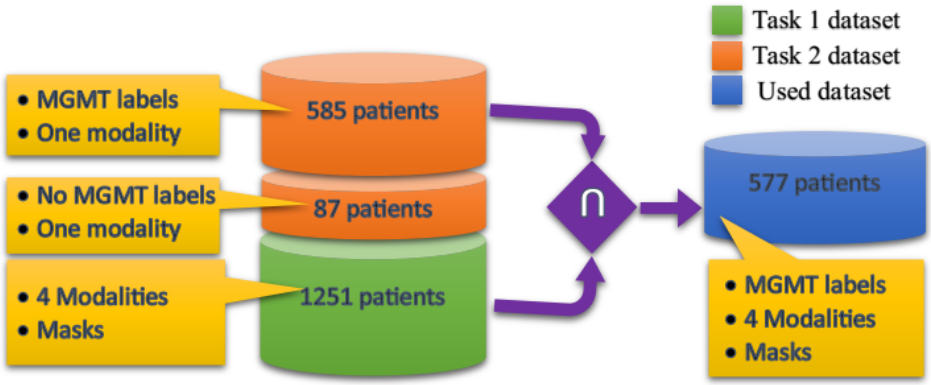


Figure IV.2: Data Selection from Brats 2021.

tumors, facilitating the development of a machine learning pipeline for predicting MGMT methylation status. The combination of high-quality imaging data and molecular labels in the BraTS2021 dataset ensures a robust foundation for building scalable, non-invasive tools that can enhance precision oncology, ultimately improving clinical decision-making and patient outcomes in GBM management.

IV.2.2 Feature Extraction

In this study, radiomics features were extracted from multi-modal MRI scans, including T1W, T1CE, T2W, and FLAIR, for each patient, with a focus on tumor subregions such as necrosis, enhancing tumor, and edema. These features were systematically organized into three primary categories to capture comprehensive characteristics of the tumor and its imaging properties.

Shape-based features describe the 2D and 3D geometry of the tumor, providing insights into its structural characteristics. These features include metrics such as volume, surface area, compactness, sphericity, elongation, flatness, maximum diameter, and surface-to-volume ratio, which collectively define the tumor's physical form and spatial properties.

Intensity-based features, or first-order statistics, capture the distribution of voxel intensities within tumor subregions. These features include metrics like mean, median, standard deviation, skewness, kurtosis, entropy, and energy, which reflect the statistical properties of the imaging data and provide a foundational understanding of intensity variations.

Texture-based features, encompassing second-order and higher-order statistics, analyze spatial patterns and relationships among voxel intensities. These features are further divided into five subcategories: Gray Level Co-occurrence Matrix (GLCM), which quantifies pixel intensity pair frequencies (e.g., contrast, correlation, homogeneity, energy, dissimilarity, angular second moment); Gray Level Run Length Matrix (GLRLM), which measures consecutive voxels with consistent intensity (e.g., short run emphasis, long run emphasis, gray level

non-uniformity, run percentage); Gray Level Size Zone Matrix (GLSZM), which identifies sizes of homogeneous regions (e.g., small area emphasis, large area emphasis, zone entropy); Neighboring Gray Tone Difference Matrix (NGTDM), which evaluates intensity differences between a voxel and its neighbors (e.g., coarseness, contrast, busyness, complexity); and Gray Level Dependence Matrix (GLDM), which focuses on voxel dependencies with similar intensities (e.g., dependence non-uniformity, large dependence emphasis).

Table IV.1: Radiomics Feature Distribution.

Features		Per Modality	Per Region × (4 Modalities)	Total × (3 Regions + Whole)
Shape		–	14	56
Intensity		18	72	288
Texture	GLCM	24	96	384
	GLRLM	16	64	256
	GLSZM	16	64	256
	NGTDM	5	20	80
	GLDM	14	56	224
Grand Total		93	386	1544

The extraction of these radiomics features enables a comprehensive characterization of tumor heterogeneity, facilitating the development of machine learning models for non-invasive prediction of molecular biomarkers in GBM. By systematically capturing geometric, intensity-based, and textural properties across all four MRI modalities and tumor subregions, this approach provides a robust dataset for subsequent feature selection and classification tasks. The use of advanced radiomics tools ensures high-dimensional feature sets that can effectively represent the complex imaging data, enhancing the potential for accurate and scalable diagnostic solutions in precision oncology for GBM management. The next figure shows the steps of radiomics features extraction from selected IRM scans and features selection:

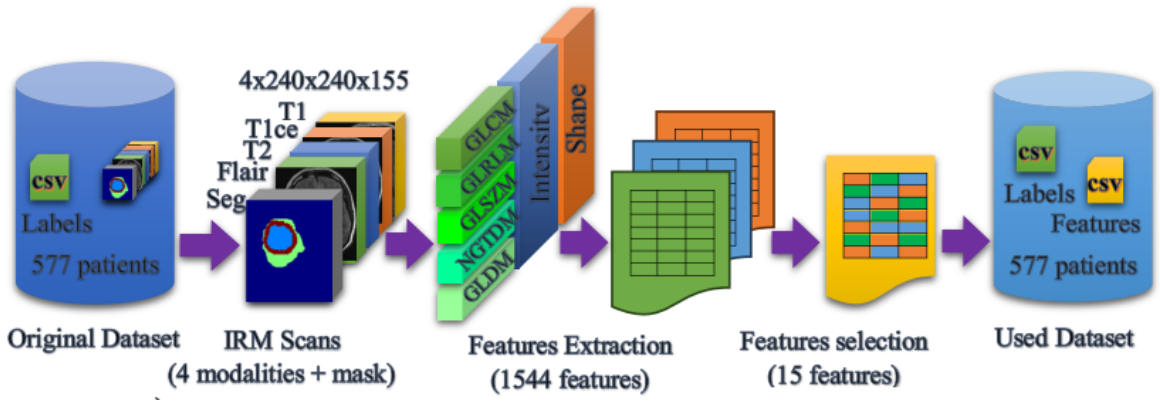


Figure IV.3: Radiomics Features Extraction from IRM Scans.

IV.2.3 Feature Selection

Feature selection is a pivotal step in machine learning, designed to enhance model performance, reduce computational complexity, and improve interpretability by identifying the most relevant features for predictive modeling. In this study, we employ a robust feature selection strategy that leverages the strengths of LightGBM and CatBoost, two advanced gradient-boosting algorithms renowned for their efficiency and accuracy in processing complex datasets. Each algorithm independently evaluates and ranks features based on their importance, which quantifies their contribution to the model's predictive power. By integrating the rankings from both LightGBM and CatBoost, we select the top 15 features that consistently exhibit high importance across both models, ensuring a consensus-driven and reliable selection process that captures the most impactful predictors for our analysis.

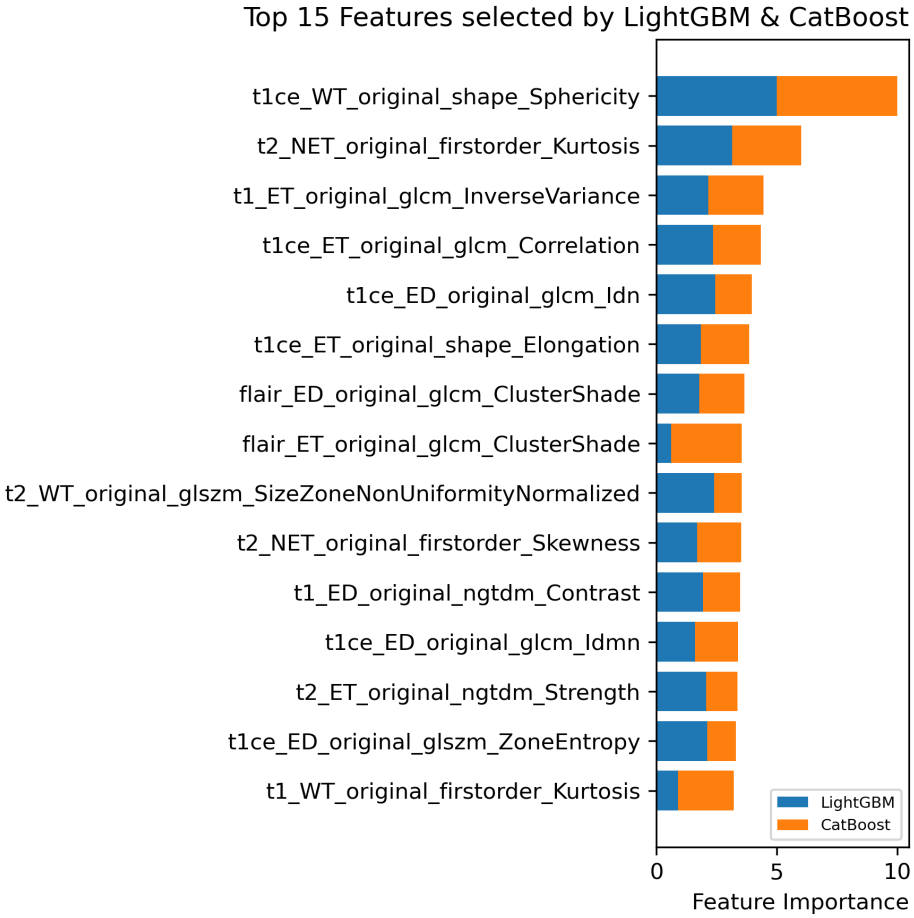


Figure IV.4: Results of Features Selection Steps Using LightGBM and CatBoost.

This combined approach to feature selection enhances the robustness of our machine learning pipeline by mitigating the risk of overfitting and focusing on features with strong predictive value. The selected features, derived from radiomics data extracted from multimodal MRI scans, provide a concise yet comprehensive representation of tumor character-

istics, facilitating accurate prediction of MGMT methylation status in glioblastoma patients. The results of this feature selection process, highlighting the top 15 features and their relative importance, are illustrated in the accompanying figure, offering clear insights into the key drivers of model performance and their relevance to clinical applications in precision oncology.

IV.2.4 Dataset Splitting

The dataset used for predicting MGMT methylation status, stored in CSV format, underwent rigorous preprocessing to ensure quality and reliability for machine learning analysis. Rows and columns with excessive missing values were removed to eliminate incomplete or unreliable entries, while duplicate columns were excluded to reduce redundancy. This process resulted in a refined dataset containing 1,530 features across 559 patients, providing a robust foundation for subsequent analysis by preserving the essential radiomic characteristics extracted from multi-modal MRI scans while minimizing noise and inconsistencies.

After cleaning, the dataset was partitioned into training (85%, ∼475 patients) and test (15%, ∼84 patients) subsets using stratified sampling to maintain the balance of MGMT methylation status, a binary outcome of clinical importance. Stratification ensured that the proportion of methylated and unmethylated cases was preserved in both subsets, thereby supporting representative and unbiased model development. The resulting distribution of MGMT methylation status is illustrated in the accompanying figure, demonstrating the effectiveness of the stratified sampling strategy in maintaining class balance and enabling reliable prediction of methylation status in glioblastoma patients.

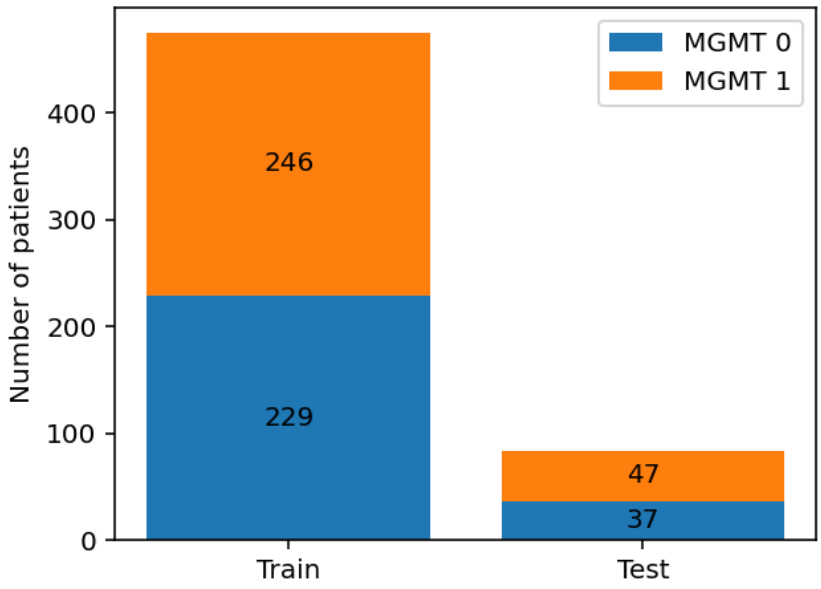


Figure IV.5: Dataset Distribution over Subsets.

IV.2.5 Data Normalization

In the data normalization phase, all feature values in the dataset were converted to the float32 data type to ensure computational efficiency and uniformity across numerical operations. This conversion optimizes memory usage and facilitates seamless processing in machine learning pipelines. Subsequently, the features were normalized using the StandardScaler, which transforms the data to a standardized range of $[-1, 1]$. This process involves centering the data by subtracting the mean and scaling it by the standard deviation, resulting in a dataset with zero mean and unit variance. Such normalization enhances the stability of machine learning algorithms, mitigates the impact of varying feature scales, and accelerates training convergence, thereby improving model performance in predicting MGMT methylation status.

The normalization process is mathematically defined by the following formula:

$$x_{\text{Normalized}} = \frac{x - \mu}{\sigma} \tag{IV.1}$$

where x represents the original feature value, μ is the mean of the feature, and σ is the standard deviation of the feature. This transformation ensures that all features contribute equally to the model, preventing those with larger scales from disproportionately influencing the results. By applying this standardization, the dataset is optimally prepared for subsequent feature selection and classification tasks, supporting robust and reliable predictions in the context of glioblastoma management.

IV.2.6 Class Balancing

To mitigate class imbalance in the MGMT methylation status classification task, we applied the Synthetic Minority Over-sampling Technique (SMOTE). This approach generates new minority class samples by interpolating between existing observations and their nearest neighbors, thereby creating a more balanced training distribution. By enhancing the representation of the minority class, SMOTE reduces model bias toward the majority class and strengthens the ability of machine learning algorithms to capture meaningful patterns.

The use of SMOTE improves the training process by ensuring equitable representation of both methylated and unmethylated cases, which is essential for robust model development. This balance not only enhances model sensitivity and generalization but also provides a stronger foundation for subsequent feature selection and classification, ultimately supporting the design of accurate and reliable predictive tools for glioblastoma management within the framework of precision oncology.

IV.2.7 Data Augmentation

To improve model generalization and overcome the limitations of a constrained dataset, we applied data augmentation by introducing controlled noise into the training subset. This strategy generates subtle perturbations that preserve the semantic integrity of the original radiomic features while increasing sample diversity. By enriching the training data in this way, the models are better equipped to capture underlying patterns and variability relevant to MGMT methylation status prediction in glioblastoma patients.

Incorporating controlled noise as an augmentation method reduces the risk of overfitting and enhances the robustness of the machine learning models. Exposure to a broader range of feature variations strengthens their ability to generalize to unseen data, thereby improving predictive stability. Integrating this augmentation step within the preprocessing pipeline supports the development of scalable and reliable machine learning tools for precision oncology.

IV.2.8 Training Models

In this study, we trained four advanced machine learning models (CatBoost, XGBoost, LightGBM, and Random Forest) on the augmented training subset to predict MGMT methylation status in glioblastoma patients. These models were chosen for their proven effectiveness in handling tabular data and their capability to capture complex, non-linear feature interactions inherent in radiomics datasets. Each model was individually optimized to leverage its unique strengths, such as CatBoost's handling of categorical features, XGBoost's and LightGBM's efficiency in gradient boosting, and Random Forest's robustness through decision tree ensembles, ensuring comprehensive learning from the diverse radiomic features extracted from multi-modal MRI scans.

To further enhance predictive performance and robustness, we developed two ensemble approaches that integrate the predictions of these four models using a soft voting strategy and a stacking strategy. This voting-based ensemble leverages the complementary strengths of the individual models, mitigating their weaknesses and improving overall accuracy in classifying MGMT methylation status. By combining the diverse predictive capabilities of CatBoost, XGBoost, LightGBM, and Random Forest, the ensemble approach ensures a more stable and reliable classification, making it well-suited for clinical applications in precision oncology for glioblastoma management.

IV.2.9 Evaluation & Metrics

To assess the performance of the machine learning models for predicting MGMT methylation status, we evaluated the test subset using a comprehensive set of metrics: accuracy, precision,

area under the receiver operating characteristic curve (AUC-ROC), F1 score, and Cohen's Kappa. These metrics provide a robust evaluation of the models' predictive capabilities, capturing various aspects of performance such as overall correctness, class-specific accuracy, and agreement beyond chance. Additionally, we generated visualizations, including plots of evaluation accuracy, AUC, the ROC curve, and the confusion matrix for the test set, to provide intuitive insights into the models' effectiveness in distinguishing between MGMT methylated and unmethylated cases in glioblastoma patients.

The evaluation metrics are defined as follows:

- **Accuracy.** Accuracy expresses the proportion of correctly predicted instances (both positives and negatives) among the total number of cases. It is intuitive and widely reported, but can be misleading for highly imbalanced datasets because a model can appear "accurate" by simply predicting the majority class:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{IV.2}$$

  where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

- **F_1 Score.** The F_1 score is the harmonic mean of precision and recall. It emphasizes a balance between the two, making it useful when classes are imbalanced or when both false positives and false negatives carry significant cost:

$$\text{F}_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{IV.3}$$

  with:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{IV.4}$$

  High F_1 indicates the model retrieves most positives while maintaining few false alarms.

- **AUC-ROC.** The Area Under the Receiver Operating Characteristic curve evaluates the ability of the model to distinguish between positive and negative classes across different threshold settings. It plots the True Positive Rate (Recall) against the False Positive Rate:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad \text{AUC} = \int_0^1 TPR(FPR)\, d(FPR) \tag{IV.5}$$

  The AUC summarizes the ROC curve as a single scalar; values closer to 1 indicate strong separability, while 0.5 reflects random guessing.

- **Cohen's Kappa.** Cohen's κ accounts for agreement occurring by chance, which is especially informative for imbalanced datasets. A κ of 1 means perfect agreement, 0

means chance-level agreement, and negative values indicate worse than chance:

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \tag{IV.6}$$

where:

$$\begin{cases} P_0 = \frac{\text{TP+TN}}{\text{Total}} \\ P_e = \frac{\text{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}}{\text{Total}^2} \end{cases} \tag{IV.7}$$

This statistic is valuable for measuring how much better the model performs than random chance.

- **Confusion Matrix.** A confusion matrix provides a complete picture of classification results by showing how many instances fall into each predicted/actual category:

Table IV.2: Confusion Matrix.

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Examining the confusion matrix can highlight systematic errors, showing whether the classifier tends to over-predict positives, under-predict negatives, or both.

These metrics and visualizations provide a comprehensive assessment of the models, capturing both strengths and weaknesses beyond a single performance indicator. In precision oncology for glioblastoma management, such evaluation is essential, as treatment decisions often depend on highly sensitive predictive models. By validating performance through multiple metrics and visual tools, researchers and clinicians can trust the models' robustness and generalizability, strengthening their clinical applicability and supporting personalized therapeutic strategies to improve patient outcomes.

IV.3 Contributions

IV.3.1 Random Forest

Random Forests, pioneered by Breiman (2001), advance bagging by incorporating randomness in data sampling and feature selection, fostering diverse decision trees that significantly improve prediction accuracy [35]. This ensemble method is highly effective for classifying MGMT promoter methylation from radiomic features, as it adeptly manages complex, high-dimensional datasets with exceptional robustness and precision. Its ability to minimize

overfitting while maintaining high accuracy makes it particularly valuable for medical applications, such as precision diagnostics, where reliable classification is critical.

The algorithm trains multiple decision trees on bootstrapped subsets of patient data, selecting random feature subsets at each split to reduce inter-tree correlation. Predictions from all trees are combined through majority voting to yield a stable final classification. Out-of-bag (OOB) samples, excluded from individual tree training, provide an unbiased estimate of generalization error, eliminating the need for a separate validation set. This efficient strategy ensures accurate and reliable results, making Random Forests a strong model for MGMT methylation prediction.

Random Forests effectively balance bias and variance, exhibit strong resistance to overfitting, and produce feature importance rankings that identify the most influential radiomic predictors of MGMT status. These interpretable insights improve model transparency and can uncover biologically meaningful patterns within the data. Consequently, Random Forests represent a dependable, interpretable, and robust approach for non-invasive MGMT methylation status classification.

Algorithm 19 Random Forest for MGMT Status Prediction

function FIT(Radiomic features X, MGMT labels y, number of trees M, number of features d_{try})
  Initialize forest $\mathcal{F} = \emptyset$
  for $m = 1, 2, \ldots, M$ **do**
    Sample bootstrap dataset D_m of size N from (X, y) ▷ sample with replacement
    Grow an unpruned decision tree T_m on D_m:
    for each node in the tree **do**
      Randomly select d_{try} features from all p features
      Choose the best split among these d_{try} features
      Split the node into two child nodes
    end for
    Add tree T_m to the forest $\mathcal{F}$
  end for
  return forest $\mathcal{F} = \{T_1, T_2, \ldots, T_M\}$
end function
function PREDICT($\mathcal{F}, X$)
  for each sample $x_i \in X$ **do**
    Obtain predictions $\hat{y}_i^{(m)} = T_m(x_i)$ for all $m = 1, \ldots, M$
    $\hat{y}_i = \begin{cases} \text{mode}\{\hat{y}_{i,1}, \ldots, \hat{y}_{i,M}\} & \text{mode = “classification”} \\ \frac{1}{M}\sum_{m=1}^{M} \hat{y}_{i,m} & \text{mode = “regression”} \end{cases}$
  end for
  return aggregated predictions $\hat{y}$ as MGMT methylation prediction.
end function

IV.3.2 LightGBM

LightGBM, a gradient boosting framework optimized for high-dimensional datasets, offers faster training and lower memory usage compared to traditional boosting methods [38]. Its efficiency makes it ideal for predicting MGMT promoter methylation status from radiomic features in medical applications. By leveraging innovative techniques, LightGBM handles complex datasets with speed and precision. This enables robust classification in precision diagnostics. Its low resource demands enhance scalability for large datasets. Overall, LightGBM is a powerful tool for accurate medical predictions.

The method builds trees sequentially, with each tree correcting errors from the previous one. Key features include: (i) histogram-based splitting to accelerate feature partitioning, (ii) leaf-wise growth for optimal loss reduction, (iii) Gradient-based One-Side Sampling (GOSS) to emphasize hard-to-classify cases, and (iv) Exclusive Feature Bundling (EFB) to reduce dimensionality by merging sparse features. These innovations enable efficient and accurate predictions. LightGBM's design makes it particularly effective for MGMT methylation classification tasks, balancing precision and computational speed. LightGBM offers a strong combination of speed, scalability, and predictive accuracy, while also providing feature importance metrics that reveal the most relevant radiomic biomarkers associated with MGMT methylation. This interpretability enhances its clinical relevance, supporting radiogenomic studies aimed at non-invasive molecular profiling. Overall, LightGBM stands out as a robust, interpretable, and efficient approach for MGMT methylation status detection.

Algorithm 20 LightGBM for MGMT Status Prediction

function FIT(Radiomic features X, MGMT labels y, loss function L, number of rounds T, learning rate η)
  Bin continuous features into histograms.
  Initialize predictions: $\hat{y}_i^{(0)} = 0$ or a constant.
  for $t = 1, 2, \ldots, T$ **do**
    Compute gradients g_i and Hessians h_i.
    Apply GOSS to select samples with large gradients Equation (II.50).
    Apply exclusive feature bundling to bundle mutually exclusive features.
    Find optimal splits using histogram-based search and Equation (II.49).
    Grow tree leaf-wise $T_m(x)$, selecting the leaf with maximum loss reduction, subject to depth/leaf constraints.
    Update predictions: $f_m(x) = f_{m-1}(x) + \eta\, T_m(x)$.
  end for
  return $\{f_1, \ldots, f_T\}$
end function
function PREDICT($\{f_1, \ldots, f_T\}, X$)
  $\hat{y} = \hat{y}^{(0)} + \eta \sum_{t=1}^{T} f_t(X)$.
  return Final MGMT methylation prediction $\hat{y}$.
end function

IV.3.3 XGBoost

XGBoost, a highly efficient gradient boosting framework, excels in structured data tasks due to its scalability and robust performance [37]. Its computational speed and strong regularization make it ideal for biomedical applications, such as predicting MGMT promoter methylation status from radiomic features. By effectively handling complex datasets, XGBoost ensures accurate and reliable classifications. Its design supports large-scale data processing with minimal resource demands. This makes it a top choice for precision medicine tasks. XGBoost's versatility enhances its utility in medical diagnostics.

The algorithm constructs regression trees sequentially, each correcting prior errors by optimizing a regularized loss function. Key innovations include: (i) L1 and L2 regularization to prevent overfitting; (ii) second-order optimization using gradient and m for precise updates; (iii) efficient split finding with weighted quantile sketches; and (iv) native support for missing and sparse data, ideal for radiomic datasets. Parallelization and GPU acceleration further boost efficiency. These features ensure XGBoost delivers fast, accurate predictions for MGMT methylation classification.

Algorithm 21 XGBoost for MGMT Status Prediction

function FIT(Radiomic features X, MGMT labels y, loss function L, number of rounds T, learning rate η)
  Initialize predictions: $\hat{y}_i^{(0)} = 0$ or a constant.
  for $t = 1, 2, \ldots, T$ **do**
    Compute gradients g_i and Hessians h_i (Equation (II.49)).
    Use weighted quantile sketch to find candidate split points.
    For each node, evaluate splits using Equation (II.49) and select the best.
    Grow tree level-wise, assigning samples to leaves based on splits.
    Compute optimal leaf weights using Equation (II.53).
    Update predictions: $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$.
  end for
  return $\{f_1, \ldots, f_T\}$
end function
function PREDICT($\{f_1, \ldots, f_T\}, X$)
  $\hat{y} = \hat{y}^{(0)} + \eta \sum_{t=1}^{T} f_t(X)$.
  return Final MGMT methylation prediction $\hat{y}$
end function

XGBoost also provides *sparsity-aware split finding*, which assigns missing radiomic values to the most informative branch during training. This property is particularly advantageous for clinical datasets that may suffer from incomplete acquisition or preprocessing. Furthermore, the model's feature importance measures (based on split gain and frequency) can help identify radiomic biomarkers most predictive of MGMT methylation, supporting interpretability and clinical validation.

By combining regularization, second-order optimization, and computational scalability, XGBoost represents a robust and interpretable tool for non-invasive prediction of MGMT methylation status from high-dimensional radiomic data.

IV.3.4 CatBoost

CatBoost is a gradient boosting framework designed to improve performance on tabular datasets, particularly those with categorical features [39]. While radiomic features are predominantly numerical, CatBoost's innovations (such as ordered boosting and symmetric trees) make it well-suited for biomedical prediction tasks like determining the methylation status of the MGMT gene promoter. Its ability to reduce overfitting, handle noisy features, and deliver robust predictions is particularly relevant in medical imaging contexts, where datasets may be heterogeneous and moderately sized.

Algorithm 22 CatBoost for MGMT Status Prediction

function FIT(Radiomic features X, MGMT labels y, loss function L, number of rounds T, learning rate η)
  Generate random permutations of training data (X, y).
  Initialize predictions: $\hat{y}_i^{(0)} = 0$ or a constant.
  for $t = 1, 2, \ldots, T$ **do**
    Compute gradients g_i and Hessians h_i (Equation (II.49)).
    Encode categorical features using ordered target statistics Equation (II.59).
    Find optimal splits for symmetric trees using Equation (II.49).
    Grow symmetric tree f_t, applying the same split at each level.
    Compute optimal leaf weights using Equation (II.57).
    Update predictions: $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$.
  end for
  return $\{f_1, \ldots, f_T\}$
end function
function PREDICT($\{f_1, \ldots, f_T\}, X$)
  $\hat{y} = \hat{y}^{(0)} + \eta \sum_{t=1}^{T} f_t(X)$.
  return Final MGMT methylation prediction $\hat{y}$
end function

The method constructs an ensemble of decision trees sequentially, with each new tree minimizing a regularized loss function by correcting errors of its predecessors. Its key innovations include: (i) ordered boosting, which prevents target leakage by computing target statistics across different permutations of the data; (ii) ordered target statistics for efficient categorical encoding, ensuring unbiased feature transformations; and (iii) symmetric (oblivious) trees, which apply the same split at each level, leading to simpler models and faster inference. CatBoost also supports GPU acceleration, making it scalable for larger radiomic cohorts.

In the context of MGMT methylation detection, CatBoost's ordered boosting reduces the risk of overfitting on limited patient cohorts, while symmetric trees ensure computational efficiency and model interpretability. Feature importance analysis derived from split gains can highlight radiomic biomarkers most associated with MGMT status, aiding biological insight and potential clinical translation. Its robustness to noisy features and balanced predictive performance make CatBoost a valuable candidate for non-invasive glioblastoma biomarker prediction.

IV.3.5 Voting Ensemble

The Voting Ensemble is one of the simplest yet effective ensemble strategies, combining predictions from multiple base learners to improve robustness and generalization. For the task of predicting the methylation status of the MGMT gene promoter, we integrate four complementary classifiers: Random Forest (RF), CatBoost, XGBoost, and LightGBM. Each of these learners captures different aspects of the radiomic feature space: RF provides robustness to noise and high-dimensionality, CatBoost mitigates overfitting and handles complex feature interactions, XGBoost incorporates regularization and second-order optimization, while LightGBM offers high efficiency and scalability.

The principle of voting is to aggregate individual predictions into a final decision, either by *hard voting* (majority class) or *soft voting* (averaged predicted probabilities). This strategy is computationally inexpensive and leverages the diversity of the base models without requiring a meta-learner. In biomedical applications such as MGMT methylation detection, this can provide more stable predictions across heterogeneous patient populations.

Algorithm 23 Voting Ensemble for MGMT Status Prediction

function FIT(Radiomic features X, MGMT labels y, base learners $\{RF, CatBoost, XGBoost, LightGBM\}$)
  for each $model_i$ in base learners **do**
    h_i = Train the $model_i$ on (X, y) independently to others base learners.
  end for
  return $\{h_1, h_2, h_3, h_4\}$
end function
function PREDICT($\{h_1, h_2, h_3, h_4\}$, X, mode)
  if mode = "hard" **then**
    $\hat{y}_m = h_m(X)$ for each trained base learner h_m
    $\hat{y} = \text{mode}(\hat{y}_1, \ldots, \hat{y}_M)$
  else if mode = "soft" **then**
    Obtain probability estimates $p_m(X)$ from each trained base learner h_m
    $\hat{y} = \frac{1}{M} \sum_{m=1}^{M} p_m(X)$
  end if
  return $\hat{y}$
end function

By combining RF, CatBoost, XGBoost, and LightGBM through voting, the ensemble balances variance reduction, bias control, and computational efficiency. Although less flexible than stacking, this method offers an interpretable and practical approach for non-invasive MGMT status determination. Its effectiveness stems from the complementary strengths of the base learners, which together enhance predictive reliability in radiomic-based glioblastoma biomarker analysis.

IV.3.6 Stacking Ensemble

Stacking is an ensemble learning strategy that combines the predictive strengths of diverse base learners by training a higher-level *meta-learner*. Unlike bagging, which reduces variance, or boosting, which reduces bias, stacking emphasizes the integration of heterogeneous models to improve generalization. For the task of predicting MGMT promoter methylation status from radiomic features, stacking enables the complementary use of Random Forest, CatBoost, XGBoost, and LightGBM within a unified predictive framework.

Algorithm 24 Stacking Ensemble for MGMT Methylation Status Prediction

function FIT(Radiomic features X, MGMT labels y, base learners $\{RF, CatBoost, XGBoost, LightGBM\}$, Train meta-learner $\mathcal{H}$ ($Logistic\ Regression$))
  Split (X, y) into k folds for cross-validation.
  for each base learner $model_i$ in $\{RF, CatBoost, XGBoost, LightGBM\}$ **do**
    for each fold $f = 1, \ldots, k$ **do**
      Train $model_i$ on training folds $(X^{(f)}_{train}, y^{(f)}_{train})$.
      Predict probabilities $\hat{y}^{(f)}_{val,i}$ on validation fold $X^{(f)}_{val}$.
    end for
    Concatenate all $\hat{y}^{(f)}_{val,i}$ to form meta-feature column Z_i.
  end for
  Form meta-feature matrix $Z = [Z_{RF}, Z_{CatBoost}, Z_{XGBoost}, Z_{LightGBM}]$.
  Train meta-learner $\mathcal{H}$ on (Z, y).
  return Trained base learners $\{RF, CatBoost, XGBoost, LightGBM\}$ and meta-learner $\mathcal{H}$.
end function
function PREDICT($\{RF, CatBoost, XGBoost, LightGBM\}$, $\mathcal{H}$, new data X_{test})
  for each trained base learner h_i in $\{RF, CatBoost, XGBoost, LightGBM\}$ **do**
    Obtain predicted probabilities $p_i(X_{test})$.
  end for
  Construct meta-feature matrix $Z_{test} = [p_{RF}, p_{CatBoost}, p_{XGBoost}, p_{LightGBM}]$.
  Predict final probabilities $\hat{y} = \mathcal{H}(Z_{test})$.
  return Final MGMT methylation prediction $\hat{y}$.
end function

The key idea is to use the predictions of the base learners as inputs for the meta-learner, which learns how to optimally combine them. This allows the ensemble to exploit different modeling biases: Random Forest contributes robustness to noise and interpretability; Cat-

Boost improves handling of categorical-like encodings and reduces overfitting; XGBoost provides strong regularization and precise optimization; and LightGBM contributes efficiency and scalability on high-dimensional radiomic data.

Stacking often achieves superior predictive accuracy for MGMT detection, since each base model captures distinct radiomic patterns. The meta-learner (commonly logistic regression or a shallow neural network) assigns appropriate weights to the base predictions, reducing individual model weaknesses. However, stacking is computationally more demanding than simple ensembles, requires robust cross-validation to prevent information leakage, and must be carefully regularized to avoid overfitting. Despite these challenges, stacking provides a powerful and interpretable strategy for non-invasive MGMT status prediction.

IV.4 Conclusion

In this chapter, we introduced a complete framework for detecting MGMT promoter methylation status in glioblastoma patients. Starting from dataset selection and radiomic feature extraction, followed by preprocessing steps such as feature selection, normalization, class balancing, and augmentation, we established a reliable dataset for experimentation. On this basis, we trained and evaluated several models of machine learning, including individual classifiers (Random Forest, LightGBM, XGBoost, and CatBoost) as well as ensemble strategies (Voting and Stacking). The outcomes of these approaches, along with their comparative analysis, are presented and discussed in the following chapter.

CHAPTER

5 Results and Discussions

V.1 Introduction

In this study, we trained and evaluated four machine learning models (CatBoost, XGBoost, LightGBM, and Random Forest) as base learners on identical data subsets to predict MGMT methylation status in glioblastoma patients. We then developed two ensemble learning approaches (voting, and stacking) to integrate these models, aiming to enhance predictive performance beyond that of any single model. By leveraging the complementary strengths of the individual models, our goal was to achieve superior accuracy, robustness, and generalization. We present and discuss the performance of each base learner and compare the effectiveness of the ensemble methods.

In this evaluation, the first step focused on radiomics feature selection to identify the most informative predictors for classifying MGMT methylation status in glioblastoma patients. To achieve this, we ranked radiomics features according to their importance scores and selected the top-performing ones. These features were then examined with respect to their distribution across different families, including shape descriptors, first-order intensity statistics, and higher-order texture features. The results of this selection process are presented in the following section.

V.2 Radiomics Feature Selection

The study employed a recursive feature elimination strategy guided by feature importance scores to identify the most relevant radiomics features for binary classification. Six machine learning algorithms (CatBoost, XGBoost, LightGBM, Random Forest, Voting Ensemble, and

Stacking Ensemble) were systematically tested to evaluate the robustness of the selected features. During this process, features deemed less informative were iteratively removed, while the predictive performance of each model was assessed at different feature counts, ranging from 30 down to 5. The evolution of performance metrics, particularly the Area Under the Curve (Figure V.1, Figure V.3) and accuracy (Figure V.2, Figure V.4), illustrates the impact of feature selection on classification outcomes.

Results revealed that across all models, the most stable and optimal predictive performance was consistently achieved when 15 radiomics features were retained. This feature subset balanced model complexity with predictive strength, avoiding overfitting from excessive features while preventing performance degradation from too few. The final selection of the top 15 features is detailed in Table V.1 and visually represented in Figure IV.4, providing a concise yet powerful set of predictors for downstream analysis. These findings highlight the importance of rigorous feature selection in improving classification accuracy and ensuring the generalizability of radiomics-based machine learning models.

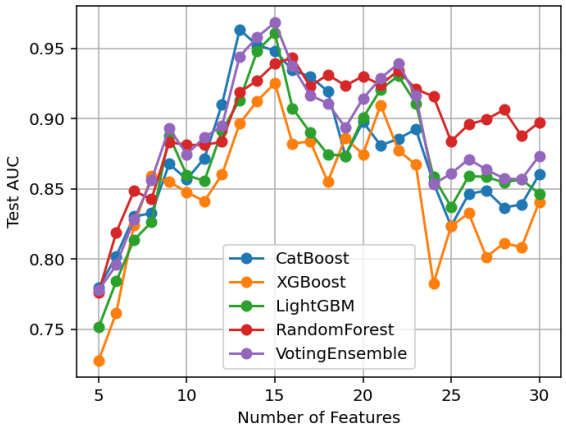


Figure V.1: AUC of Voting Model Using Different Numbers of Features.

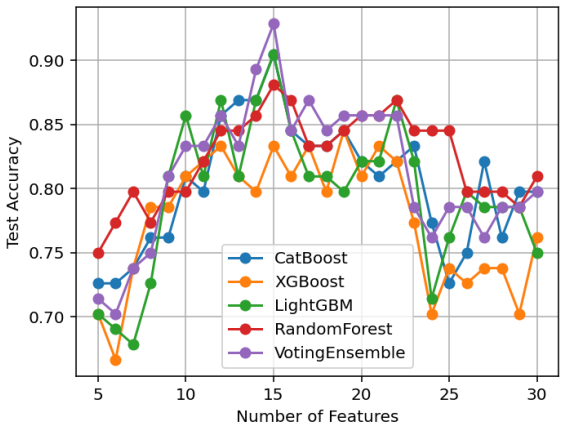


Figure V.2: Accuracy of Voting Model Using Different Numbers of Features.

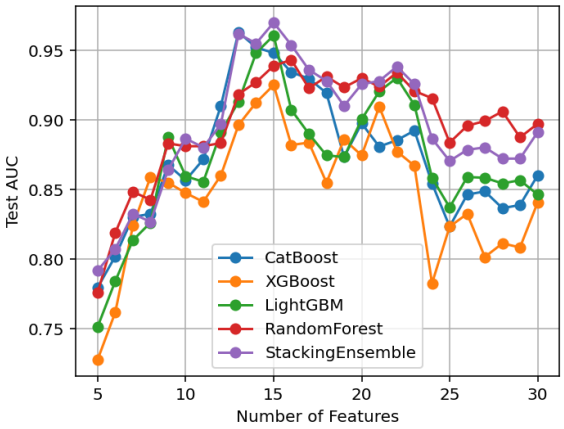


Figure V.3: AUC of Stacking Model Using Different Numbers of Features.

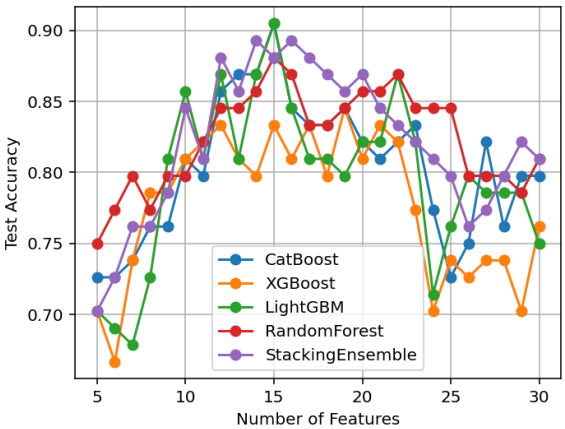


Figure V.4: Accuracy of Stacking Model Using Different Numbers of Features.

V.3 Feature Distribution

The Table V.1 lists the fifteen radiomics features ultimately retained for the modeling stage, showing how they are distributed across imaging modalities, tumor sub-regions, and feature families. Several MRI sequences contribute complementary information: T1-weighted with contrast appears most frequently, underscoring its value in highlighting tumor vascularity and enhancing regions, while T2-weighted and FLAIR add sensitivity to edema and tissue heterogeneity. The inclusion of features from the whole tumor, non-enhanced tumor, enhanced tumor, and edema regions ensures that both the core lesion and peritumoral environment are quantitatively described, which is essential for characterizing glioma heterogeneity and potential methylation patterns.

Table V.1: Top 15 Radiomics Features Selected.

Order	Modality	Region	Feature Class	Feature Name
1	T1CE	Whole Tumor	Shape	Sphericity
2	T2W	Non-Enhanced Tumor	First-order	Kurtosis
3	T1W	Enhanced Tumor	GLCM	Inverse Variance
4	T1CE	Enhanced Tumor	GLCM	Correlation
5	T1CE	Edema	GLCM	Idn
6	T1CE	Enhanced Tumor	Shape	Elongation
7	FLAIR	Edema	GLCM	Cluster Shade
8	FLAIR	Enhanced Tumor	GLCM	Cluster Shade
9	T2W	Whole Tumor	GLSZM	Size Zone Non-Uniformity Normalized
10	T2W	Non-Enhanced Tumor	First-order	Skewness
11	T1W	Edema	NGTDM	Contrast
12	T1CE	Edema	GLCM	Idmn
13	T2W	Enhanced Tumor	NGTDM	Strength
14	T1CE	Edema	GLSZM	Zone Entropy
15	T1W	Whole Tumor	First-order	Kurtosis

From a feature-type perspective, the selection is balanced across shape, first-order intensity statistics, and higher-order texture families (GLCM, GLSZM, NGTDM), indicating that both geometry and intra-lesional texture contribute to discriminative power. For instance, “Sphericity” and “Elongation” capture morphological complexity, while “Inverse Variance,” “Correlation,” and “Cluster Shade” reflect subtle gray-level dependencies. The presence of “Zone Entropy” and “Size Zone Non-Uniformity Normalized” highlights the importance of heterogeneity measures, and first-order metrics like “Kurtosis” and “Skewness” provide insights into intensity distribution asymmetry. Overall, the table demonstrates a deliberate fea-

ture selection strategy aimed at integrating spatial, statistical, and textural descriptors across sequences, thereby maximizing the ability of the subsequent models to detect methylation-related imaging phenotypes.

V.4 Models Performances

In the following analysis, the performance of the selected machine learning models was rigorously examined to determine their ability to classify methylation status with reliability and precision. CatBoost, XGBoost, and LightGBM represent gradient-boosting approaches known for effectively handling non-linear relationships and feature interactions, while Random Forest provides a bagging-based ensemble benchmark. The inclusion of a Voting Ensemble further leverages the collective strengths of individual models to potentially enhance overall predictive stability. By applying these techniques to the curated set of 15 radiomics features from 84 patients, the evaluation aimed to reflect real-world predictive scenarios, balancing model complexity with clinical interpretability.

A comprehensive suite of metrics was employed to ensure a multidimensional performance assessment. Accuracy and AUC were used to measure overall correctness and the ability to discriminate between methylated and unmethylated cases across different thresholds. Precision, recall, and F1 score offered a closer examination of how well the models managed false positives and false negatives, particularly important when class distribution or clinical consequences vary between classes. Cohen's kappa was incorporated to quantify the agreement between model predictions and the reference standard beyond random chance, adding robustness to the evaluation. Table V.2 consolidates these results, enabling clear comparison of individual models and the ensemble approach in terms of both discriminative power and reliability.

Table V.2: Test Performance Metrics (%).

Model	Accuracy	AUC	Precision	Recall	F1 Score	Kappa
CatBoost	90.48	94.82	86.79	**97.87**	92.00	80.34
XGBoost	83.33	92.52	83.67	87.23	85.42	65.99
LightGBM	90.48	96.09	88.24	95.74	91.84	80.45
Random Forest	88.10	93.90	83.64	**97.87**	90.20	75.28
Voting Ensemble	**92.86**	96.84	**90.20**	**97.87**	**93.88**	**85.34**
Stacking Ensemble	88.10	**97.01**	83.64	**97.87**	90.20	75.28

The results presented in Table V.2 provide a comprehensive comparison of the performance of several machine learning models, including individual classifiers and ensemble methods, across multiple evaluation metrics. Looking first at accuracy, the Voting Ensemble

achieved the highest score (92.86%), outperforming both CatBoost and LightGBM (90.48% each). In contrast, XGBoost obtained the lowest accuracy (83.33%), suggesting weaker overall classification capability on this dataset. The Stacking Ensemble, despite its more complex design, reached only 88.10% accuracy, indicating that stacking may not have added meaningful value compared to simpler ensembles.

When considering the Area Under the Curve (AUC), the Stacking Ensemble achieved the best result (97.01%), reflecting its strength in distinguishing between positive and negative cases in a threshold-independent manner. LightGBM followed closely with 96.09%, while Random Forest showed the weakest performance at 93.90%. This finding indicates that while stacking may not yield superior classification accuracy, it improves the reliability of probability estimation and ranking, which can be valuable in risk stratification contexts.

Precision results highlight the advantage of the Voting Ensemble, which achieved the highest value (90.20%), reducing the likelihood of false positives. CatBoost (86.79%) and LightGBM (88.24%) performed comparably well, whereas Random Forest (83.64%) and XGBoost (83.67%) lagged behind. On the other hand, recall values were very strong for most models, with CatBoost, Random Forest, Stacking Ensemble, and Voting Ensemble all reaching 97.87%, demonstrating high sensitivity to positive cases. LightGBM followed slightly lower at 95.74%, while XGBoost trailed with 87.23%, suggesting a higher tendency to miss positive cases.

The F1 score, which balances precision and recall, again favored the Voting Ensemble (93.88%), marking it as the most well-rounded model. LightGBM (91.84%) and CatBoost (92.00%) also demonstrated strong balance, whereas Random Forest and Stacking Ensemble both achieved 90.20%. XGBoost remained the weakest with an F1 of 85.42%, further underscoring its relative underperformance. Cohen's Kappa values supported this trend: the Voting Ensemble achieved the highest agreement with true labels beyond chance (85.34%), followed by LightGBM (80.45%) and CatBoost (80.34%). Random Forest and Stacking Ensemble were lower at 75.28%, while XGBoost scored only 65.99%, highlighting its limited reliability.

Overall, the results indicate that the Voting Ensemble consistently delivers the most robust and balanced performance across all metrics, making it the most effective approach for this task. Stacking, while excelling in AUC, failed to outperform Voting in terms of accuracy, F1, or Kappa, which limits its practical utility in direct classification. LightGBM and CatBoost proved to be strong individual learners, with complementary strengths in precision, recall, and AUC, whereas Random Forest provided solid recall but lower overall balance. XGBoost, however, underperformed across all dimensions and appears less suitable without further optimization. These findings highlight that ensemble learning, particularly through Voting, provides the best trade-off between sensitivity and reliability, which is especially critical in applications such as medical diagnosis.

V.4.1 Base Learners Performances

CatBoost achieved strong performance overall base learner models, with an accuracy of 90.48% and an AUC of 94.82%. Its recall was among the highest at 97.87%, indicating that the model was very effective at identifying positive cases. Precision, however, was slightly lower at 86.79%, suggesting a greater tendency toward false positives compared to ensemble methods. The F1 score of 92.00% and Kappa value of 80.34% confirm that CatBoost maintains a good balance between sensitivity and reliability. This makes it a competitive individual learner, particularly in scenarios where recall is prioritized.

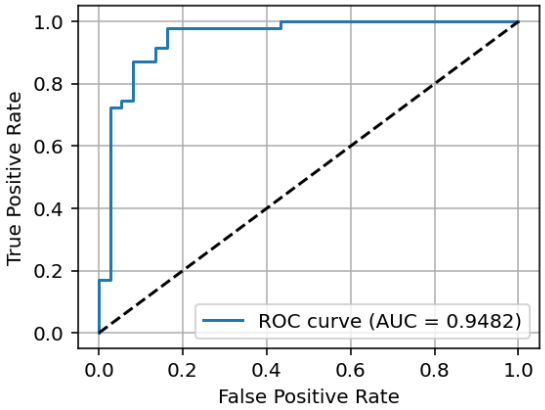


Figure V.5: CatBoost ROC Visualization.

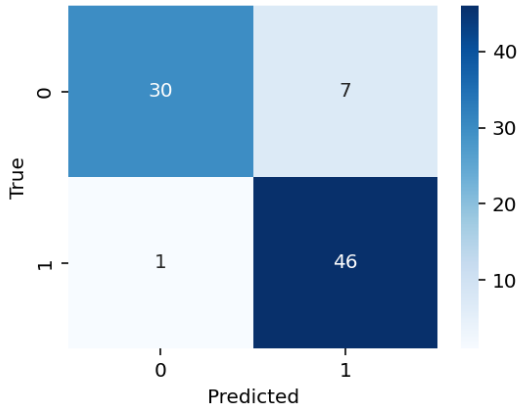


Figure V.6: CatBoost Confusion Matrix Visualization.

This ROC curve (Figure V.5) of the CatBoost model shows strong classification ability, with an AUC of 0.9482. The curve is close to the upper-left corner, indicating excellent sensitivity and specificity. This confirms that CatBoost effectively distinguishes between classes, though complementary evaluations such as the confusion matrix are needed for a complete assessment.

This confusion matrix (Figure V.6) shows that the CatBoost model correctly classified most cases, with 30 true negatives and 46 true positives. Only 7 false positives and 1 false negative were observed, highlighting its high recall (very few missed positives) and strong overall reliability. These results align with the model's high ROC-AUC score and F1 score, confirming its effectiveness for this classification task.

XGBoost , in contrast, showed the weakest performance across nearly all metrics. With an accuracy of 83.33% and an AUC of 92.52%, it underperformed relative to the other gradient boosting methods. Its recall of 87.23% was notably lower than all other models, indicating a higher risk of missing positive cases. Precision was 83.67%, also modest, leading to the lowest F1 score (85.42%) and Kappa (65.99%). These results suggest that XGBoost, in its current configuration, may not be well suited for this dataset without further hyperparameter

tuning or feature optimization.

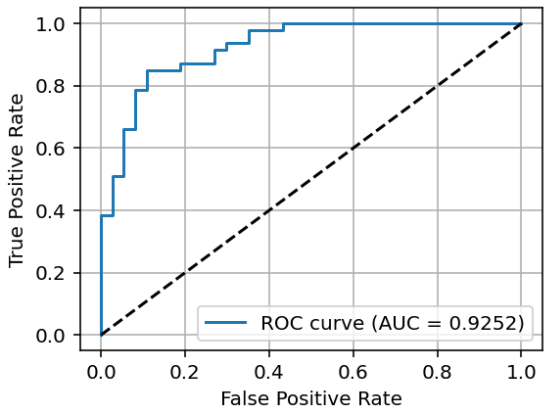


Figure V.7: XGBoost ROC Visualization .

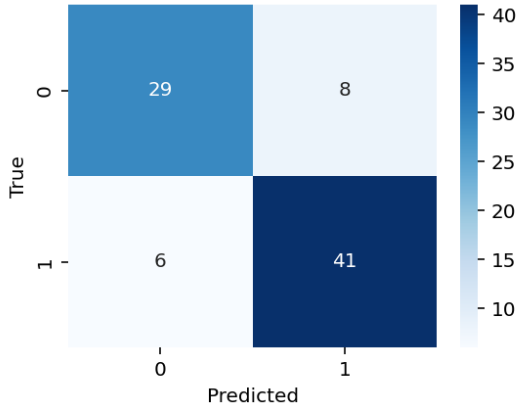


Figure V.8: XGBoost Confusion Matrix Visualization.

The ROC curve in Figure V.7 shows that the XGBoost model has strong discriminative power, with an AUC of 0.9252. The curve rises sharply toward the top-left corner, indicating high sensitivity and specificity across thresholds. This high AUC value confirms the model's robustness and reliability for predicting MGMT promoter methylation status.

Figure V.8 complements this by detailing classification outcomes. The model achieved 29 true negatives and 41 true positives, with only 8 false positives and 6 false negatives. This balance reflects both high sensitivity and specificity, confirming good predictive accuracy. While some errors remain, the overall performance supports XGBoost as a strong tool for aiding clinical decision-making.

LightGBM delivered competitive results, achieving the same accuracy as CatBoost (90.48%) but with a superior AUC of 96.09%. Its recall of 95.74% was slightly lower than CatBoost, but it compensated with a higher precision of 88.24%. The resulting F1 score of 91.84% and Kappa of 80.45% confirm that LightGBM provides a balanced trade-off between sensitivity and specificity. Its strong AUC also highlights its effectiveness in ranking and probability estimation, making it a solid choice as a standalone model.

The ROC curve in Figure V.9 highlights the excellent discriminative ability of the LightGBM model, with an AUC of 0.9609. The curve approaches the top-left corner, confirming very high sensitivity and specificity across thresholds. This performance surpasses typical benchmarks, showing that LightGBM is highly effective in distinguishing between methylated and unmethylated MGMT promoter cases, making it a strong candidate for clinical applications.

In Figure V.10, the confusion matrix provides a detailed breakdown of predictions. The model correctly classified 31 true negatives and 45 true positives, with only 6 false positives

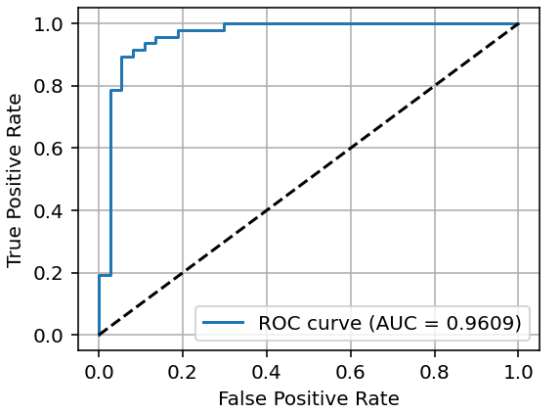


Figure V.9: LightGBM ROC Visualization .

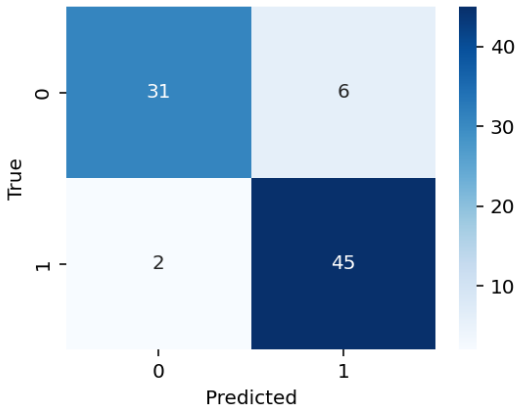


Figure V.10: LightGBM Confusion Matrix Visualization.

and 2 false negatives. These results demonstrate strong precision and recall, with very few misclassifications. Compared to XGBoost, LightGBM achieves slightly better balance and fewer errors, reinforcing its reliability and robustness for non-invasive biomarker prediction in glioma studies.

Random Forest offered more moderate performance. Its accuracy was 88.10%, with the lowest AUC of all models (93.90%). However, like CatBoost, it achieved very high recall (97.87%), showing strong sensitivity to positive cases. This came at the expense of precision, which was only 83.64%, leading to a lower F1 score of 90.20% and a Kappa of 75.28%. These results suggest that while Random Forest is effective at minimizing false negatives, it tends to generate more false positives, limiting its balance compared to boosting methods.

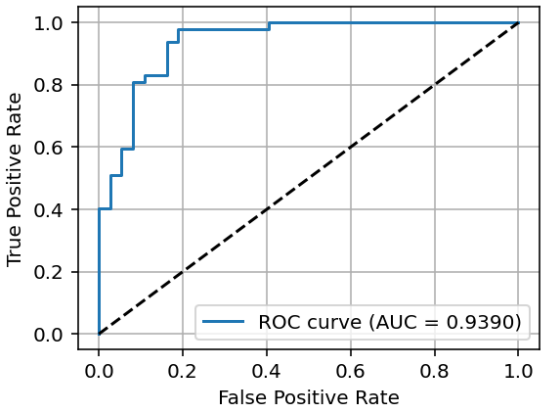


Figure V.11: Random Forest ROC Visualization.

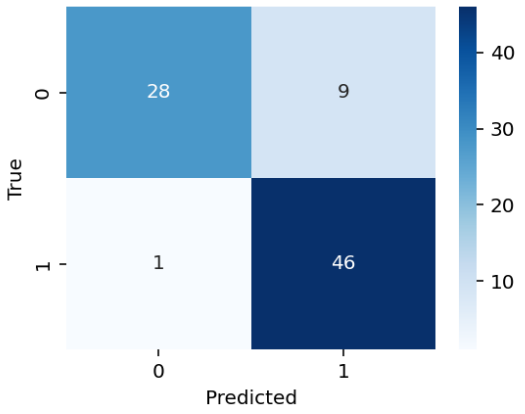


Figure V.12: Random Forest Confusion Matrix Visualization.

The ROC curve in Figure V.11 shows that the Random Forest model performs well, with an AUC of 0.9390. The curve rises steeply toward the upper-left corner, reflecting a strong balance of sensitivity and specificity. Although the AUC is slightly lower than that of Light-

GBM, it still indicates excellent discriminative capability, making Random Forest a reliable approach for MGMT promoter methylation prediction.

The confusion matrix in Figure V.12 reveals that the model correctly identified 28 true negatives and 46 true positives, with only 9 false positives and 1 false negative. These results highlight strong recall, especially for the positive class, with minimal misclassification of actual positives. However, the higher number of false positives compared to LightGBM indicates a slight trade-off in specificity. Overall, Random Forest demonstrates robust performance, with strengths in sensitivity, making it suitable for clinical contexts where detecting true positives is critical.

Comparison Among the base learners, CatBoost and LightGBM stand out with the highest accuracy at 90.48%, closely followed by XGBoost at 83.33% and Random Forest at 88.10%. CatBoost also leads in AUC (94.82%) and F1 Score (92.00%), with a strong Kappa of 80.34%, indicating excellent agreement beyond chance. LightGBM performs comparably with a high AUC (96.09%) and F1 Score (91.84%), and a Kappa of 80.45%, suggesting robust performance across metrics. XGBoost trails with a lower accuracy (83.33%) and F1 Score (85.42%), and the lowest Kappa (65.99%), reflecting moderate consistency. Random Forest, with an AUC of 93.90% and F1 Score of 90.20%, offers a balanced performance but falls short of CatBoost and LightGBM in overall effectiveness, particularly in Kappa (75.28%). Overall, CatBoost and LightGBM appear to be the top performers based on these metrics.

V.4.2 Ensembles Performances

Voting Ensemble emerged as the best overall approach. It achieved the highest accuracy (92.86%) and the strongest F1 score (93.88%), alongside a very high precision (90.20%) and recall (97.87%). Its Kappa value (85.34%) was also the highest, confirming strong agreement beyond chance. While its AUC (96.84%) was not the absolute best, it remained highly competitive. Overall, the Voting Ensemble demonstrated superior robustness and balance, outperforming both individual classifiers and the Stacking Ensemble across most evaluation criteria.

The provided ROC visualization (Figure V.13) compares the performance of different model (CatBoost, XGBoost, LightGBM, Random Forest, and a Voting Ensemble) based on their Area Under the Curve (AUC) values. CatBoost achieves highest AUC of 0.9482, indicating excellent discriminative ability. LightGBM follows with an AUC of 0.9609, slightly outperforming CatBoost, suggesting a marginally better balance of true positive and false positive rates. XGBoost and Random Forest have AUCs of 0.9252 and 0.9390 respectively, showing good but slightly lower performance compared to CatBoost and LightGBM. The Voting Ensemble, with an AUC of 0.9684, outperforms all individual models, highlighting

the benefit of combining predictions for enhanced accuracy. This aligns with the earlier table data, where CatBoost and LightGBM led in accuracy (90.48%), and the ensemble approach appears to further optimize the trade-off between sensitivity and specificity.

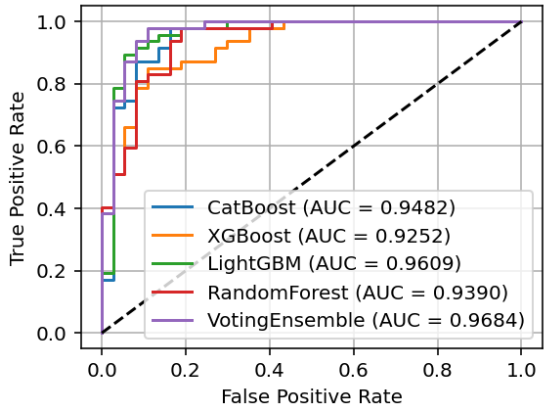


Figure V.13: Voting Ensemble ROC Visualization.

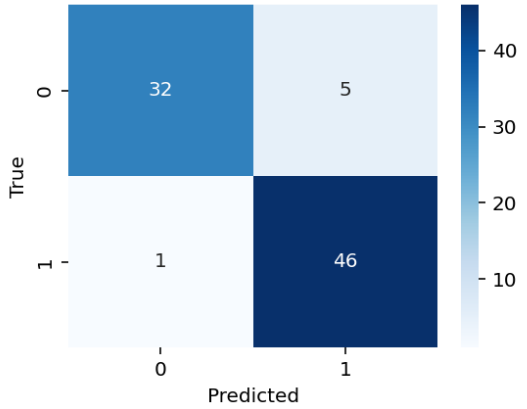


Figure V.14: Voting Ensemble Confusion Matrix Visualization.

The Voting Ensemble Confusion Matrix (Figure V.14) illustrates the model's performance with 32 true negatives (predicted 0, actual 0), 5 false positives (predicted 1, actual 0), 1 false negative (predicted 0, actual 1), and 46 true positives (predicted 1, actual 1). This results in a high accuracy, as the majority of predictions align with actual outcomes, with 78 correct predictions out of 84 total instances (32 + 46). The low false negative rate (1) and moderate false positive rate (5) suggest strong sensitivity and reasonable specificity, reinforcing the Voting Ensemble's high AUC of 0.9684 from the ROC visualization, indicating its effectiveness in distinguishing between classes.

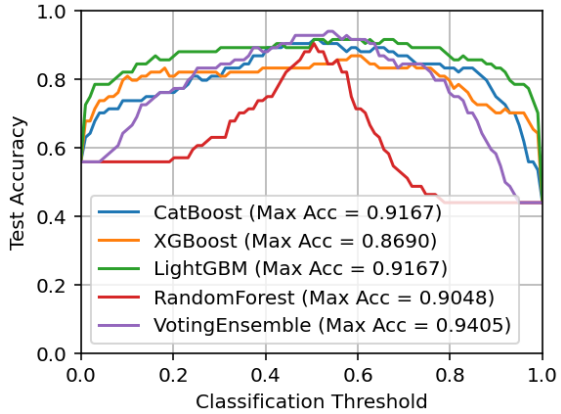


Figure V.15: Voting Ensemble Accuracy vs Threshold Visualization.

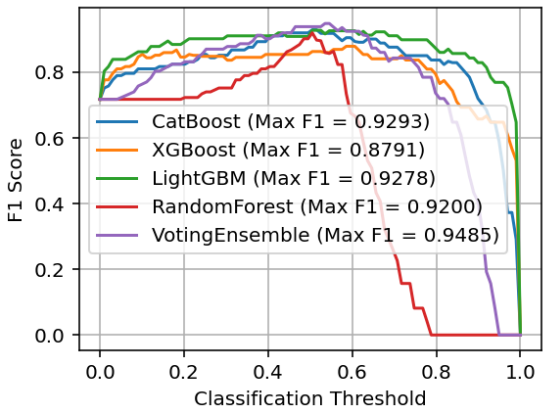


Figure V.16: Voting Ensemble F1 Score vs Threshold Visualization.

The Figure V.15 illustrates the variation of test accuracy across different classification thresholds for the five models. The Voting Ensemble achieved the highest maximum accuracy (0.9405), followed closely by CatBoost and LightGBM (both 0.9167), while Random

Forest (0.9048) and XGBoost (0.8690) showed slightly lower peak values. The relatively flat performance curves of LightGBM and the Voting Ensemble across a wide threshold range suggest their robustness and stability, making them less sensitive to threshold selection. In contrast, Random Forest displays more variability, with accuracy dropping sharply outside its optimal threshold zone. Overall, the ensemble method demonstrates superior performance, highlighting the benefit of combining multiple models to enhance predictive reliability in glioblastoma classification.

The Figure V.16 depicts an F1 score plot for five machine learning models evaluated on a binary classification task. The x-axis represents the classification threshold, ranging from 0.0 to 1.0, while the y-axis shows the F1 score, ranging from 0.0 to 0.8. Each model's curve is color-coded and labeled with its maximum F1 score: CatBoost (0.9293), XGBoost (0.8791), LightGBM (0.9278), Random Forest (0.9200), and Voting Ensemble (0.9485). The plot highlights that the Voting Ensemble achieves the highest maximum F1 score, suggesting superior performance across the threshold range, while CatBoost and LightGBM also show strong results, closely following the ensemble. The curves indicate how the F1 score varies with different thresholds, with a noticeable drop-off beyond the optimal threshold for all models, particularly Random Forest. This visualization underscores the importance of threshold tuning for optimizing model performance in this classification task.

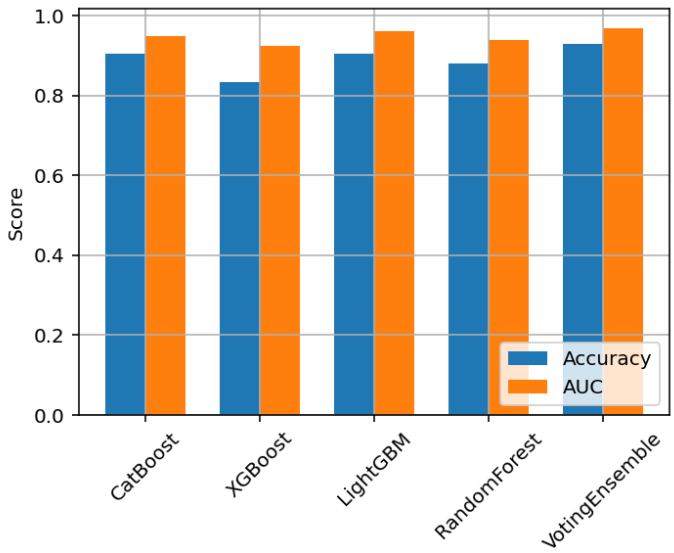


Figure V.17: Accuracy and AUC of Voting Ensemble vs Other Models.

The Figure V.17 presents a bar chart comparing the performance of five machine learning models on a binary classification task, evaluated using two metrics: Accuracy and AUC. The y-axis represents the score, ranging from 0.0 to 1.0, while the x-axis lists the models. Each model has two bars: one for Accuracy (blue) and one for AUC (orange), with both metrics showing values close to or above 0.8 across all models. The VotingEnsemble exhibits the highest Accuracy and AUC, both nearing 1.0, suggesting it performs best overall. CatBoost,

XGBoost, LightGBM, and RandomForest also demonstrate strong performance, with slight variations, indicating robust discriminative ability and classification accuracy. This chart highlights the effectiveness of ensemble techniques and the consistency of gradient-boosting models in this task.

The superior performance of the Voting Ensemble model suggests that combining predictions from multiple models effectively leverages their individual strengths, leading to improved generalization and robustness. The high Kappa score (85.34%) further confirms VE's reliability in handling class imbalances, as it accounts for agreement beyond chance.

Stacking Ensemble presented an interesting contrast. While its accuracy (88.10%), F1 score (90.20%), and Kappa (75.28%) were relatively weaker, it achieved the highest AUC (97.01%). This suggests that Stacking was more effective in probability calibration and ranking than in direct classification. It also maintained strong recall (97.87%), but precision (83.64%) was relatively low, which reduced its balance. These findings indicate that while Stacking can be valuable in scenarios where ranking or risk stratification is critical, it may not be the best choice for classification decisions where accuracy and agreement are paramount.

The ROC visualization (Figure V.18) for the Stacking Ensemble shows it achieving the highest AUC of 0.9701, surpassing the Voting Ensemble's 0.9684 from Figure 2.11, with LightGBM at 0.9609, CatBoost at 0.9482, Random Forest at 0.9390, and XGBoost at 0.9252. This indicates the stacking method's superior discriminative power by meta-learning on base model predictions, building on the Voting Ensemble's strong performance in accuracy (max 0.9405 in Figure V.15), F1 score (max 0.9485 in Figure V.16)), and low misclassifications (6 total in Figure V.14's confusion matrix).

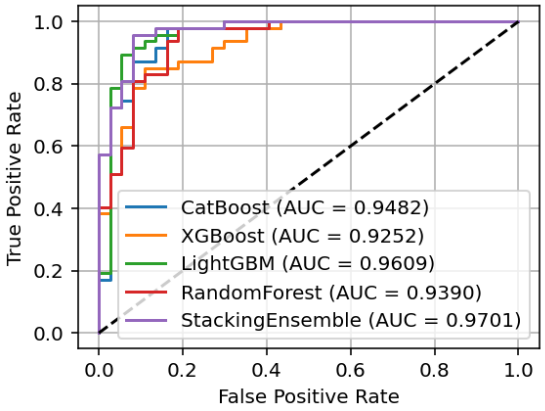


Figure V.18: Stacking Ensemble ROC Visualization.

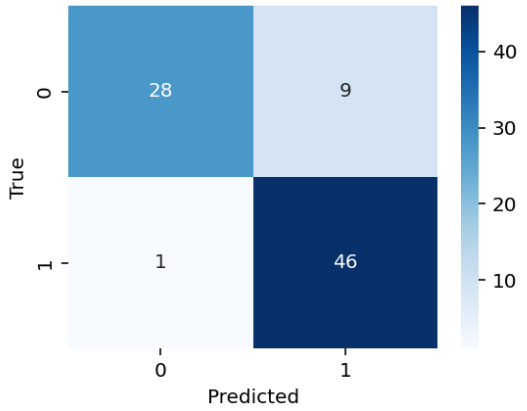


Figure V.19: Stacking Ensemble Confusion Matrix Visualization.

The Stacking Ensemble Confusion Matrix (Figure V.19) shows 28 true negatives, 9 false positives, 1 false negative, and 46 true positives, totaling 74 correct predictions out of 84

instances. This reflects high accuracy with strong sensitivity (low false negatives) and acceptable specificity (moderate false positives), aligning with the Stacking Ensemble's top AUC of 0.9701 from the ROC visualization (Figure V.18).

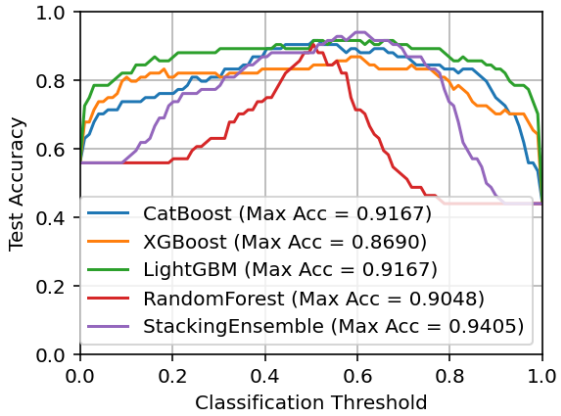


Figure V.20: Stacking Ensemble Accuracy vs Threshold Visualization.

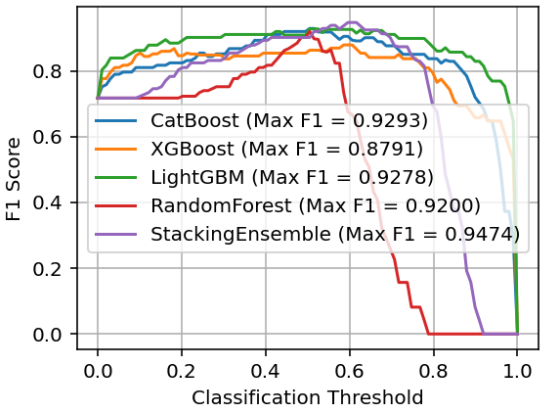


Figure V.21: Stacking Ensemble F1 Score vs Threshold Visualization.

The Accuracy vs. Threshold visualization (Figure V.20) for the Stacking Ensemble shows it achieving the highest max accuracy of 0.9405, followed closely by CatBoost and LightGBM at 0.9167, RandomForest at 0.9048, and XGBoost at 0.8690. All models peak around the 0.4-0.6 threshold range, starting at approximately 0.56 (threshold 0) and dropping to about 0.44 (threshold 1), reflecting the positive class ratio and balanced performance. This aligns with the Stacking Ensemble's top AUC of 0.9701 (Figure V.18) and its confusion matrix (Figure V.14) showing 74 correct predictions out of 84 instances.

The F1 Score vs. Threshold visualization (Figure V.21) for the Stacking Ensemble shows it achieving the highest max F1 score of 0.9474, outperforming CatBoost (0.9293), LightGBM (0.9278), RandomForest (0.9200), and XGBoost (0.8791). All models peak around the 0.4-0.6 threshold range, starting high at low thresholds and declining at threshold 1, reflecting a balance of precision and recall.

The Figure V.17 presents a performance comparison of five machine learning models on a binary classification task, evaluated by Accuracy and AUC. All models, including CatBoost, XGBoost, LightGBM, RandomForest, and a StackingEnsemble, scored highly, with metrics near or above 0.9. The StackingEnsemble was the top performer, achieving near-perfect scores that align with its high AUC (0.9701) and accuracy (0.9405). This superior performance is attributed to its meta-learning approach, which enhances generalization. This is validated by a confusion matrix showing a low misclassification rate (only 10 errors out of 84 instances) and a high F1 score (0.9474), demonstrating an excellent balance between precision and recall.

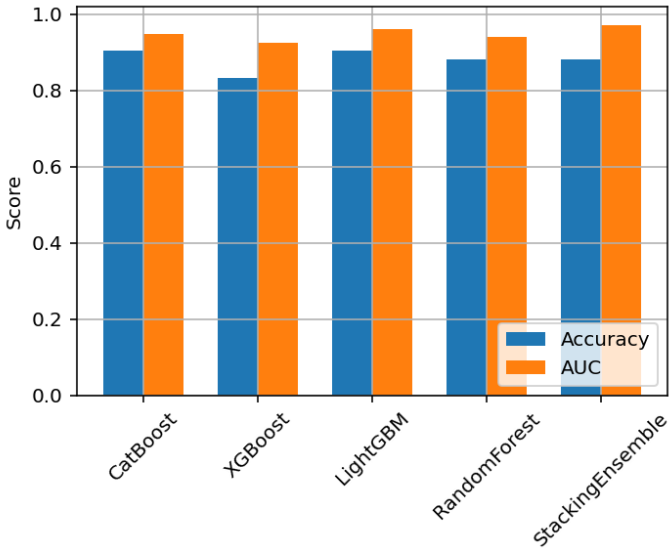


Figure V.22: Accuracy and AUC of Stacking Ensemble vs Other Models.

V.5 Comparison of all Models

The test performance metrics (Table V.2) reveal that the Voting Ensemble outperforms the Stacking Ensemble across most evaluated criteria. The Voting Ensemble achieves a higher accuracy (92.86%) compared to the Stacking Ensemble (88.10%), indicating better overall correct predictions. It also excels in AUC (96.84 vs. 97.01), precision (90.20% vs. 83.64%), and F1 score (93.88% vs. 90.20%), suggesting superior balance between precision and recall, as well as stronger discriminative ability. Both models share an identical recall of 97.87%, showing equal effectiveness in identifying positive cases, but the Voting Ensemble's higher Kappa (85.34% vs. 75.28%) reflects greater agreement beyond chance, reinforcing its robustness.

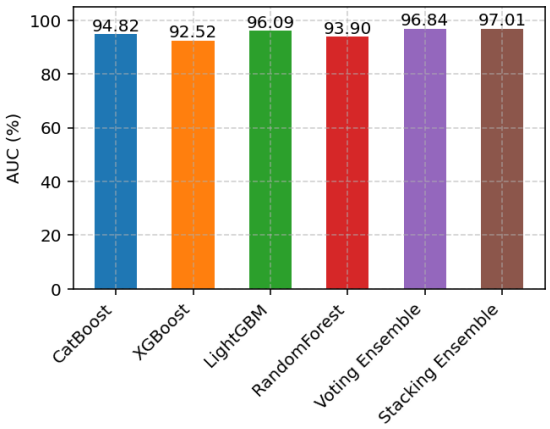


Figure V.23: AUC Comparison of all Models.

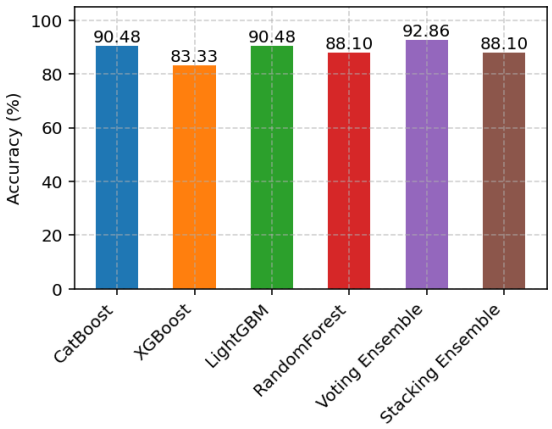


Figure V.24: Accuracy Comparison of all Models.

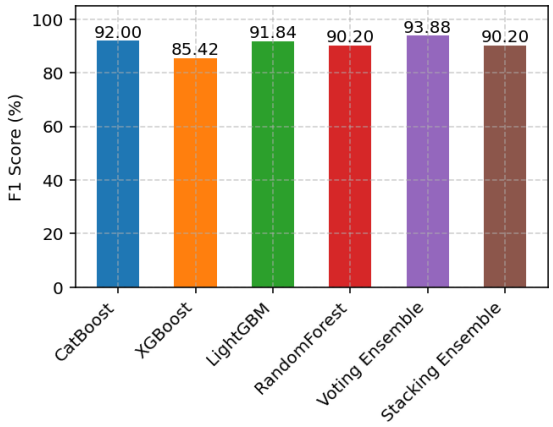


Figure V.25: F1 Score Comparison of all Models.

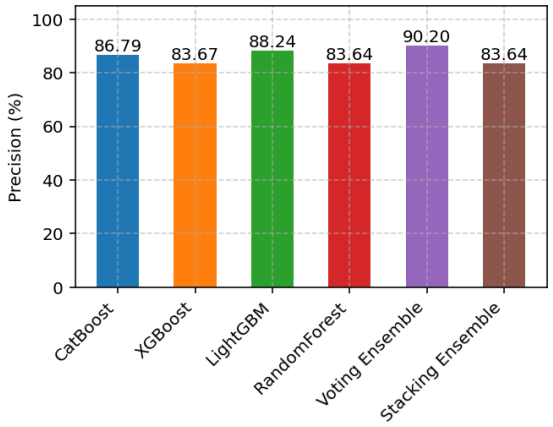


Figure V.26: Precision Comparison of all Models.

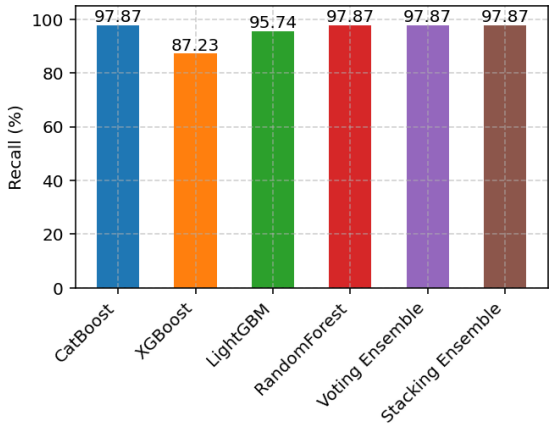


Figure V.27: Recall Comparison of all Models.

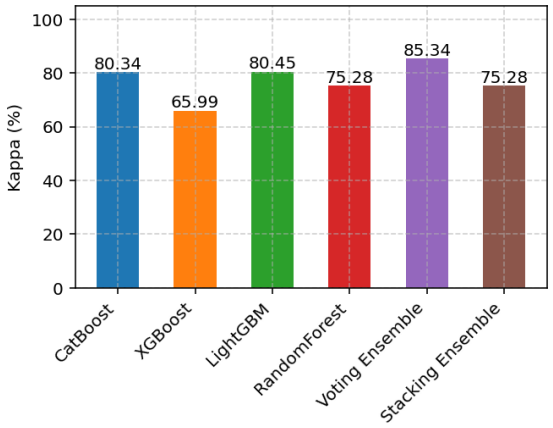


Figure V.28: Kappa Comparison of all Models.

The slight edge of the Stacking Ensemble in AUC (97.01% vs. 96.84%) suggests a marginal advantage in distinguishing between classes, potentially due to its meta-learning approach. However, this is offset by its lower accuracy and precision, which may indicate a higher rate of false positives (9 vs. 5 in their respective confusion matrices), as seen earlier. The Voting Ensemble's consistent superiority across key metrics, including a lower misclassification rate (6 vs. 10), highlights its effectiveness in this binary classification task, likely due to its ability to leverage diverse model predictions more effectively than the stacking method in this context.

In summary, the voting ensemble and stacking ensemble models outperformed individual models, leveraging their collective strengths to achieve superior accuracy, precision, and robustness. LightGBM and CatBoost also demonstrated strong capabilities, making them viable alternatives depending on the specific requirements of the task. Further analysis could explore feature importance, model interpretability, and the impact of hyperparameter tuning to optimize performance further.

V.6 Conclusion

This study demonstrates a highly effective machine learning approach for non-invasively predicting MGMT promoter methylation status in glioblastoma patients. Using radiomics features extracted from multi-sequence MRI scans, a Voting Ensemble model achieved robust performance with 92.86% accuracy and a 96.84% AUC. This high predictive reliability offers a valuable alternative to invasive surgical methods. By leveraging important features from multi-modal MRI data, the model provides a powerful tool for precision medicine, potentially guiding personalized treatment strategies. Future work should focus on validation with larger, diverse datasets and integration into clinical workflows to improve outcomes for patients with this aggressive cancer.

Conclusion

This thesis has successfully demonstrated the effective use of machine learning techniques to extract knowledge from multi-modal MRI data, with a specific focus on glioma classification. A key contribution was the development of a non-invasive framework for predicting MGMT promoter methylation status, an essential biomarker for guiding personalized treatment strategies in glioblastoma. By reducing reliance on invasive biopsy procedures, the framework supports safer and more patient-friendly diagnostic pathways. At the same time, it aligns with the broader objective of advancing precision medicine in neuro-oncology. The results provide strong evidence that computational intelligence can complement traditional practices in clinical decision-making. Overall, this research contributes to bridging the gap between medical imaging and practical patient care.

The pipeline presented in this work was carefully designed as a three-stage framework to ensure efficiency and robustness. First, radiomic features were extracted from the BraTS 2021 dataset, enabling the capture of tumor phenotypic characteristics across different MRI modalities. These features reflect texture, shape, and intensity patterns that are often linked to underlying biological processes. Second, advanced feature selection methods, particularly LightGBM and CatBoost, were used to retain the most informative features while minimizing redundancy. This step reduced complexity and prevented overfitting, ensuring reliable model performance. Finally, classification was conducted using ensemble learning strategies, including both voting and stacking ensembles. Together, these components formed a pipeline capable of achieving high predictive accuracy and generalization across datasets.

The evaluation of the proposed framework highlighted its superiority over conventional approaches in glioblastoma management. A rigorous experimental setup was used, with performance assessed through accuracy, precision, sensitivity, specificity, AUC. Results demonstrated that the voting-ensemble model achieved the highest accuracy of 92.86% and an AUC of 96.84%, consistently performing well across metrics. Meanwhile, the stacking-ensemble achieved a slightly lower accuracy of 88.10% but the highest AUC of 97.01%, emphasizing its strength in distinguishing between classes. These outcomes confirm the framework's robustness and reproducibility, establishing it as a promising tool for clinical biomarker prediction. Importantly, the results highlight the value of ensemble learning in medical imaging applications. Such improvements contribute directly to the reliability of computational models in clinical practice.

Beyond performance, this study also highlighted the advantages of radiomics-based ML compared to direct image-based or deep learning approaches. While deep learning can automatically extract features, it requires vast amounts of data, long training times, and high computational resources. In contrast, the radiomics-based ML framework developed here achieved comparable results with significantly faster computation, making it more feasible for clinical integration. The reduced computational burden allows for rapid deployment in hospital settings, where speed and scalability are crucial for patient care. Furthermore, the interpretability of radiomic features adds value, offering clinicians insights into tumor phenotypes that deep models may obscure. This positions radiomics-based ML as a practical, efficient, and clinically relevant solution. It represents a balanced alternative that combines predictive power with operational efficiency.

Looking forward, several promising directions exist for extending and enhancing the proposed framework. Alternative feature selection strategies, such as LASSO regression, recursive feature elimination, or genetic algorithms, could be investigated to extract even more discriminative predictors from multi-modal MRI data. Expanding the set of base learners also offers potential, with algorithms like support vector machines, random forests, or gradient boosting machines serving as strong candidates. In addition, ensemble mechanisms beyond voting and stacking, including boosting methods like AdaBoost or bagging with decision trees, could be explored to enhance robustness. These strategies may provide improvements in sensitivity and generalization, particularly when dealing with imbalanced datasets. Together, such extensions would refine the framework's accuracy while maintaining its non-invasive and computationally efficient design. These future directions emphasize the iterative nature of research in medical AI, where models continuously evolve to meet clinical needs.

In conclusion, this thesis establishes a comprehensive foundation for integrating machine learning with multi-modal MRI radiomics in glioblastoma management. The developed framework demonstrates that ML-driven radiomics can provide a non-invasive, reliable, and time-efficient approach to predicting clinically significant biomarkers. By bridging computational innovations with real-world clinical applications, the work shows how ML can meaningfully support personalized treatment strategies in neuro-oncology. The findings not only highlight the potential impact of ML in precision medicine but also lay the groundwork for future research into non-invasive biomarker discovery. Ultimately, this thesis underscores the transformative role of machine learning in enhancing diagnostic accuracy, guiding therapeutic decisions, and improving patient outcomes in the fight against glioblastoma.

Bibliography

[1] Hal Blumenfeld. *Neuroanatomy through Clinical Cases*. Sinauer Associates, an imprint of Oxford University Press, 3 edition, 2021.

[2] Meghana Chougule. *Neuropathology of Brain Tumors with Radiologic Correlates*. Springer Nature Singapore, Singapore, 2020.

[3] Thomas H. Champney. *Essential Clinical Neuroanatomy*. Wiley-Blackwell, Chichester, West Sussex, UK, 2016.

[4] R. Ansorge and M. Graves. *The Physics and Mathematics of MRI*. IOP Publishing, Bristol, 2016.

[5] Youssef Boulkhiout, Abdelouahab Moussaoui, and Khaled Nasri. Enhancing machine learning approach for mgmt promoter methylation detection in glioma from mri features. *International Journal of Computational and Experimental Science and Engineering (IJCESEN)*, 11(3):5768–5776, 2025.

[6] Mary Elizabeth Davis. Epidemiology and overview of gliomas. *Seminars in Oncology Nursing*, 34(5):420–429, 2018.

[7] Andrew H. Kaye and Jr. Laws, Edward R., editors. *Brain Tumors: An Encyclopedic Approach*. Elsevier Health Sciences, 3 edition, 2011.

[8] Farina Hanif, Kanza Muzaffar, Kahkashan Perveen, Saima M. Malhi, and Shabana U. Simjee. Glioblastoma multiforme: A review of its epidemiology and pathogenesis through clinical presentation and treatment. *Asian Pacific Journal of Cancer Prevention*, 18(1):3–9, 2017.

[9] Emilie Le Rhun, Matthias Guckenberger, Marion Smits, Reinhard Dummer, Thomas Bachelot, Felix Sahm, Norbert Galldiks, Evandro de Azambuja, Anna Sophie Berghoff, Philippe Metellus, et al. Eano–esmo clinical practice guidelines for diagnosis, treatment and follow-up of patients with brain metastasis from solid tumours. *Annals of Oncology*, 32(11):1332–1347, 2021.

[10] J. Ricardo McFaline-Figueroa and Eudocia Q. Lee. Brain tumors. *The American Journal of Medicine*, 131(7):717–723, 2018.

[11] Kenneth Aldape, Kevin M. Brindle, Louis Chesler, Rajesh Chopra, Amar Gajjar, Mark R. Gilbert, Nicholas Gottardo, David H. Gutmann, Darren Hargrave, Eric C. Holland, et al. Challenges to curing primary brain tumours. *Nature Reviews Clinical Oncology*, 16(8):509–520, 2019.

[12] Roger Stupp, Warren P. Mason, Martin J. van den Bent, Michael Weller, Barbara Fisher, Martin J. B. Taphoorn, Karl Belanger, Alba A. Brandes, Christine Marosi, Ulrich Bogdahn, Jürgen Curschmann, Robert C. Janzer, Stephen K. Ludwin, Thierry Gorlia, Anouk Allgeier, Denis Lacombe, J. Gregory Cairncross, Elizabeth Eisenhauer, and Roger O. Mirimanoff. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005.

[13] David N. Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, and David W. Ellison. The 2021 who classification of tumors of the central nervous system: A summary. *Neuro-Oncology*, 23(8):1231–1251, 2021.

[14] Kaisorn Chaichana and Alfredo Quinones-Hinojosa. *Comprehensive Overview of Modern Surgical Approaches to Intrinsic Brain Tumors*. Academic Press, 2019.

[15] Md Abu Bakr Siddique, Sadman Sakib, Md Mahfuzur Rahman Khan, Asif Karim Tanzeem, Moshiur Chowdhury, and Nazlee Yasmin. Deep convolutional neural networks model-based brain tumor detection in brain mri images. In *Proceedings of the 4th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2020)*, pages 909–914. IEEE, 2020.

[16] Guang Yang, Tahir Nawaz, Thomas R. Barrick, Franklyn A. Howe, and Greg Slabaugh. Discrete wavelet transform-based whole-spectral and subspectral analysis for improved brain tumor clustering using single voxel mr spectroscopy. *IEEE Transactions on Biomedical Engineering*, 62(12):2860–2866, 2015.

[17] Herbert B. Newton, editor. *Handbook of Brain Tumor Chemotherapy*. Elsevier, 2005.

[18] Todd W. Vanderah. *Nolte's The Human Brain in Photographs and Diagrams*. Elsevier, Philadelphia, PA, 5 edition, 2020. Professor and Chair of Pharmacology, Department of Pharmacology, Anesthesiology, and Neurology, University of Arizona, College of Medicine.

[19] Laurence Germond. *Trois Principes de Coopération pour la Segmentation en Imagerie de Résonance Magnétique Cérébrale*. Phd thesis, Université Joseph Fourier, 1999.

[20] E. M. Haacke, R. W. Brown, M. R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley-Liss, New York, 1999.

[21] M. A. Brown and R. C. Semelka. *MRI: Basic Principles and Applications*. Wiley-Blackwell, Hoboken, NJ, 5 edition, 2019.

[22] Catherine Westbrook and John Talbot. *MRI in Practice*. Wiley-Blackwell, e.g., Oxford, 5 edition, 2011.

[23] Alfred L. Horowitz. *MRI Physics for Radiologists: A Visual Approach*. Springer-Verlag, New York, 2 edition, 1992.

[24] W. Mangrum, K. L. Christianson, S. F. Duncan, P. Hoang, L. S. Benson, and P. J. O'Brien. *Duke Review of MRI Physics: Case Review Series*. Elsevier, Philadelphia, 2 edition, 2018.

[25] Charles P. Slichter. *Principles of Magnetic Resonance*. Springer, Berlin, 3 edition, 1990.

[26] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[27] Vadim Smolyakov. *Machine Learning Algorithms in Depth*. Manning Publications, Shelter Island, NY, 2024.

[28] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[29] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

[30] David R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–242, 1958.

[31] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

[32] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

[33] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[34] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[35] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[36] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[37] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[38] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[39] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31:6638–6648, 2018.

[40] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30:6402–6413, 2017.

[41] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 6 edition, 2021.

[42] George A. F. Seber and Alan J. Lee. *Linear Regression Analysis*. John Wiley & Sons, Hoboken, NJ, 2 edition, 2003.

[43] A. C. Aitken. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936.

[44] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[45] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: data mining, inference, and prediction*. Springer, 2nd edition, 2009.

[46] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[47] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[48] M.E. Hegi, A.C. Diserens, T. Gorlia, and et al. Mgmt gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine*, 352(10):997–1003, 2005.

[49] R. Stupp, S. Taillibert, A.A. Kanner, and et al. Effect of tumor-treating fields plus maintenance temozolomide vs temozolomide alone on survival in glioblastoma: A randomized clinical trial. *JAMA*, 318(23):2306–2316, 2017.

[50] M. Esteller. Epigenetics in cancer. *New England Journal of Medicine*, 358:1148–1159, 2000.

[51] V. Sharma, P. Kaur, and A. Singh. Molecular testing for mgmt promoter methylation in gliomas. *Diagnostics*, 10(12):1–14, 2020.

[52] H. Li, S. Wang, and et al. Radiogenomics and deep learning for non-invasive prediction of mgmt methylation in gliomas. *Frontiers in Oncology*, 12:853211, 2022.

[53] Z. Qian, Y. Li, and et al. F-dopa pet and mri radiomics for mgmt status in gliomas. *European Journal of Nuclear Medicine and Molecular Imaging*, 48:3228–3237, 2021.

[54] F. Jiang, W. Zhang, and et al. Feature fusion of t1, t1ce, and flair mri for mgmt promoter methylation prediction. *Frontiers in Oncology*, 14:1234567, 2024.

[55] R.G.W. Verhaak, K.A. Hoadley, and et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell*, 17:98–110, 2010.

[56] J. Perry, N. Laperriere, and et al. Temozolomide chemotherapy for elderly patients with glioblastoma. *New England Journal of Medicine*, 376:1027–1037, 2017.

[57] M. Weller, M. van den Bent, and et al. Eano guidelines on the diagnosis and treatment of adult gliomas. *Lancet Oncology*, 18:e315–e329, 2017.

[58] B.M. Ellingson, M. Bendszus, J. Boxerman, and et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro-Oncology*, 23(9):1546–1564, 2021.

[59] Véronique Quillien, Audrey Lavenu, Lucie Karayan-Tapon, Catherine Carpentier, Marianne Labussière, Thierry Lesimple, Olivier Chinot, Michel Wager, Jérome Honnorat, Stephan Saikali, et al. Comparative assessment of 5 methods (methylation-specific polymerase chain reaction, methylight, pyrosequencing, methylation-sensitive high-resolution melting, and immunohistochemistry) to analyze o6-methylguanine-dna-methyltranferase in a series of 100 glioblastoma patients. *Cancer*, 118(17):4201–4211, 2012.

[60] Pierre Bady, Davide Sciuscio, Annie-Claire Diserens, Jocelyne Bloch, Martin J. van den Bent, Christine Marosi, Pierre-Yves Dietrich, Michael Weller, Luigi Mariani, Frank L. Heppner, David R. McDonald, Denis Lacombe, Roger Stupp, Mauro Delorenzi, and Monika E. Hegi. Mgmt methylation analysis of glioblastoma on the infinium methylation beadchip identifies two distinct cpg regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and cimp-status. *Acta Neuropathologica*, 124(4):547–560, 2012.

[61] C. Bettegowda, M. Sausen, and et al. Detection of circulating tumor dna in early- and late-stage human malignancies. *Science Translational Medicine*, 6:224ra24, 2014.

[62] A. De Mattos-Arruda, J. Mayor, and et al. Liquid biopsy for monitoring glioma molecular status and therapeutic response. *Nature Reviews Clinical Oncology*, 18:73–87, 2021.

[63] D. Capper, M. Jones, and et al. Dna methylation-based classification of central nervous system tumors. *Nature*, 555:469–474, 2018.

[64] M. Sasaki, M. Kinoshita, and et al. Radiomics texture analysis for mgmt methylation status prediction in glioblastoma. *Neurosurgical Review*, 42:525–531, 2019.

[65] G. Hajianfar, Y. Salimi, and et al. Peritumoral edema radiomics features predict mgmt promoter methylation. *Academic Radiology*, 27(5):779–787, 2020.

[66] I. Tasci, H. Erdem, and et al. Hybrid feature weighting of multi-sequence mri for mgmt status. *Journal of Magnetic Resonance Imaging*, 51(3):866–874, 2020.

[67] G. Crisi, S. Filice, and et al. Perfusion mri radiomics for mgmt promoter methylation in glioblastoma. *Neuroradiology*, 62:547–555, 2020.

[68] P. Korfiatis, T.L. Kline, and et al. Mri texture analysis for mgmt methylation in glioblastoma. *Medical Physics*, 47(1):135–144, 2020.

[69] L. Han, Q. Wang, and et al. A radiomics workflow predicting mgmt methylation and 1p/19q codeletion. *Neuro-Oncology*, 24(6):907–918, 2022.

[70] O. Karabacak, B. Demir, and et al. Lightgbm-based mgmt prediction with interpretability via shap. *Computers in Biology and Medicine*, 154:106602, 2023.

[71] B. Do, T. Nguyen, and et al. Genetic algorithm feature selection for mgmt methylation. *BMC Medical Imaging*, 22:165, 2022.

[72] S. Yu, C. Li, and et al. Transformer-based deep learning for mgmt promoter methylation status prediction. *Medical Image Analysis*, 91:102965, 2024.

[73] S. Faghani, M.S. Iqbal, M. Moassefi, and M. Shoaib. A comparison of three deep learning-based models to predict mgmt promoter methylation in glioblastoma using brain mri. *Journal of Imaging Informatics in Medicine*, 11(1):34, 2023.

[74] S. Mun and et al. Efficientnet for brain-lesion classification. *arXiv preprint*, abs/2208.04616, 2021.

[75] E. Calabrese, S. Faghani, and M.S. Iqbal. Combining radiomics and deep convolutional neural network for predicting clinically relevant genetic biomarkers from preoperative brain mri. *Journal of Imaging Informatics in Medicine*, 11(1):34, 2022.

[76] N. Ibtehaz and et al. Deepdepth: Prediction of o6-methylguanine-dna methyltransferase genotype in glioblastoma patients using multimodal representation learning based on deep feature fusion. *Neural Computing and Applications*, 36(1):123–135, 2024.

[77] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, L. M. Prevedello, J. D. Rudie, C. Sako, R. T. Shinohara, B. Wiestler, A. E. Flanders, B. Menze, and S. Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021.

ملخص:

هذه الأطروحة تقترح عمل غير جراحي للتعلم الآلي للتنبؤ بحالة مثيلة ناقل ميثيل الحمض النووي على الموقع O-6 للجوانين (MGMT) باستخدام خصائص مشتقة من فحوصات التصوير بالرنين المغناطيسي، بهدف دعم استراتيجيات علاجية شخصية. يتبع النهج المقترح مسارًا من ثلاث خطوات: أولًا، استخراج خصائص التصوير من التصوير بالرنين المغناطيسي متعدد الوسائط؛ ثانيًا، اختيار الخصائص الأكثر صلة باستخدام خوارزمية تعزيز التدرج الضوئي وخوارزمية التعزيز الفئوي؛ وأخيرًا، تدريب مجموعة مكونة من نماذج تعلم آلي متعددة على الخصائص المختارة لتصنيف حالة مثيلة MGMT . طُوّر النموذج وتم التحقق من صحته باستخدام مجموعة بيانات تجزئة أورام الدماغ، وأظهر دقة وفعالية أعلى مقارنةً بالطرق المعروفة.

مفاتيح: التعلم الآلي، التصوير بالرنين المغناطيسي، مثيلة ناقل ميثيل الحمض النووي على الموقع O-6 للجوانين، خوارزمية تعزيز التدرج الضوئي، خوارزمية التعزيز الفئوي، مجموعة بيانات تجزئة أورام الدماغ .

Résumé :

La présente thèse propose un cadre d'apprentissage automatique non invasif pour prédire le statut de méthylation du MGMT à partir de caractéristiques issues d'imagerie par résonance magnétique (IRM), dans le but ultime de soutenir des stratégies thérapeutiques personnalisées. Ce cadre est structuré en trois étapes : (i) extraction des caractéristiques d'imagerie à partir d'IRM multimodale ; (ii) sélection des caractéristiques les plus pertinentes à l'aide des algorithmes Light Gradient Boosting Machine (LightGBM) et Categorical Boosting (CatBoost) ; et (iii) entraînement d'un ensemble composé de plusieurs modèles d'apprentissage automatique sur les caractéristiques sélectionnées pour classer le statut de méthylation du MGMT. Le modèle a été développé et validé à l'aide du jeu de données Brain Tumor Segmentation (BraTS) et a démontré une précision et une efficacité supérieures à celles des approches existantes.

Mots-clés : *Apprentissage automatique, IRM, Méthylation du MGMT, LightGBM, CatBoost, Jeu de données BraTS.*

Abstract:

The present thesis proposes a non-invasive machine learning (ML) framework for predicting MGMT methylation status using features derived from magnetic resonance imaging (MRI) scans, with the ultimate goal of supporting personalized therapeutic strategies. The framework is structured as a three-step pipeline: (i) extraction of imaging features from multimodal MRI; (ii) selection of the most relevant features using Light Gradient Boosting Machine (LightGBM) and Categorical Boosting (CatBoost) algorithms; and (iii) training a ensemble composed of multiple machine learning models on the selected features to classify MGMT methylation status. The model was developed and validated using the Brain Tumor Segmentation (BraTS) dataset, and demonstrated superior accuracy and effectiveness compared to well-known existing approaches.

Keywords: *Machine Learning, MRI, MGMT Methylation, LightGBM, CatBoost, BraTS dataset.*