POLITECNICO DI TORINO

**Master's Degree Course in**
**Data Science and Engineering**

Master's Degree Thesis

# 3D Multi-Input Deep Learning for Brain Lesion Classification: Attention-Based Analysis of Stable vs. Recurrent Lesions


**Supervisors**
Prof. Santa DI CATALDO
Prof. Francesco PONZIO

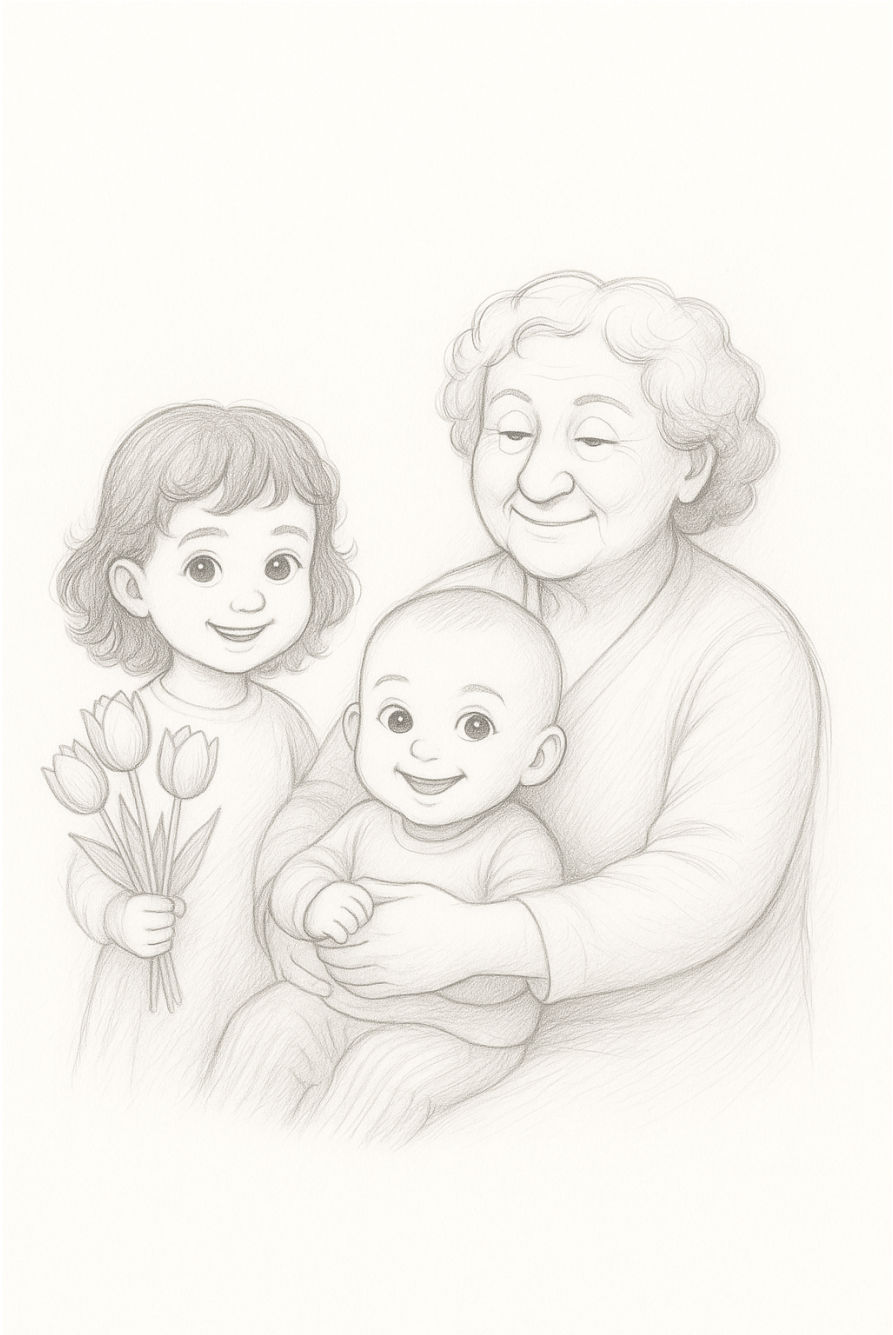
**Candidate**
Kuerxi GULISIDAN

. . . . . .

. . . . . .

Academic Year 2024-2025

*To my Grandmother Kembernisa. I hope I've made you proud.*



**Abstract**

Brain metastases are a common and serious complication among cancer patients and are often treated by stereotactic radiosurgery, such as Gamma Knife therapy.
While this treatment does an excellent job of controlling the local tumor, it is necessary to discriminate between stable and recurrent disease.

Existing evaluation strategies are largely based on expert visual interpretation of serial MRI and Radiotherapy-Planning Images; however, this approach requires intensive manual handling of Radiotherapy Dose data and, as such, **introduces variability across operators** and observers.

This thesis aims to evaluate the current literature on the topic and introduces an effective deep-learning architecture designed to classify brain lesions as stable or recurrent after Gamma Knife radiosurgery.
The resulting model integrates multimodal information such as Magnetic Resonance Imaging (MRI), Radiotherapy Dose distributions (RTDose), and highly structured clinical parameters into a single multi-input neural network architecture.
Particular consideration is given to handling the large class imbalance in recurrence prediction through the application of selective augmentation and balanced sampling schemes to enhance learning efficiency.

Extensive experimentation and validation demonstrate meaningful improvements over existing baselines, with greater robustness across patients and different data.
Compared to previously established metrics (10% recall, 18.2% F1 score), the model achieved a significant improvement with a recall of 50% and an F1 score of 28.6%.
This sensitivity improvement is clinically significant and may help avoid critical interventions for recurrent cases from being delayed.

The proposed framework contributes meaningfully to the evolving field of automated neuro-oncology, laying the foundation for consistent, data-driven monitoring of patients undergoing radiosurgical treatment for brain metastases.
Although the results and findings are encouraging, demonstrating the potential of combining deep learning techniques, multimodal imaging data, and structured clinical information, they also indicate that this fundamental and important topic requires further and greater focus.

**Acknowledgements**

Desidero esprimere il mio più sentito ringraziamento alla **Prof.ssa Santa Di Cataldo** e al **Prof. Francesco Ponzio** per la loro preziosa guida e dedizione durante l'intero percorso della mia tesi. Sono particolarmente grata al Prof. Francesco Ponzio per il suo costante sostegno, incoraggiamento e per gli approfonditi feedback che hanno arricchito lo sviluppo di questo lavoro.

My academic journey in Italy began when I started my Bachelor's in Electronic and Communication Engineering. At first, I didn't speak any Italian, so everything felt new and difficult. It took time and hard work to learn the language and get used to a different way of studying and living. Slowly, I found my rhythm and learned to face the challenges at university.

After I finished my Bachelor's, I went on to a Master's program in Data Science and Engineering. Thanks to the skills I had already built, that transition was smoother and less stressful. I'm proud of how far I've come and that I never gave up, even when things were tough.

Completing this journey means so much to me, and I couldn't have done it without the support of many people. I want to thank them all:

**Bahargül,** 我的妈妈，我想对您说一声最真心的感谢。您一个人把我和姐姐拉扯大，从来不逼我们做自己不想做的事，也不拿我们和别人比。您总是默默支持我们的选择，给我们自由去追逐自己的梦想，哪里做得不好也从不责怪，只会给我们鼓励和温暖。能成为您的女儿，是我这辈子最幸运的事。妈妈，谢谢您这些年为我们操心、付出和辛苦，我爱您！ (Bahargül, My mother, I want to express my most sincere gratitude to you. You raised my sister and me on your own, never forcing us to do anything we didn't want to, and never comparing us to anyone else. You have always quietly supported our choices, giving us the freedom to chase our own dreams. Whenever we struggled, you never blamed us, instead you offered encouragement and warmth. Being your daughter is the greatest blessing of my life. Mom, thank you for all your care, sacrifice, and hard work over the years. I love you!)

I want to thank my sister, **Güzel,** who has always been like a twin to me despite our four-year age gap. She arrived in this unfamiliar country before I did and walked through challenges she never should have faced. Because of everything she went through, my own path was so much smoother. Having her by my side in Italy eased my homesickness and turned this foreign place into a second home. She supported me in practical ways, she was my emotional rock, always cheering me on and helping me believe in myself. I love her more than words can say, and no matter where life leads us, I hope we'll always be side by side.

I would also like to thank my brother-in-law **Gaetano,** who has truly become like a brother to me. He is an incredibly kind Italian guy who always lends a hand to others and firmly believes there are more good people than bad in the world. His support, both academically and personally has meant so much to me. After completing his bachelor's degree, he spent several years working before returning to pursue his master's, and that's

how we ended up as classmates. We studied side by side, tackled problem sets together, and collaborated on projects, often fighting a lot over them. Gaetano often jokes that I'll forget him and my sister, but I know how vital they are in my life, NO ONE could ever take their place.

Vorrei anche esprimere la mia sincera gratitudine alla famiglia di Gaetano: la mamma **Marinella**, il papà **Bruno**, la sorella **Laura** e **Giuseppe**. Mi hanno accolta come parte della loro famiglia, facendomi sentire a casa qui in Italia. Questo senso di calore e appartenenza è stato particolarmente prezioso durante ogni festività e nel periodo della pandemia. Ringrazio in modo speciale il papà Bruno, che con le sue conversazioni quotidiane in italiano ha contribuito in modo significativo al mio miglioramento linguistico. Sono profondamente grata a tutti loro per l'affetto, la gentilezza e il sostegno che mi hanno sempre dimostrato.

Desidero inoltre ringraziare la nonna di Gaetano, nonna **Rita**, i suoi peperoncini sono i più buoni che abbia mai assaggiato, e da amante del piccante quei sapori mi hanno fatto sentire meno la nostalgia dei piatti di casa. Un grazie speciale va anche a **Rocco** e **Gloria**, insieme alle loro figlie **Aurora** e **Miriam**, li ho conosciuti quando erano ancora piccole, ma già riuscivamo a capirci perfettamente. Ho amato ogni momento trascorso nel **Salento**, dove mi hanno donato tanto calore.

Sono loro la mia famiglia qui in Italia.

I would also like to extend my deepest gratitude to all the **friends** I've been lucky enough to meet on this journey, from those early days of my Bachelor's studies through the challenges and triumphs of my Master's program. You have each left an indelible mark on my heart.

To the friends who borrow me their lecture notes or summary after each lessions, thank you for believing in me even when I doubted myself. To those late-night study sessions in Study room of Michelangelo where we fueled our brains on coffee. you reminded me that I was never alone in pushing toward our goals. When group projects felt impossible, you stayed by my side, patiently working through every glitch, every bug, every "we'll never finish in time" panic, until we crossed the finish line together and telling me I am the Number 1.

I'm deeply grateful for the laughter we shared in moments of pure joy, the gelato breaks, the picnics, the silly inside jokes that still make me smile on the toughest days. I also want to thank you for inviting me into your homes and sharing every meal with me. Each dinner, lunch, or simple snack you prepared gave me exactly the boost of warmth and encouragement I needed to carry on.

Because of you, the burdens felt lighter, the nights felt brighter, and the victories tasted sweeter. Thank you for walking beside me, cheering me on, and sharing in every step of this adventure. My life and my heart are richer for having each of you in it, and I will carry our memories and your unwavering support with me always.

**Un sincero grazie a tutti coloro che hanno fatto parte di questo percorso.**

# Contents

# List of Tables

# List of Figures

*The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery every day. Never lose a holy curiosity.*

[Albert Einstein]

# Acronyms

| | |
|---|---|
| **MRI** | Magnetic Resonance Imaging |
| **CT** | Computed Tomograph |
| **MRS** | Magnetic Resonance Spectroscopy |
| **DWI** | Diffusion-Weighted Imaging |
| **WBRT** | Whole Brain Radiotherapy |
| **SRS** | Stereotactic Radiosurgery |
| **CNN** | Convolutional Neural Network |
| **ViTs** | Vision Transformer |
| **3D CNN** | 3D Convolutional Neural Network |
| **RTDOSE** | Radiotherapy Dose Distribution |
| **ROI** | Region of Interes |
| **DICOM** | Digital Imaging and Communications in Medicine |
| **RTSTRUCT** | Radiotherapy Structure Set |
| **BBB** | Blood-Brain Barrier |
| **T1WI** | T1-Weighted Imagin |
| **T2WI** | T2-Weighted Imagin |
| **FLAIR** | Fluid-Attenuated Inversion Recovery |
| **PWI** | Perfusion-Weighted Imaging |
| **PET-CT** | Positron Emission Tomograph |
| **CBV** | cerebral blood volume |

# Chapter 1

# Introduction

## 1.1 Brain Metastases: Clinical Significance and Current Challenges

Brain metastases represent the most common intracranial malignancies in adults, occurring in approximately 20–40% of all cancer patients[5][6][7] and significantly surpassing primary brain tumors in prevalence. This high incidence makes brain metastases a critical challenge in modern oncology and neurology. Over the past twenty years, the reported incidence of brain metastases has risen significantly. This increase is chiefly due to two factors: more effective systemic cancer therapies that prolong patient survival thereby allowing micrometastases to develop and widespread adoption of high-resolution neuroimaging techniques that detect asymptomatic lesions much earlier. Current estimates suggest that the annual incidence of brain metastases in the United States ranges from 170,000 to 200,000 cases[6].

Brain metastases typically arise through hematogenous spread and are predominantly located at the gray–white matter junction, with the cerebral hemispheres being the most frequently affected areas[8]. The distribution and characteristics of these metastases largely depend on the primary tumor type. Lung cancer accounts for 40–50% of all brain metastases cases, particularly from small-cell and non-small-cell lung cancer[9]. Breast cancer contributes 15–25% of cases, with higher incidence rates observed in HER2-positive and triple-negative subtypes[9]. Melanoma, while representing 5–20% of cases, shows a strong propensity for hemorrhagic metastases[9], adding complexity to both diagnosis and treatment.

The clinical presentation of brain metastases varies significantly depending on lesion location, size, and related edema. Patients may have headaches, cognitive changes, seizures, focal neurological deficits, or remain asymptomatic until advanced stages. This differences in presentation, together with the rapid progression of metastatic disease, necessitates prompt and accurate diagnosis followed by appropriate therapeutic intervention.

| Primary Tumor Type | Incidence in Brain Metastases (%) | Common Sites of Metastases | Imaging Characteristics |
|---|---|---|---|
| Lung Cancer | 40-50% | Frontal, Temporal, Parietal | Multiple, Edematous |
| Breast Cancer | 15-25% | Occipital, Cerebellum | Ring-Enhancing Lesions |
| Melanoma | 5-10% | Hemorrhagic Sites | Hyperintense, Bleeding |

Table 1.1: Incidence of Brain Metastases by Primary Tumor Type

## 1.2 Stereotactic Radiosurgery and Gamma Knife Therapy

Stereotactic Radiosurgery (SRS) has become as a key treatment modality for patients with brain metastases, offering a highly targeted, non-invasive approach for delivering precise, high-dose radiation to tumor lesions[10]. Unlike conventional radiation therapy, which irradiates large areas of brain tissue over multiple sessions, SRS concentrates radiation beams to a specific target area in a single or limited number of sessions. This targeted approach minimizes damage to surrounding healthy brain tissue, making SRS particularly suitable for treating small to medium-sized lesions typically measuring less than 3 cm in diameter[11].

Gamma Knife surgery is one of the most advanced forms of SRS technology. The apparatus comprises 200 individual cobalt-60 sources in a hemispheric configuration each radiating the gamma rays converging at the desired lesion site[12]. The convergence achieves a steep dose gradient such that the tumor is given an ablative dose of radiation with sparing of the rest of the surrounding normal brain tissue. The precision of Gamma Knife therapy is because of a complex multi-step workflow wherein stereotactic frame placement is followed by the acquisition of high-resolution imaging along with computerized treatment planning and precise dose administration

### 1.2.1 Clinical Applications and Indications

Gamma Knife therapy has proven particularly effective for several clinical scenarios:

- **Oligometastatic Disease:** Patients with a limited number of metastatic lesions ($\leq$3–5) benefit significantly from SRS as it provides localized treatment with minimal impact on overall brain function[13].
- **Radioresistant Tumors:** Tumors such as melanoma and renal cell carcinoma, which are less responsive to conventional radiation, showes improved control rates with SRS due to its ability to deliver ablative doses of radiation.

- **Lesions in Eloquent Brain Areas:** For metastases located in regions critical for neurological function (e.g., motor cortex, speech centers), SRS offers a safer alternative to surgical resection, minimizing the risk of neurological deficits.
- **Recurrent or Residual Lesions:** Salvage therapy with SRS may be applied as a non-surgical option for patients with poor performance status in cases of previously treated lesions showing recurrence.

### 1.2.2 Technical Implementation

The high precision of Gamma Knife is achieved through a comprehensive workflow:

- **Patient Positioning and Immobilization:** A stereotactic frame is attached to the patient's skull to eliminate head movement and ensure accurate targeting.
- **Imaging and Treatment Planning:**
  - High-resolution MRI or CT imaging is performed to delineate lesion boundaries, enabling creation of three-dimensional treatment plans.
  - RTStruct data is used to outline tumor margins, while RTDose data quantifies the planned radiation dose distribution.
- **Radiation Delivery:** Several beams of gamma radiation are focused at the lesion site to provide high dose focal therapy with minimal exposure to surrounding tissues
- **Post-Treatment Monitoring:** Follow-up MRI scans are performed at regular intervals to assess treatment response and detect potential complications such as radiation necrosis.

## 1.3 Post-Treatment Monitoring: The Recurrence vs. Stability Challenge

While Gamma Knife radiosurgery provides effective local tumor control, it is associated with potential complications, most notably radiation necrosis. Radiation necrosis is a delayed adverse effect characterized by cell death and inflammation within the irradiated area. This complication can occur months to years post-treatment and may present clinically with new or worsening neurological deficits including headache, cognitive decline, seizures, or focal weakness.

On MRI, radiation necrosis typically presents as a contrast-enhancing lesion with surrounding vasogenic edema, mimicking the appearance of recurrent tumor. Distinguishing between tumor recurrence and radiation necrosis based solely on conventional imaging is challenging due to overlapping radiological features[14][15].

### 1.3.1 Diagnostic Complexity

The primary challenge in post-Gamma Knife monitoring lies in distinguishing between tumor recurrence and radiation necrosis, as both conditions can present with remarkably similar imaging characteristics on conventional MRI:

- **Ring Enhancement:** Both conditions may present as ring-enhancing lesions with central necrosis.
- **Edema and Mass Effect:** Perilesional edema is often observed in both radiation necrosis and recurrent tumor.
- **Hemorrhage:** Metastases from melanoma and renal cell carcinoma are especially susceptible to hemorrhagic transformation, making interpretation of post-SRS imaging difficult

| Imaging Modality | Recurrent Tumor | Radiation Necrosis |
|---|---|---|
| T1-Weighted MRI | Solid, Nodular Enhancement | Ring Enhancement, Central Necrosis |
| DWI | Restricted Diffusion | Facilitated Diffusion |
| PWI | Elevated CBV | Decreased CBV |
| MRS | Increased Choline | Elevated Lactate and Lipid Peaks |

Table 1.2: Imaging Features of Recurrent Tumor vs. Radiation Necrosis

### 1.3.2 Current Assessment Limitations

Current assessment protocols rely predominantly on expert interpretation of serial MRI and radiotherapy planning images. This approach, while clinically established, is both labor-intensive and subjective, leading to significant inter-observer variability. The subjective nature of image interpretation can result in diagnostic uncertainty, particularly in borderline cases where imaging features are ambiguous. This variability in assessment underscores the need for more objective, reproducible, and automated approaches to post-treatment lesion classification.

This diagnostic ambiguity has profound clinical implications. Misclassification of radiation necrosis as recurrent tumor can lead to unnecessary interventions such as re-irradiation or surgical resection, increasing the risk of adverse effects and patient morbidity. Conversely, mistaking viable tumor for necrosis may result in delayed treatment, allowing tumor progression and potentially compromising patient outcomes.

## 1.4 Advanced Imaging Techniques for Lesion Characterization

### 1.4.1 Conventional MRI Techniques

Magnetic Resonance Imaging (MRI) serves as the gold standard for detecting and evaluating brain metastases due to its exceptional soft-tissue contrast and spatial resolution[16]. Several MRI sequences provide complementary information for lesion characterization:

- **T1-Weighted Imaging (T1WI):** Provides high-resolution anatomical detail and is particularly useful for visualizing hemorrhagic lesions[17][18]. When combined with contrast agents, post-contrast T1WI becomes essential for detecting contrast-enhancing brain metastases, indicating blood-brain barrier disruption[19].
- **T2-Weighted Imaging (T2WI):** Valuable for detecting perilesional changes such as vasogenic edema and cystic components around metastatic brain lesions. T2 hyperintense signals commonly reflect extracellular fluid accumulation, often indicating vasogenic edema caused by tumor-related blood-brain barrier disruption[20].
- **FLAIR (Fluid-Attenuated Inversion Recovery):** Suppresses cerebrospinal fluid (CSF) signals, enhancing visibility of lesions adjacent to ventricles and cortical sulci. By nulling CSF, FLAIR improves detection of periventricular metastases, subtle edema, and leptomeningeal involvement[21][22].

### 1.4.2 Advanced MRI Techniques

To improve differential diagnosis beyond conventional imaging, several advanced MRI techniques have been developed:

- **Diffusion-Weighted Imaging (DWI):** Highly sensitive to water molecule movement within tissue, making it valuable for characterizing brain metastases. Restricted diffusion on DWI can help differentiate hypercellular tumors from abscesses or cystic lesions. Most brain metastases show facilitated (non-restricted) diffusion, while DWI helps detect acute ischemic changes, necrosis, or treatment-related effects[23].
- **Perfusion-Weighted Imaging (PWI):** Measures cerebral blood volume and flow and provides information on vascular properties of brain metastases. Metastatic lesions have characteristic sharply outlined regions of increased perfusion due to tumor angiogenesis. PWI helps in the distinction from high-grade gliomas and aids in assessing response to therapy[24].

### 1.4.3 Multimodal Advanced Techniques

Advanced MRI and nuclear medicine techniques provide additional information beyond conventional contrast-enhanced imaging:

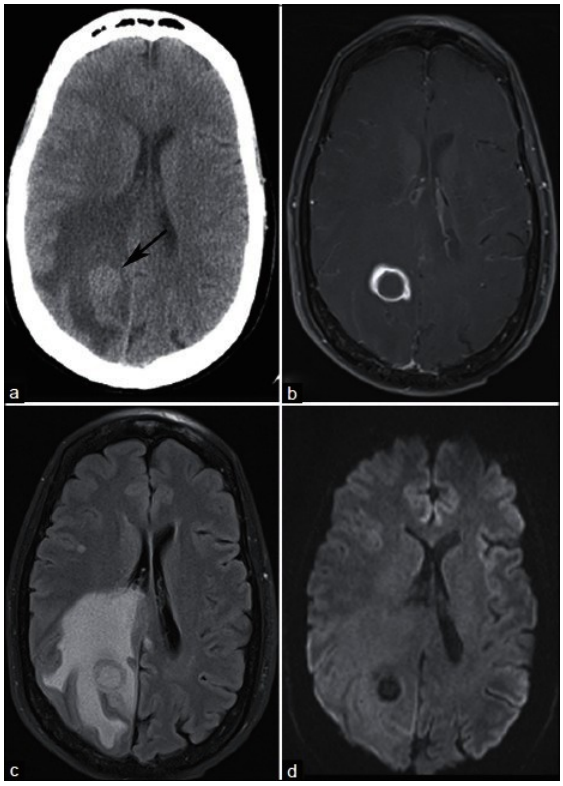



Figure 1.1: "A 59-year-old smoker with headache and balance issues. (a) NECT shows a right parietal mass at the gray–white junction with vasogenic edema. (b) Post-contrast T1 MRI reveals ring enhancement. (c) FLAIR confirms extensive edema. (d) DWI shows no central diffusion restriction, excluding abscess.Lung biopsy confirmed non-small cell lung cancer, and the brain lesion underwent stereotactic radiosurgery."[1]

- **Magnetic Resonance Spectroscopy (MRS):** Provides metabolic profiling by detecting elevated choline, lactate, and lipid peaks associated with tumor recurrence. High choline-to-creatine ratios suggest cellular proliferation, while increased lactate indicates necrosis[25][26].
- **Positron Emission Tomography (PET-CT):** Detects metabolic activity via FDG uptake, distinguishing viable tumor tissue from necrotic or fibrotic regions through metabolic rather than anatomical criteria[27].

## 1.5 Problem Statement and Research Objectives

Despite notable progress in imaging technology and DL tools, some important gaps still exist in automated assessment of lesions after Gamma Knife treatment. Present DL methods for brain metastasis classification often concentrate on one imaging type, not taking

full advantage of the wealth of multimodal data in clinical settings. Existing models frequently struggle with the severe class imbalance inherent in recurrence prediction, where stable lesions vastly outnumber recurrent cases. Furthermore, most published studies report limited sensitivity for detecting recurrent lesions, which is clinically the most critical metric for patient management.

The Brain-TR-GammaKnife dataset[4] provides an opportunity to address these limitations, containing comprehensive multimodal data including MRI scans, RTStruct files, RTDose distributions, and detailed clinical information from 47 patients with 244 lesions. However, this dataset exhibits severe class imbalance with only 23 recurrent lesions (9.4%) compared to 221 stable lesions (90.6%).

### 1.5.1 Research Objectives

This thesis addresses these limitations by developing a robust deep learning framework specifically designed to automatically classify brain lesions as stable or recurrent following Gamma Knife radiosurgery. The primary objectives include:

1. **Multimodal Integration:** Develop a unified neural network architecture that effectively combines Magnetic Resonance Imagin (MRI), radiotherapy dose distributions (RTDose), and structured clinical variables to leverage all available diagnostic information.
2. **Class Imbalance Mitigation:** Implement specialized techniques including focal loss[28], selective data augmentation, and class-balanced sampling strategies to enhance the model's ability to detect rare recurrent cases.
3. **Clinical Validation:** Demonstrate substantial improvements in sensitivity and F1-score compared to existing baselines, with particular emphasis on recurrence detection performance that would have meaningful clinical impact.
4. **Practical Implementation:** Ensure the framework is compatible with standard clinical data formats (DICOM) and can be integrated into existing radiological workflows

### 1.5.2 Expected Contributions

This research contributes to the growing field of automated neuro-oncology by providing a data-driven, objective approach to post-treatment lesion assessment. The proposed framework has the potential to reduce inter-observer variability, improve consistency in clinical decision-making, and ultimately enhance patient care by enabling more timely and accurate detection of lesion recurrence. By addressing the specific challenges of multimodal data integration and class imbalance in recurrence prediction, this work establishes a foundation for more sophisticated and clinically relevant automated diagnostic tools in stereotactic radiosurgery follow-up care.

# Chapter 2

# Related Work and State-of-the-Art

## 2.1 Deep Learning in Medical Imaging: Foundations and Evolution

Deep learning (DL) has fundamentally transformed medical imaging, enabling automated feature extraction and pattern recognition from complex, high-dimensional datasets. Unlike traditional machine learning methods that rely on handcrafted features engineered by domain experts, DL models learn hierarchical representations directly from raw data, leading to unprecedented performance improvements across various imaging tasks[29]. This paradigm shift has been particularly impactful in medical imaging, where the complexity and variability of anatomical structures, pathological presentations, and imaging artifacts present significant challenges for conventional image analysis approaches.

### 2.1.1 From 2D to 3D: Architectural Evolution

Medical imaging deep learning has undergone an evolutionary process from natural image processing to the unique requirements of medical images. The first was the use of the 2D Convolutional Neural Networks (CNNs) to process single volumetric medical image slices as separate images. This approach failed essentially in preserving the extensive spatial context inherent in three-dimensional medical images where slice-to-slice relations tend to contain important diagnostically relevant information.

The introduction of 3D CNNs marked a pivotal advancement, allowing models to process volumetric data holistically and capture spatial relationships across all three dimensions[29]. This evolution has been particularly beneficial for tasks involving brain imaging, where lesion characteristics, spatial relationships, and contextual information spanning multiple

slices are crucial for accurate diagnosis and classification. The ability to process entire volumes simultaneously has enabled deeper analysis of tumor morphology, growth patterns, and spatial distribution.

### 2.1.2 Architectural Innovations in Medical Imaging

Some dedicated architectures have been developed specifically for medical imaging tasks and deal with specific issues in the medical application space:

- **Residual Networks (ResNet):** The introduction of skip connections in ResNet addressed the vanishing gradient problem, enabling the training of much deeper networks[30]. In medical imaging, this has allowed for more sophisticated feature extraction and better representation of complex pathological patterns. ResNet architectures have been successfully adapted for various medical imaging tasks, from lesion detection to disease classification.
- **U-Net and Encoder-Decoder Networks:** First proposed for biomedical image segmentation, U-Net was the first to employ skip connections from encoder to decoder layers for the purpose of accurate localization with retention of contextual information[31]. It has since become the de-facto standard architecture for medical image segmentation tasks with several variants having been proposed for varying imaging modalities and anatomical locations.
- **Attention Mechanisms:** Attention-based networks have emerged as dominant architectures in medical imaging by allowing networks to selectively attend to clinically significant regions and attenuate irrelevant background information. Self-attention mechanisms enable the model to extract long-range dependencies in medical images, and channel attention enables the model to focus on significant feature maps[32][33][34]. These modules have been particularly useful in brain imaging cases where the subtle abnormalities can be spread across various anatomical regions.
- **Transformer-Based Architectures:** Recent advances in transformer models, originally developed for natural language processing, have been successfully adapted for medical imaging. Vision Transformers (ViTs) and hybrid CNN-Transformer architectures offer new approaches to modeling spatial relationships and have shown promising results in various medical imaging tasks[35].

### 2.1.3 Transfer Learning and Domain Adaptation

The scarcity of large, well-annotated medical datasets has made transfer learning a critical component of medical imaging applications. Pre-training on large-scale natural image datasets (such as ImageNet) followed by fine-tuning on medical data has become a standard practice, leveraging learned low-level features while adapting to domain-specific characteristics.

However, the domain gap between natural images and medical data has led to the development of specialized transfer learning strategies. Medical-specific pre-training, where

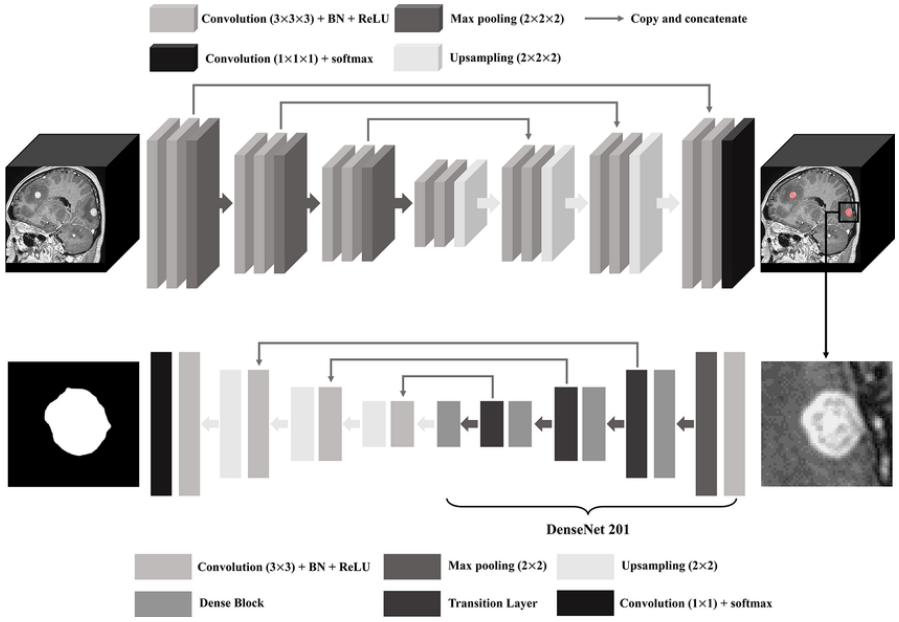



Figure 2.1: U-Net framework for detecting and segmenting brain metastases (BM), where BN means batch normalization and ReLU means rectified linear unit.[2].

models are first trained on large medical imaging datasets before fine-tuning on specific tasks, has shown superior performance compared to natural image pre-training. Additionally, domain adaptation techniques help bridge the gap between different imaging modalities, acquisition protocols, and institutional variations.

### 2.1.4 Data Augmentation in Medical Imaging

Data augmentation has really gained significance for enhancing model robustness and generalization in medical imaging, where the training datasets are typically limited. Medical-oriented augmentation techniques need to strike an intricate balance between data diversification and anatomical believability. Typical methods encompass geometric transformations (rotation, scaling, elastic deformation), intensity modifications (contrast shifting, noise injection), and higher-order techniques such as Generative Adversarial Networks (GANs) for the creation of synthetic data.

However, medical data augmentation must be done with consideration of anatomical limits and clinical utility. Certain transformations may violate anatomical rules or create unrealistic disease manifestations and thus adversely impact model performance and clinical utility.

## 2.2 Brain Metastases Detection and Classification: Current Approaches

The application of deep learning to brain metastases detection and classification has emerged as one of the most active areas of research in medical imaging AI[36][37]. The complexity of brain anatomy, the variability of metastatic presentations, and the critical importance of accurate diagnosis have driven substantial research efforts toward automated detection and classification systems.

### 2.2.1 Methods of Detection and Segmentation

Current approaches to brain metastases detection primarily rely on supervised learning with manually annotated datasets. Convolutional Neural Networks form the backbone of most detection systems, with architectures ranging from simple 2D CNNs applied slice-by-slice to sophisticated 3D networks that process entire brain volumes.

#### Object Detection Methods

Some works applied object detection methodologies formulated on natural images to the detection of brain metastases. R-CNN (Region-based CNN) variants, YOLO (You Only Look Once) structures, and single-shot detection architectures have been used successfully to detect and localize brain metastases. These methods usually attain detection sensitivity from 85% to 95% depending on the lesion dimensions, image quality, and the nature of the dataset[36].

#### Segmentation-Based Detection

The detection task receives an alternative treatment through segmentation techniques that detect metastatic regions by outlining them at the pixel scale. The use of U-Net and its derivatives has dominated this task and recent research demonstrates Dice scores above 0.90 when dealing with well-defined lesions[38]. The implementation of 3D U-Net models demonstrates strong potential for brain volume processing because these models extract slice-based contextual data while preserving exact spatial placement.

#### Cascade and Multi-Stage Approaches

Researchers have developed multiple cascade methods because they want to detect small metastases without increasing computational costs. The methods start with initial basic detection and then perform detailed analysis on suspected regions. The approaches achieve a balance between high detection accuracy and low false positive rates through effective computational management[39].

### 2.2.2 Classification Challenges: Stable vs. Recurrent Lesions

The classification of brain lesions as stable or recurrent following treatment represents one of the most challenging problems in neuro-oncology imaging. This task requires

distinguishing between subtle changes that may indicate tumor progression and treatment-related effects such as radiation necrosis[14][15].

#### Single-Modality Approaches

The initial automated classification systems used T1-weighted contrast-enhanced MRI because it represents the standard clinical imaging sequence for brain metastases follow-up. These early approaches focused on training CNN models through lesion patches that were taken from follow-up scan images. The methods achieved reasonable overall accuracy yet they demonstrated poor sensitivity to detect recurrent lesions that showed subtle changes in imaging. changes.

#### Radiomics and Handcrafted Features

Medical images underwent feature extraction through radiomics methods during the pre-deep learning era by calculating shape descriptors and texture measures and intensity statistics[14][15]. Traditional machine learning classifiers processed these features after their extraction. Radiomics approaches produced understandable results but their dependence on feature engineering proved restrictive because they struggled to detect complex patterns which deep learning methods easily recognized.

#### Longitudinal Analysis

Several approaches have attempted to leverage temporal information by comparing follow-up scans with baseline post-treatment images. These methods aim to detect changes over time rather than classifying individual time points. However, registration challenges, variable scan timing, and differences in acquisition parameters have limited the success of longitudinal approaches.

### 2.2.3 Baseline Models and Performance Benchmarks

The development of brain metastases detection and classification methods requires researchers to build performance benchmarks as their primary foundation. Several baseline architectures have emerged as standard comparison points:

- **3D CNN Baselines:** Simple 3D CNN architectures with progressively increasing filter sizes have served as foundational baselines. These models typically consist of 3-5 convolutional layers followed by global pooling and dense classification layers. While computationally efficient, these basic architectures often lack the representational power needed for complex classification tasks.
- **ResNet Adaptations:** 3D ResNet architectures have been widely adopted as stronger baselines, offering improved gradient flow and enabling deeper networks. ResNet-18 and ResNet-50 variants adapted for 3D medical data have shown consistent performance improvements over simpler CNN architectures.

- **Medical-Specific Architectures:** Several architectures have been specifically designed for medical imaging applications. Med3D, a 3D CNN pre-trained on medical imaging data, has provided strong baseline performance across multiple medical imaging tasks. Similarly, architectures like DenseNet, adapted for 3D medical data, have offered competitive baseline performance while maintaining parameter efficiency.

## 2.3 Multimodal Deep Learning in Medical Imaging

The integration of multiple data modalities represents a significant trend in medical imaging AI, driven by the recognition that different imaging sequences and data types provide complementary diagnostic information[40]. In the context of brain metastases, the combination of structural imaging, treatment planning data, and clinical information offers the potential for more comprehensive and accurate assessment.

### 2.3.1 Multimodal Fusion Strategies

#### Early Fusion

The early fusion technique merges various imaging sequences during the initial input stage through channel concatenation. This method requires minimal computation but fails to leverage individual modality strengths effectively because dominant signal modalities tend to control the process.

#### Late Fusion

Late fusion processes each modality independently through separate network branches before combining the learned representations at the decision level. This approach allows each modality-specific network to learn optimal representations while enabling flexible combination strategies. Most successful multimodal medical imaging systems employ variants of late fusion[40].

#### Intermediate Fusion

Intermediate fusion combines modalities at various levels within the network architecture,allowing for both modality-specific and cross-modal feature learning. Attention mechanisms are often employed to weight the contribution of different modalities dynamically based on input characteristics.

### 2.3.2 Imaging and Non-Imaging Modality Integration

#### Structural and Functional Imaging

The combination of structural MRI sequences (T1, T2, FLAIR) with functional information (DWI, PWI, MRS) has shown significant promise for brain metastases characterization. Each modality provides unique information: structural sequences reveal anatomical

details, diffusion imaging provides cellular density information, perfusion imaging reveals vascular characteristics, and spectroscopy offers metabolic insights[25][26].

#### Treatment Planning Integration

Post-treatment brain metastases analysis benefits from having treatment planning data which includes RTDose and RTStruct files[4]. Post-treatment imaging combined with this information allows researchers to analyze dose-response relationships and spatial correlations which link treatment delivery to outcomes.

#### Clinical Data Integration

The incorporation of structured clinical variables (patient demographics, treatment parameters, temporal information) with imaging data has shown promise for improving classification performance. However, the integration of heterogeneous data types requires careful architectural design and often involves embedding layers for categorical variables and normalization strategies for continuous variables.

### 2.3.3 Challenges in Multimodal Learning

#### Modality Imbalance

Different modalities may contribute unequally to the final decision, with some modalities dominating the learning process. Balancing contributions requires careful architectural design and often involves modality-specific loss functions or attention mechanisms.

#### Missing Modalities

Multiple clinical datasets include missing modalities because of acquisition failures together with protocol variations and temporal constraints during data collection. The system needs to support incomplete data inputs by either utilizing imputation methods or designing architectures that operate without full information.

#### Temporal Alignment

When combining data acquired at different time points or with different protocols, temporal and spatial alignment becomes critical. Registration errors and acquisition differences can significantly impact multimodal fusion performance.

## 2.4 Datasets and Benchmarks for Brain Metastases Research

The development of robust deep learning systems for brain metastases analysis has been significantly facilitated by the availability of curated datasets with expert annotations. These datasets serve not only as training resources but also as benchmarks for comparing different algorithmic approaches.

### 2.4.1 Public Datasets and Their Characteristics

#### Brain-TR-GammaKnife Dataset

The Cancer Imaging Archive (TCIA) stores a dataset with 47 patient records which contain 244 lesions and their associated MRI images along with radiation therapy dose plans (RTDose) and structure sets (RTStruct) and clinical data[4]. This dataset exists to help predict tumor recurrence after Gamma Knife radiosurgery which connects it to post-treatment observation needs.

#### UCSF-BMSR Dataset

The University of California San Francisco Brain Metastases Stereotactic Radiosurgery dataset contains 560 multimodal brain MRI scans with expert annotations for 412 patients who received Gamma Knife treatment[41]. This dataset contains more patient data than others and it serves mostly for detection and segmentation analysis.

#### Comprehensive Annotated Brain Metastasis Dataset

The dataset published in Scientific Data that contains 637 high-resolution imaging studies from 75 patients includes 260 brain metastasis lesions with clinical data and semi-automatic segmentations[42]. The dataset contains multiple imaging sequences along with detailed clinical annotations.

### 2.4.2 Dataset Limitations and Challenges

#### Limited Size and Diversity

Most available datasets are relatively small by deep learning standards, often containing fewer than 1000 lesions. This limitation is particularly challenging for training robust deep learning models and may lead to overfitting and poor generalization.

#### Institutional Bias

Many datasets originate from single institutions, potentially limiting generalizability across different imaging protocols, patient populations, and clinical practices. Multi-institutional datasets are rare due to data sharing constraints and standardization challenges.

#### Annotation Quality and Consistency

Manual annotation of brain metastases requires significant expertise and is subject to inter-observer variability. Different datasets may employ different annotation protocols, making cross-dataset evaluation challenging.

#### Class Imbalance

Virtually all brain metastases datasets exhibit severe class imbalance, with recurrent lesions representing a small minority of cases[4]. This imbalance reflects the clinical reality

but poses significant challenges for machine learning algorithms.

### 2.4.3 Preprocessing and Standardization Challenges

#### Image Acquisition Variability

The characteristics of images become different when MRI scanners use various field strengths along with various acquisition protocols. The standardization process across datasets needs advanced preprocessing pipelines which fail to completely remove systematic variations.

#### ROI Extraction Consistency

The method used to define and extract regions of interest around lesions has a substantial effect on model performance. The results of studies show inconsistent outcomes because different methods of ROI extraction (fixed vs. adaptive size, margin inclusion, background handling) exist.

#### Normalization Strategies

The process of intensity normalization between different imaging sessions and scanners proves to be difficult. Research employs various normalization approaches which range from basic histogram matching to complex tissue-based normalization but results vary depending on the method.

## 2.5 Current Limitations and Research Gaps

Despite significant advances in deep learning applications for brain metastases analysis, several critical limitations and research gaps remain that limit the clinical translation and widespread adoption of these technologies.

#### Limited Sensitivity for Recurrence Detection

Most published studies report suboptimal sensitivity for detecting recurrent lesions, which is clinically the most critical metric[4]. The severe class imbalance and subtle imaging changes associated with early recurrence contribute to this limitation.

#### Single-Modality Focus

Many approaches focus on single imaging modalities, failing to leverage the rich multimodal information routinely available in clinical practice. The integration of treatment planning data, multiple imaging sequences, and clinical variables remains underexplored.

#### Temporal Modeling Limitations

The majority of existing approaches process imaging time points as separate entities which prevents them from using temporal data and change patterns that could help predict recurrence.

#### Generalization Challenges

The models demonstrate strong performance on data from their origin institution yet they struggle to generalize across different scanners and protocols and patient populations. The limitation creates major obstacles to using these models in clinical settings.

#### Interpretability and Trust

Deep learning models often function as "black boxes," providing limited insight into decision-making processes. The lack of interpretability poses significant barriers to clinical acceptance and regulatory approval.

#### Integration with Clinical Workflows

Most research focuses on algorithmic performance while neglecting practical aspects of clinical integration, including computational requirements, user interfaces, and workflow compatibility.

#### Validation Standards

The medical imaging community lacks standardized evaluation protocols for comparing different approaches, making it difficult to assess relative performance and clinical utility.

#### Regulatory Pathways

The path to regulatory approval for AI-based medical imaging tools remains complex and uncertain, particularly for applications involving treatment decision support.

### 2.5.1 Methodological Gaps

#### Uncertainty Quantification

Most approaches provide point predictions without uncertainty estimates, limiting their clinical utility where confidence assessments are crucial for decision-making.

#### Robustness Analysis

Limited attention has been paid to model robustness against acquisition variations, artifacts, and adversarial examples that may occur in clinical practice.

### Long-term Outcome Correlation

Most studies focus on short-term imaging endpoints rather than correlating predictions with long-term clinical outcomes such as survival and quality of life.

### Multi-institutional Validation

Large-scale, multi-institutional validation studies are lacking, which reduces trust in clinical utility and generalizability in a variety of contexts.

There are a lot of areas for improvement, especially in multimodal integration, class imbalance handling, and clinical translation, according to this thorough review of the state-of-the-art. By creating a unique multimodal framework especially for post-Gamma Knife recurrence prediction, the work presented in this thesis fills in a number of these gaps.

# Chapter 3

# Dataset and Preprocessing

## 3.1 Dataset Overview and Characteristics

The main source of data for this study is the Brain-TR-GammaKnife dataset, which is one of the largest publicly accessible sets of multimodal brain metastases data created especially for recurrence prediction studies [4]. Together, the University of Mississippi Medical Center (UMMC) and Mississippi State University (MSU) created this dataset. Strict institutional review board approval (IRB-2017-0266) and complete HIPAA compliance were used to protect patient privacy during data collection.

The dataset includes complete follow-up imaging and clinical documentation for 47 patients who had Gamma Knife stereotactic radiosurgery for brain metastases. The dataset is publicly available through The Cancer Imaging Archive (TCIA) under DOI: 10.7937/xb6d-py67, and all patient identifiers have been fully anonymized to enable reproducible research and algorithm comparison across the research community.

### 3.1.1 Patient Demographics and Clinical Characteristics

The dataset shows a balanced gender distribution with 26 female patients (55.3%) and 21 male patients (44.7%), reflecting the typical demographics of brain metastases populations where breast cancer metastases contribute significantly to the overall incidence[9]. The patient class represents diverse primary cancer types, with lung cancer, breast cancer, and melanoma being the most common sources of brain metastases, consistent with established epidemiological patterns[43].

Follow-up duration varies significantly across patients, with 30 patients (63.8%) having single-visit data, 10 patients (21.3%) with two visits, 6 patients (12.8%) with three visits, and one patient (2.1%) with eight documented visits. This distribution reflects real-world clinical practice where follow-up intensity depends on patient prognosis, treatment response, and clinical presentation.

| Visits per Patient | Number of Patients | Percentage (%) |
|---|---|---|
| 1 | 30 | 63.8% |
| 2 | 10 | 21.3% |
| 3 | 6 | 12.8% |
| 4–7 | 0 | 0.0% |
| 8 | 1 | 2.1% |

Table 3.1: Distribution of Hospital Visits per Patient

### 3.1.2 Lesion-Level Characteristics and Distribution

The dataset contains 244 individual lesions across all patients, with the critical clinical challenge being the severe class imbalance: only 23 lesions (9.4%) are classified as recurrent, while 221 lesions (90.6%) remain stable during the follow-up period. Although this distribution presents major obstacles for the development of machine learning algorithms, it accurately depicts the clinical reality of post-Gamma Knife treatment outcomes, where the majority of lesions achieve durable local control.

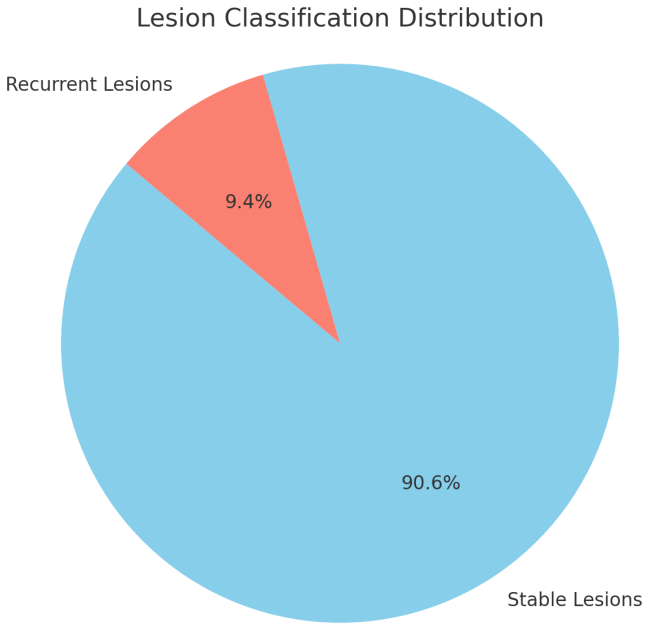



Figure 3.1: Class distribution showing severe imbalance between stable and recurrent lesions in the Brain-TR-GammaKnife dataset

Each patient has a very different distribution of lesions; some have only one metastasis, while others have several lesions that need to be treated at the same time. Of the 47

patients, 32 had stable disease in all treated lesions, and 15 had at least one recurrent lesion at follow-up. In order to ensure appropriate data splitting and stop information leakage between training and testing sets, this patient-level information is essential.

## 3.2 Data Composition and Modalities

The Brain-TR-GammaKnife dataset distinguishes itself through comprehensive multimodal data collection, providing researchers with access to complementary information sources that collectively enable sophisticated analysis of treatment outcomes and recurrence patterns.

### MRI Volumes

16,792 DICOM files containing high-resolution structural brain MRI acquisitions are included in the dataset. Standard clinical procedures that are optimized for the detection of brain metastases and follow-up evaluation were used to obtain these images. Essential anatomical context, lesion morphology details, and baseline structural references required for treatment planning and outcome evaluation are provided by the MRI data.

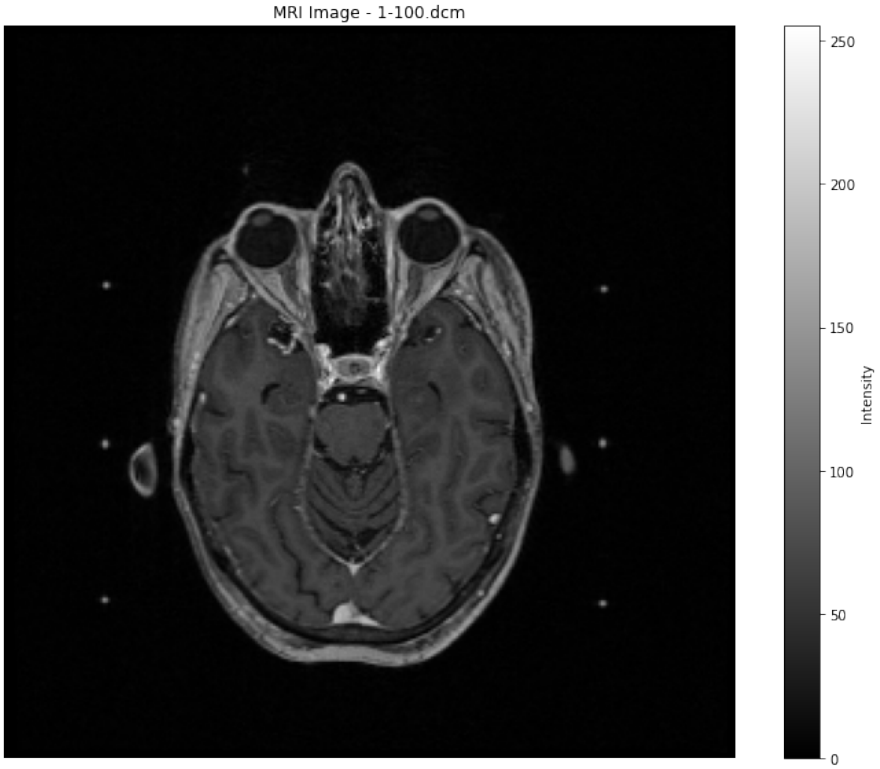



Figure 3.2: Representative MRI slice showing brain metastasis with surrounding anatomical structures

### RTDose Data

Three dimensional dose distribution maps quantify the spatial radiation exposure delivered during Gamma Knife treatment. These volumetric datasets contain voxel-level dose information in Gray (Gy) units, spatially registered to the corresponding MRI volumes. RTDose data enables analysis of dose-response relationships and spatial correlations between radiation delivery patterns and treatment outcomes.

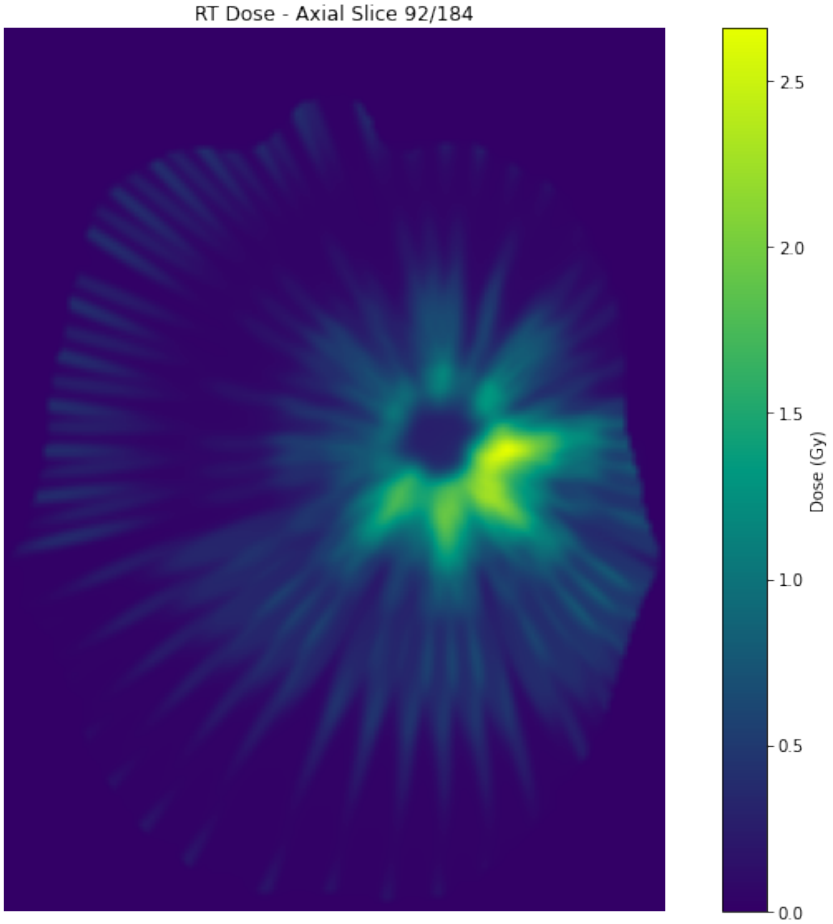



Figure 3.3: RTDose visualization showing radiation dose distribution overlaid on anatomical structures

### RTStruct Data

During treatment planning, skilled radiation oncologists manually define the critical anatomical structures and lesion boundaries using precise geometric definitions found in radiation therapy structure sets. Multiple regions of interest (ROIs) with anatomically descriptive names that correspond to lesion locations are included in each RTStruct file. These structures have two functions: they provide ground truth segmentation masks for machine

learning applications and define treatment targets during Gamma Knife planning.

### 3.2.1 Clinical Data Structure

The clinical information is systematically organized in a structured Excel spreadsheet (Brain-TR-GammaKnife-Clinical-Information.xlsx) containing three structure data sheets:

#### Patient-Level Data (pt_level)

Contains 47 records with unique patient identifiers enabling linkage across all data modalities while maintaining anonymization protocols.

#### Course-Level Data (course_level)

Provides 76 records documenting treatment courses, including primary cancer diagnosis, patient demographics (age at diagnosis, gender), and treatment-specific information. This level captures the relationship between primary cancer characteristics and brain metastases presentation.

| unique_pt_id | Course # | Diagnosis | Primary Diagnosis | Age | Gender |
|---|---|---|---|---|---|
| 103 | 1 | Mets Ovary | Serous carcinoma | 75 | Female |
| 114 | 1 | Brain Mets-Breast | Invasive ductal carcinoma | 60 | Female |
| ... | ... | ... | ... | ... | ... |

Table 3.2: Example clinical records from the course-level data sheet

#### Lesion-Level Data (lesion_level)

The most granular level contains 244 records corresponding to individual lesions, including anatomical location, recurrence status (mri_type), temporal information (duration_tx_to_imag), treatment fractionation details, and standardized lesion nomenclature for cross-referencing with imaging data.

| No. | unique_pt_id | Treatment | Course | Lesion# | Lesion Location | mri_type | duration_tx_to_imag | Fractions | Lesion Name NRRD files |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 463 | 1 | 1 | 1 | Lt Frontal | recurrence | 11 | 1 | GK.463_1_LLtFronta |
| 2 | 463 | 2 | 2 | 2 | R Motor Cortex | stable | 8 | 1 | GK.463_2_LRTMotorCortex |
| 3 | 463 | 2 | 3 | 3 | Lt Post Temporal | stable | 8 | 1 | GK.463_2_LLtPostTemporal |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

Table 3.3: Example clinical records from the lesion-level data sheet

## 3.3 Architecture of the Preprocessing Pipeline

Raw multimodal clinical data is transformed into standardized, analysis ready inputs for deep learning applications by the preprocessing pipeline. Basic issues in medical imaging preprocessing are addressed by the pipeline design, including managing missing data,

preserving clinical validity during the transformation process, handling heterogeneous data formats, and guaranteeing spatial consistency across modalities.

The preprocessing architecture follows several key principles established for medical imaging applications. Data integrity preservation ensures that all transformations maintain clinical relevance and anatomical validity. Reproducibility is achieved through deterministic processing steps and comprehensive logging of all transformations. Scalability considerations enable efficient processing of the entire dataset while supporting future expansion to larger cohorts.

Quality assurance mechanisms are integrated throughout the pipeline, including automated validation checks, error handling procedures, and comprehensive logging systems that track processing success and failure rates. This systematic approach ensures reliable data preparation for subsequent machine learning applications.

### 3.3.1 File Organization and Data Discovery

The dataset employs a hierarchical folder structure that separates data by patient and clinical visit. Each patient folder (e.g., GK_103, GK_114) contains visit-specific subfolders named with standardized patterns including visit date, imaging protocol, and internal identifiers.

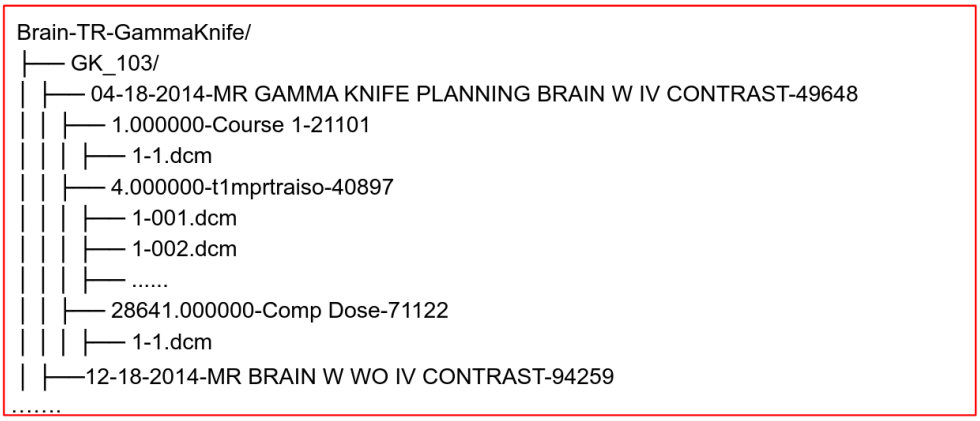



Figure 3.4: Hierarchical file structure organization of the Brain-TR-GammaKnife dataset showing patient and visit-level organization

Within each visit folder, DICOM files are organized by acquisition type, though folder names may not directly indicate content type, necessitating metadata-based file classification. Automated file discovery algorithms recursively scan the directory structure, examining DICOM headers to classify files by modality type (MR, RTSTRUCT, RTDOSE). This approach ensures robust handling of varying folder organization patterns while maintaining compatibility with standard DICOM data structures.

## 3.4 Standardization and Image Processing

### 3.4.1 Reading and Validating DICOM Data

Comprehensive DICOM file reading using specialized medical imaging libraries (pydicom, SimpleITK, nibabel) is the first step in the initial processing process. Essential metadata, such as patient identifiers, acquisition parameters, spatial positioning data, and modality classification, are extracted from each DICOM file through header validation. Files that lack important metadata or have corrupted headers are marked for manual review or removal.

MRI volume reconstruction uses spatial position information encoded in DICOM headers to merge separate DICOM slices into cohesive 3D volumes. Through the analysis of Image Position Patient and Image Orientation Patient tags, slice ordering algorithms guarantee proper anatomical orientation. Quality control checks find possible gaps or overlapping slices and confirm that the volume is complete.

### 3.4.2 Spatial Standardization and Resampling

Spatial standardization addresses variations in voxel spacing, image dimensions, and field of view across different acquisition sessions. All MRI volumes are resampled to a uniform voxel spacing and standardized dimensions ($256 \times 256 \times 256$) using trilinear interpolation for continuous data and nearest-neighbor interpolation for label data.

The resampling process preserves anatomical relationships while enabling consistent processing across all patients. Spatial transformation matrices are carefully maintained to enable accurate alignment between different modalities and proper coordinate system transformations during subsequent processing steps.

### 3.4.3 Intensity Normalization and Preprocessing

Variations in signal characteristics brought on by various scanner types, acquisition protocols, and temporal factors are addressed by intensity normalization. Within brain tissue regions, a robust normalization pipeline uses z-score standardization after histogram-based normalization. In order to focus normalization on pertinent brain tissue and eliminate non-brain structures that might skew intensity statistics, skull stripping is carried out using algorithms that have been proven to work.

Automated normalization failure detection, the detection of anomalous intensity distributions that might point to acquisition artifacts, and the comparison of normalized intensity ranges with physiologically expected values are examples of quality control procedures. against expected physiological values.

## 3.5 ROI Extraction and Lesion Localization

### 3.5.1 Metadata Matching and Validation

A crucial preprocessing challenge is ensuring accurate correspondence between imaging data and clinical metadata. To create trustworthy relationships between RTStruct ROI names and clinical lesion descriptions, the matching algorithm uses a complex two-stage process that combines rule-based filtering with similarity-based scoring.

In order to ensure temporal and patient-specific consistency, rule-based filtering first reduces possible matches by patient identifier and treatment course. Similarity scoring algorithms then use fuzzy string matching and normalized text processing to compare ROI names with clinical lesion location descriptions.

A thorough mapping dictionary that covers common variances in clinical nomenclature is used in the normalization process to standardize anatomical terminology. For instance, "Left" variants such as "Lt," "L," and "left" are standardized to a single representation. Likewise, anatomical region names like "Cerebellar" include variants like "Cereb," "Cerebe," and "Cerebellum."

| **Standardized Term** | **Accepted Variations** |
|---|---|
| Left | “Lt”, “L”, “left”, “Left” |
| Cerebellar | “Cereb”, “Cerebe”, “Cerebellar”, “Cerebllar”, “Cerbellar”, “cerebellum” |
| Occipital | “Occip”, “Occi”, “Occipit”, “occipital” |
| Right | “Rt”, “R”, “right”, “Right” |
| ... | ... |

Table 3.4: Example anatomical terminology standardization mapping for ROI matching

### 3.5.2 Similarity Scoring Algorithm

The matching algorithm computes similarity scores using both location-based and name-based comparisons. Location similarity employs set-based intersection analysis between normalized ROI location terms and clinical lesion location descriptions, computing Jaccard-like similarity coefficients:

$$\text{Location Score} = \frac{|\text{Location}_{ROI} \cap \text{Location}_{Excel}|}{|\text{Location}_{ROI} \cup \text{Location}_{Excel}|}$$

Name-based similarity compares ROI identifiers with standardized lesion names from clinical records using string similarity metrics. Combined scores integrate both similarity measures using weighted averages (typically 60% location similarity, 40% name similarity):

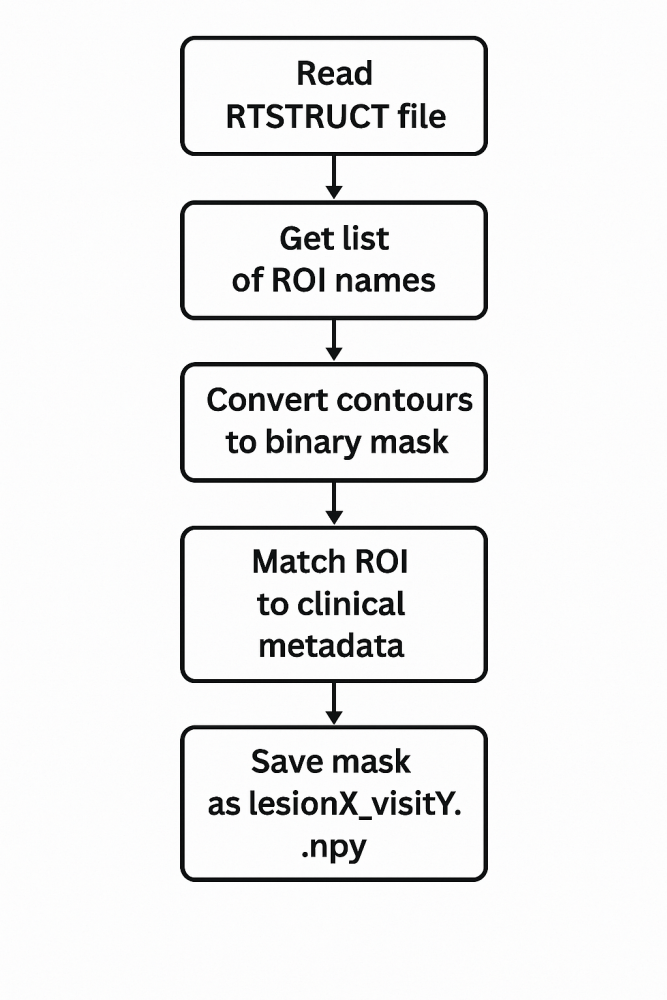



Figure 3.5: Flowchart illustrating the ROI mask generation process from RTStruct data to binary volumetric masks

$$\text{Combined Score} = 0.6 \times \text{Location Score} + 0.4 \times \text{Name Score}$$

Empirically determined thresholds (>0.2) are used for accepting matches. Low-confidence matches (<0.4) are flagged for manual review, while matches below acceptance thresholds are excluded from further processing.

### 3.5.3 ROI Mask Generation

Using the rt-utils library, validated ROI matches are converted from vector-based DICOM RTStruct contours to volumetric binary masks. Lesion regions are designated as 1 and

background regions as 0 in the binary volumes created by this conversion process, which interpolates 2D contour data across 3D space.

Validation procedures are included in mask generation to guarantee anatomical completeness and consistency. Masks that have unrealistic volumes or are empty are marked for review. To maintain binary values, all produced masks are resampled to uniform dimensions using nearest-neighbor interpolation after being spatially aligned with matching MRI volumes.

## 3.6 Multimodal Data Alignment and Integration

### 3.6.1 Cross-Modal Spatial Registration

Spatial alignment between different imaging modalities requires careful handling of coordinate systems, resolution differences, and acquisition timing variations. The alignment process establishes MRI volumes as the spatial reference, with RTDose and RTStruct data transformed to match MRI coordinate systems.

RTDose volumes undergo trilinear interpolation during resampling to the MRI spatial grid, preserving dose value relationships while achieving geometric consistency. Spatial transformation matrices are computed using DICOM spatial metadata and validated through anatomical landmark verification.

### 3.6.2 Lesion-Centered Patch Extraction

After spatial alignment, bounding box computation around binary mask regions is used to extract lesion-specific regions. In order to incorporate pertinent anatomical context while preserving computational efficiency, bounding boxes are enlarged with adjustable margins (usually 5–10 voxels).

Patch extraction provides uniform input dimensions for deep learning models by generating standardized 64×64×64 voxel regions centered on lesion locations. The extraction algorithm centers the patch on the centroid of the lesion if the lesion is larger than the patch size. Zero-filled arrays that match the dimensions of the MRI patch are used to fill in the missing RTDose data.

### 3.6.3 Clinical Data Integration

Clinical metadata undergoes preprocessing to create structured feature vectors suitable for neural network integration. Categorical variables (lesion location, primary diagnosis) are encoded using one-hot encoding schemes, while continuous variables (duration between treatment and imaging, patient age) are normalized using robust scaling techniques.

The clinical feature encoding process includes handling of missing values through appropriate imputation strategies and validation of feature distributions to identify potential

data quality issues. Encoded clinical features are stored alongside corresponding image patches to enable multimodal model training.

#### One-Hot Encoding Implementation

One-hot encoding converts categorical variables into binary vector representations suitable for neural network processing. Each possible category value is represented by a unique position in the binary vector, with a value of 1 indicating the presence of that category and 0 otherwise.

For lesion location encoding with 4 possible locations:

- "Frontal" $\rightarrow$ [1, 0, 0, 0]
- "Cerebellar" $\rightarrow$ [0, 0, 1, 0]

This encoding ensures that categorical variables are treated as distinct classes without introducing artificial ordinal relationships that could bias model learning.

## 3.7 Data Quality Assurance and Validation

Comprehensive quality control mechanisms operate throughout the preprocessing pipeline to identify and handle various data quality issues. Automated checks include validation of DICOM file integrity, verification of spatial consistency between modalities, detection of missing or corrupted data, and identification of outlier values that may indicate processing errors.

Statistical validation procedures analyze lesion size distributions, intensity characteristics, and spatial relationships to identify potential anomalies. Lesions with extremely small or large volumes compared to population distributions are flagged for manual review to ensure clinical validity.

Robust error handling ensures pipeline stability while maintaining data quality standards. Modality availability checks verify the presence of required imaging data (MRI and RT-Struct) for each patient visit, with optional RTDose data handled gracefully through zero-substitution strategies.

File reading operations employ exception handling to manage corrupted DICOM files without terminating the entire pipeline. All processing errors are logged with detailed information including patient identifiers, error types, and contextual information to facilitate debugging and quality improvement.

Validation procedures verify the consistency and completeness of processed data before machine learning applications. Cross-modal alignment is validated through anatomical landmark verification and spatial overlap analysis. Lesion mask validity is confirmed through volume analysis and boundary condition checks.

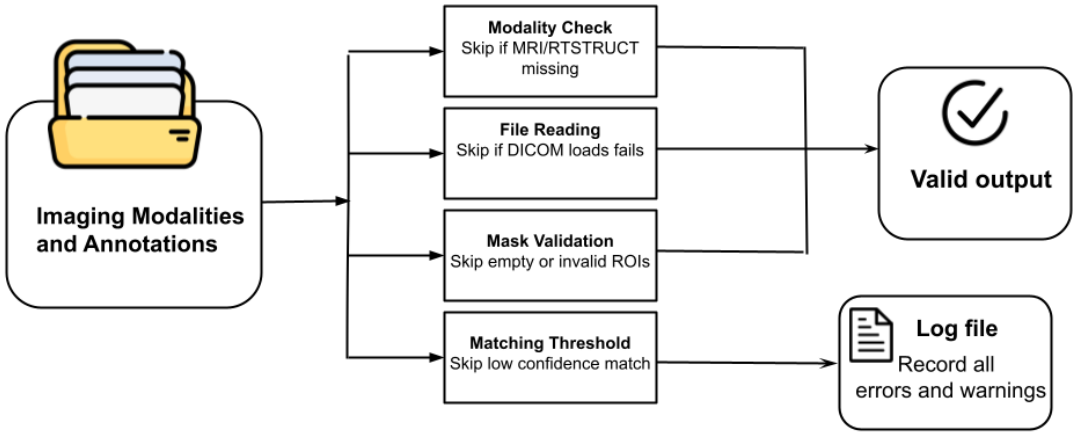



Figure 3.6: Data cleaning and exception handling workflow showing decision points and error recovery strategies

Processed data undergoes final validation including verification of feature vector completeness, confirmation of proper data type assignments, and validation of file naming conventions that enable reliable data loading during model training.

## 3.8 Dataset Splitting and Experimental Setup

### 3.8.1 Patient-Level Stratified Splitting

To ensure fair evaluation and prevent data leakage, the dataset is partitioned at the patient level using stratified sampling that preserves class distribution across training, validation, and test sets. This approach prevents any patient's lesions from appearing in multiple splits while maintaining representative samples of recurrent and stable lesions in each subset.

The splitting strategy allocates approximately 60% of patients to training, 20% to validation, and 20% to testing. Stratification ensures proportional representation of patients

with recurrent lesions across all splits, maintaining the challenging class imbalance while enabling fair model evaluation.

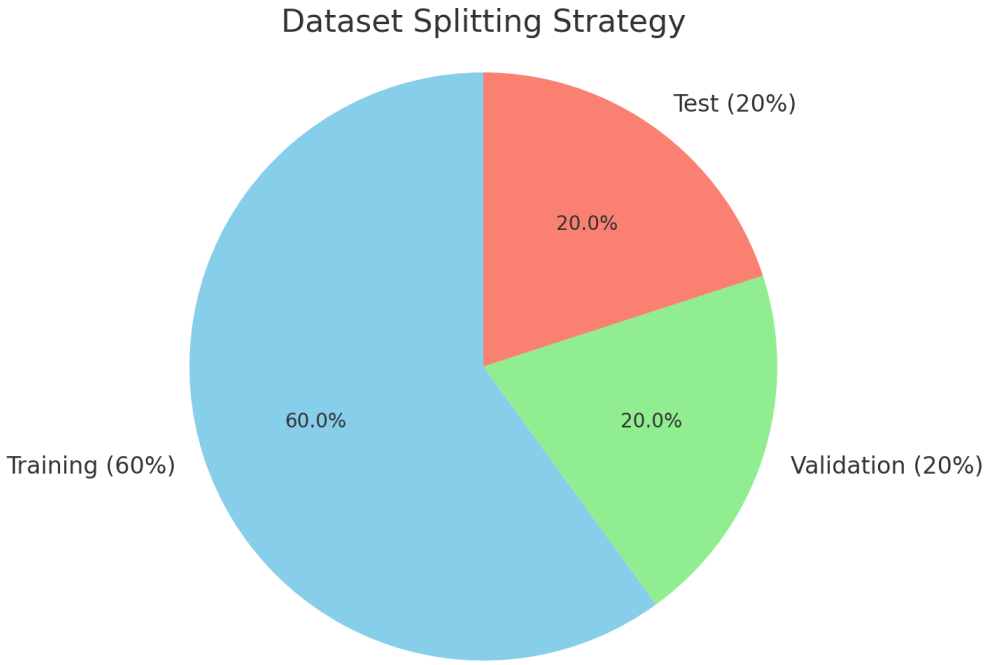



Figure 3.7: Dataset splitting strategy showing patient-level partitioning with stratified sampling to maintain class distribution

### 3.8.2 Cross-Validation Considerations

The severe class imbalance and small dataset size demand rigorous validation strategies to ensure reliable performance estimates. Instead of lesion-level splitting, which can place multiple lesions from the same patient in both training and test sets, leading to overly optimistic results, we employ patient-level splitting so that data from each individual is confined to a single partition. Within this framework, stratification by label preserves class proportions across splits, and fixing a random seed guarantees reproducible splits across experimental runs. We isolate the test set completely until the final evaluation, reserving the validation set exclusively for early stopping and hyperparameter tuning.

To counteract the roughly 10 : 90 imbalance between recurrent and stable lesions, we apply data augmentation selectively during training: only recurrent lesions are augmented, while the validation and test sets remain untouched to maintain an unbiased assessment. Augmentation techniques include spatial transformations (random rotations between $\pm 10°$ and $\pm 30°$, flips), intensity adjustments (scaling, noise injection), and elastic deformations that respect anatomical plausibility. By generating two to four augmented variants of each recurrent lesion per epoch, we achieve a more balanced class representation in each

training batch without compromising the integrity of the underlying dataset.

### 3.8.3 Output Data Structure

The preprocessing pipeline generates a standardized output structure that facilitates efficient data loading and model training. Each patient directory contains organized subdirectories with processed data:

- **Full Brain MRI Volumes:** Resampled to 256×256×256 resolution and saved as .npy files for each visit
- **RTDose Volumes:** Aligned to MRI space and saved in matching resolution
- **Lesion Masks:** Full-resolution binary masks for each identified lesion
- **Cropped Lesion Patches:** 64×64×64 patches centered on lesions for direct model input
- **Metadata Files:** JSON format containing clinical information, processing parameters, and quality metrics

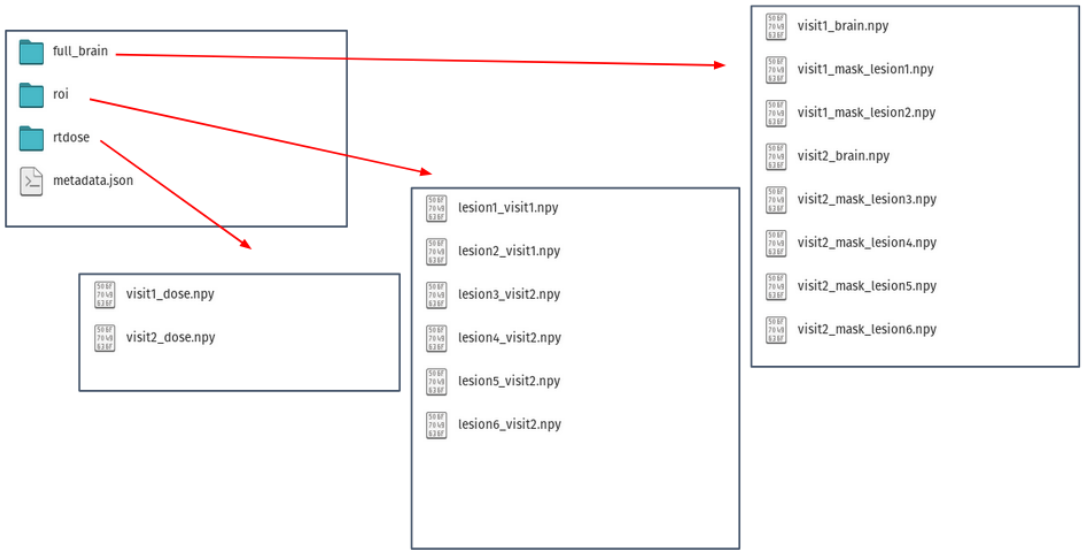



Figure 3.8: Example output directory structure showing organized processed data files for a single patient

Effective data loading during model training is made possible by this standardized output format, which also preserves complete traceability of preprocessing procedures and clinical metadata. Reproducible experimental workflows and flexible model architectures are supported by the modular structure.

The preprocessing pipeline effectively converts the unprocessed Brain-TR-GammaKnife dataset into multimodal inputs that are ready for analysis and can be used in deep learning

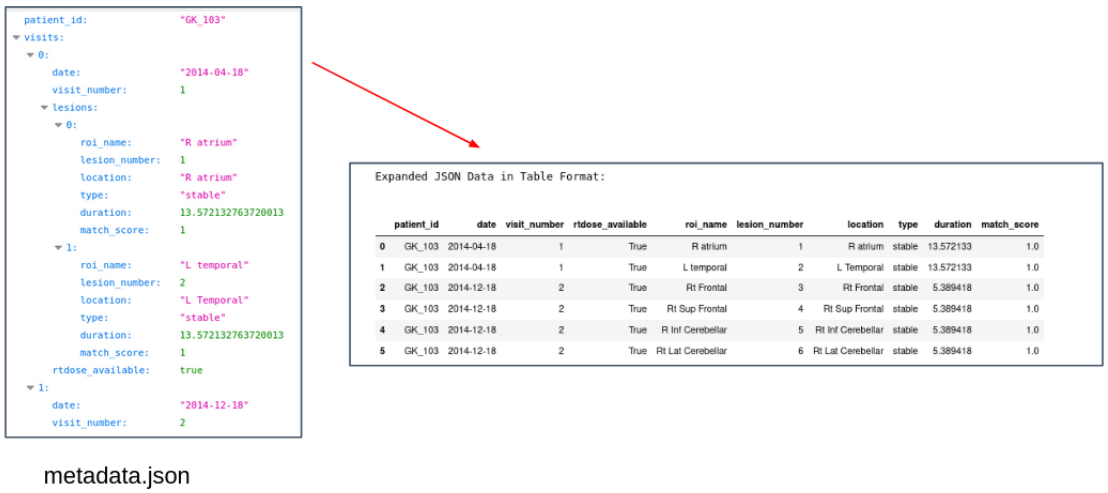



Figure 3.9: Example JSON metadata file containing structured clinical information and processing parameters

applications. This all encompassing strategy preserves data quality and clinical validity throughout the transformation process while addressing the particular difficulties of medical imaging preprocessing.

# Chapter 4

# Methodology

After stereotactic radiosurgery, it is important to correctly and quickly detect the recurrence of brain lesions. One of the most critical challenges faced by neuro-oncology is, after initial treatment, monitoring patients for signs of tumor progression or recurrence. This is especially important for patients receiving Gamma Knife radiosurgery, a highly localized, non-invasive radiation treatment frequently used to treat both primary and metastatic brain tumors. Although Gamma Knife therapy helps many patients achieve local control, some continue to experience lesion recurrence, usually months after treatment. Early detection of these recurrences may allow the management of additional therapeutic interventions, which could enhance quality of life and survival.

## 4.1 Problem Definition

In this study, the recurrence prediction task is formulated as a binary classification problem, where each lesion is labeled as either recurrent or stable. The goal is to use deep learning to automate this classification process using a combination of imaging and clinical data. Every lesion in the dataset is identified as an independent sample based on follow-up assessments recorded in the lesion-level clinical Excel sheet. The labels are determined by radiological reports and post-treatment evaluations to find out whether the lesion has remained stable or progressed.

The multimodal input for the model includes the following:

- Structural MRI volumes, which provide full brain structure.
- The radiotherapy dose distributions from RTDOSE DICOMs, which measure the distribution of exposure to radiation that the lesion received.
- RTSTRUCT based segmentation masks that define the lesion area.
- Clinical characteristics include lesion location, treatment fractionation, and time since treatment.

The strong class imbalance makes this task even more challenging: out of the 244 lesions, only 23 (or 9.4%) are recurrent, while the other 221 are stable. Because of such skewed class distributions, a model may be biased to over-predict the majority class, resulting in high accuracy but low sensitivity to recurrence cases. To counteract this, the methodology uses techniques like balanced batch sampling, selective data augmentation, and focal loss, all of which are meant to increase the model's sensitivity without sacrificing specificity.

## 4.2 Input Design

In our approach, each lesion is treated as an individual data sample. The task is to determine whether a lesion is stable or recurrent following Gamma Knife radiosurgery. To build a rich and informative input representation for each lesion, we combined imaging data with clinical context. Each sample consists of three components: a cropped MRI volume, a corresponding RTDose, and a structured clinical feature vector.

- **MRI Lesion Patch (3D Volume)**

  For each lesion, we extract a 3D MRI patch of size 64×64×64 centered on the lesion region. The lesion's location is defined by the contours in the RTSTRUCT DICOM file. We convert these contours into binary masks and compute a bounding box around the lesion. To provide context, we expand the bounding box with additional padding in each dimension. Before cropping, we resample the entire MRI volume to a uniform voxel spacing and normalize the intensity values to the [0, 1] range. This ensures consistent input dimensions and contrast across all patients.

- **RTDose Patch (3D Volume)**

  If RTDose data is available, we extract a matching 64×64×64 dose patch aligned to the same coordinates as the MRI lesion patch. The RTDose files contain 3D grids of radiation dose values, which we resample and align with the MRI space using the metadata stored in the DICOM headers. For lesions without RTDose files, we insert a zero-filled array to maintain a consistent input structure across all samples.

- **Clinical Metadata (1D Feature Vector)**

  Each lesion also includes structured clinical data, which we extract from the Excel sheet in the "lesion_level" tab. we include features such as the number of months between treatment and imaging (duration_tx_to_imag), the number of radiation fractions, and the anatomical location of the lesion. We encode these variables into a fixed-size vector, using one-hot encoding or numerical scaling as needed.

**One-Hot Encoding**

One-hot encoding is a method used to convert categorical variables, values like "Lt Frontal" or "Rt Cerebellar", into a numerical format that can be understood by machine learning models.

Unlike numbers, categorical values do not have mathematical meaning or order. For example, "Frontal" is not greater or smaller than "Parietal", they are just different. But

machine learning models need numerical input, so we convert each possible category into a binary vector, where:

- Each position in the vector represents one possible category.
- A 1 is placed at the position corresponding to the current value.
- All other positions are filled with 0.

Example:

4 possible lesion locations:

- Frontal
- Parietal
- Cerebellar
- Occipital

Then:

- "Frontal" → [1, 0, 0, 0]
- "Cerebellar" → [0, 0, 1, 0]

In the dataset, we extracted lesion location information from the clinical Excel file under the column "Lesion Location". These values were written in text and often had inconsistent abbreviations (e.g., "Lt Cerebellar" or "Left Cereb"). First, we normalized these names using a location mapping dictionary.

Once we had a consistent set of standardized lesion locations (like "Frontal", "Cerebellar", "Parietal"), we applied one-hot encoding to convert these labels into binary vectors.

This allowed each lesion's location to be represented numerically as part of the clinical metadata input to the model. The resulting vector was then concatenated with other clinical features (like duration_tx_to_imag and fractions) and passed through the metadata branch of the model.

By using one-hot encoding, we ensured that:

- The model treated each lesion location as a unique class.
- There was no unintended mathematical relationship between categories.

This encoding step was essential to allow categorical clinical data to be integrated into the neural network alongside 3D imaging features.

Therefore, for each lesion, we construct a sample with the following input format:

- A 3D MRI patch: (64×64×64)
- A 3D RTDose patch: (64×64×64)
- A 1D metadata vector: (fixed length)

To enable effective loading during model training, we save all data in the NumPy.npy format. To ensure that the model learns meaningful lesion-specific patterns rather than patient identity cues, each patient's data is stored independently to avoid patient mixing.

## 4.3 Model Architecture

To classify each brain lesion as either stable or recurrent, we implemented and evaluated three different 3D convolutional neural network (CNN) architectures. These models were designed to progressively increase the complexity and incorporate more types of information, including imaging, dosimetric, and clinical metadata. Each model builds upon the previous, allowing a comparative analysis of how multimodal integration affects performance.

### 4.3.1 First Model – Basic MRI-Only CNN

The first model, which we refer to as the Basic MRI-Only CNN, forms the basis of this study. Its primary purpose is to establish a minimal threshold for classifying recurrences at the lesion level just on anatomical details extracted from structural MRI data. This design intentionally removes other information sources, such as RTDose or clinical metadata, in order to isolate the performance contribution of volumetric imaging alone. The first model trained here:

- Act as a baseline for measuring the effect of increasing model complexity
- Focus exclusively on what can be learned from MRI structure alone
- Serve as a computationally efficient option, especially for ablation studies or when full multimodal data is not available

A 64×64×64×1 single-channel, 3D volumetric MRI patch centered on the lesion serves as the model's input. Lesion masks, that are standardized across patients to a fixed spatial resolution and intensity scale and defined in the RTSTRUCT files, are used to extract these patches. The lesion itself and a margin of surrounding brain tissue help the model to learn contextual cues.

This model uses a compact 3D CNN architecture composed of three convolutional blocks, progressively extracting increasingly abstract spatial features.

The use of Global Average Pooling instead of flattening ensures a compact representation and reduces the number of trainable parameters. This is followed by a dense layer to learn final nonlinear mappings and a dropout layer for regularization. The final neuron produces a scalar output for binary classification.

**Architectural Rationale**

This model is intentionally kept simple by avoiding deeper convolutional stacks, batch normalization, residual connections, or attention mechanisms, thereby reducing computational cost while providing a benchmark.

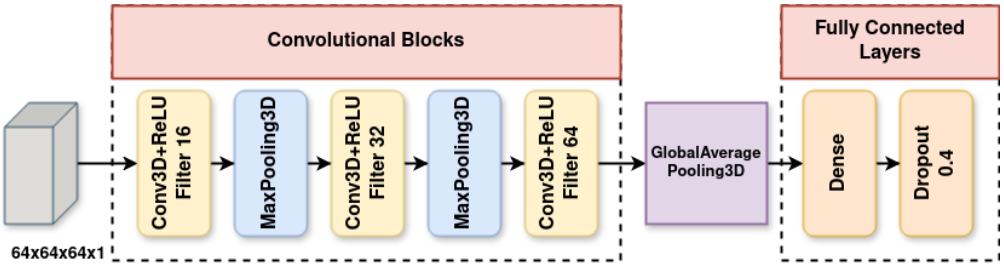



Figure 4.1: Basic MRI-Only CNN

**Training Observations**

During training, this model converged rapidly and demonstrated modest accuracy on the validation set. However, its sensitivity to recurrent lesions was limited. The absence of RTDose and clinical context restricted its capacity to capture treatment-related or patient-specific factors contributing to recurrence. Nonetheless, it learned general structural characteristics of stable vs. recurrent lesions reasonably well.

**Limitations and Role in the Study**

The Basic MRI-Only CNN is crucial because it illustrates the lower performance bound when depending solely on imaging, even though it lacks the representational depth of the later models. It offers a clear, comprehensible starting point for measuring the impact of more complex model elements in later designs.

Though its standalone performance is limited in complex cases involving post-treatment recurrence, this first model confirms that even a compact architecture using only lesion centered MRI patches can extract clinically relevant patterns.

### 4.3.2 Second Model – Deeper MRI CNN with Expanded Dense Layers

The second model builds upon the first by deepening the convolutional architecture and expanding the fully connected layers. Like the first model, it uses only MRI data, but it is designed to better capture complex spatial features by increasing the number of filters in each convolutional layer and introducing additional dense layers for richer decision boundaries.

The input remains a 3D MRI patch of size 64×64×64×1, centered on the lesion and standardized in size and intensity. The extraction process is the same as described for the first model.

This model deepens the network and includes more trainable parameters.

The deeper convolution stack allows the model to extract more sophisticated hierarchical features. The two fully connected layers and added dropout help improve generalization and allow the network to learn more abstract relationships.

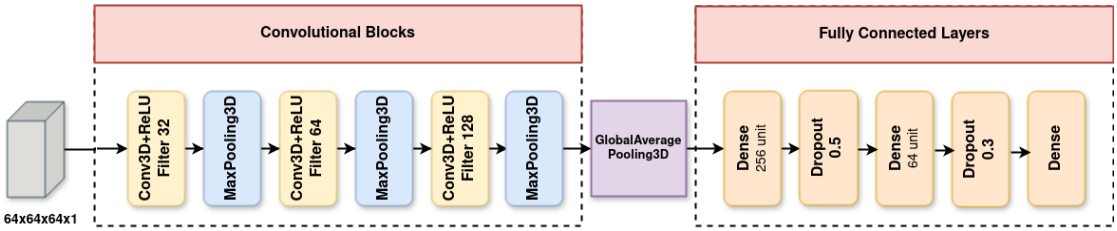



Figure 4.2: Deeper MRI CNN with Expanded Dense Layers

**Architectural Rationale**

This architecture reflects a common strategy in deep learning: increase model capacity and regularization simultaneously. The three convolutional layers with increasing filters allow deeper spatial representation. The larger dense layer (256 units) offers a wider space for decision logic, while dropout helps preventing overfitting.

**Training Observations**

This model demonstrated improved learning over the first model, with better validation performance and increased sensitivity to recurrent lesions. However, the lack of RTDose or clinical metadata means it still has a limited view of treatment context. It benefits from its depth and more sophisticated structure, but remains restricted by its unimodal design.

**Limitations and Role in the Study**

Although this model offers higher performance than the first, it still lacks the multi-dimensional context of later models. It serves as a midpoint in architectural complexity and is useful for assessing the impact of additional depth and larger dense layers in uni-modal (MRI-only) classification.

In conclusion, by increasing feature extraction depth and decision-layer capacity, the Deeper MRI CNN surpasses the baseline and becomes a more powerful model for learning recurrence-related imaging features from MRI alone.

### 4.3.3 Third Model – 3D Multi-Input Network

The third and final model, referred to as the Multi-Input Fusion Network, is the most advanced architecture in this study. It is designed to fully exploit the richness of the dataset by combining structural imaging (MRI), treatment data (RTDose), and clinical features. The goal of this architecture is to replicate how clinicians make informed decisions based on multiple complementary sources of data.

Three parallel processing branches are used in this model to separately extract high-level features from each input modality. To generate a final prediction, these features are then combined and run through a number of fully connected layers. In the final classification

step, this structure permits cross-modal interactions while allowing the model to learn and retain modality-specific information.

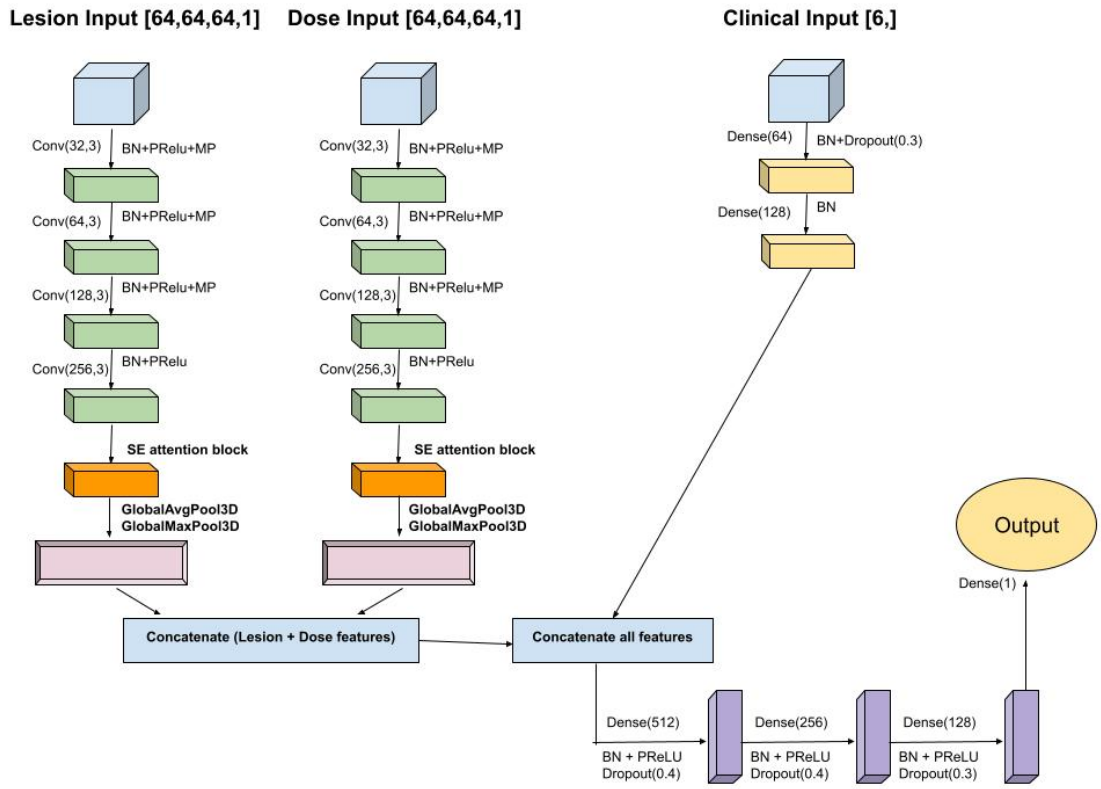



Figure 4.3: Deeper MRI CNN with Expanded Dense Layers

The model accepts three distinct inputs, each corresponding to a different data modality:

- **MRI Lesion Patch:** A 3D volume of shape 64×64×64×1 representing the lesion-centered MRI scan. This volume contains the lesion and surrounding tissue and provides anatomical context.
- **RTDose Patch:** A 3D volume of the same shape (64×64×64×1) representing the spatial radiation dose distribution aligned with the MRI volume. In cases where the RTDose is not available, a zero-filled volume is substituted to maintain consistency in input shape.
- **Clinical Metadata Vector:** A 6-dimensional feature vector that includes categorical and numerical data, including the number of dose fractions, the one-hot encoded lesion location, and the time between treatment and follow-up.

Each input is processed through a specialized sub-network tailored to its data type and then followed by a late fusion mechanism and a deep classifier head.

Going into details of each branches, there are:

**1. MRI and RTDose Branches with SE-Attention**

Both the MRI lesion patch and the aligned RTDose patch are processed by separate but architecturally identical 3D convolutional subnetworks. Each branch consists of four stacked Conv3D→BatchNorm→PReLU blocks, interleaved with MaxPooling3D layers to progressively reduce spatial dimensions and increase representational capacity. After the final 256-channel convolutional block (Conv3D(256,3)→BN→PReLU), we insert a Squeeze-and-Excitation (SE) attention[44] module to enable channel-wise feature reweighting:

- **Squeeze**
  - A GlobalAveragePooling3D operation collapses each of the 256 feature-maps to a single scalar, yielding a 256-dimensional descriptor that summarizes the global response of each channel.
- **Excitation**
  - This descriptor passes through a two-layer fully connected "bottleneck" MLP (first to 256/4 units with ReLU, then back to 256 units with a sigmoid activation), producing one weight per channel in the range (0,1).
- **Reweight**
  - The 256-vector of weights is reshaped to (1×1×1×256) and multiplied element-wise into the original 3D feature-maps. Channels deemed more informative for distinguishing stable vs. recurrent lesions are thus amplified, while less useful channels are suppressed.

After reweighting, we apply dual global pooling (GlobalAveragePooling3D + GlobalMaxPooling3D) to produce a compact 512-dimensional vector for each branch. These attention-augmented branch features, one from MRI, one from RTDose are then concatenated (along with the clinical branch) and passed to the fusion classifier head.

**2. Clinical Metadata Branch**

The clinical vector, representing structured patient and treatment data, is passed through a fully connected subnetwork designed to project low-dimensional metadata into a higher-level embedding space.

This pathway Dense, BatchNormalization, Dropout helps the model encode non-imaging data and capture patient-level risk factors that are not visually observable in the scans.

**3. Feature Fusion and Classifier Head**

The outputs of the three branches (MRI, RTDose, Clinical Metadata) are concatenated to form a unified feature vector. This vector is then passed through a deep classifier.

This structure enables the model to learn complex interdependencies between image features, dose patterns, and clinical characteristics. The PReLU activations improve learning flexibility over standard ReLU, especially in deep networks.

**Architectural Rationale**

Several important principles guided the design of this model:

- **Multimodal Fusion:** Gathers supplementary data from RTDose (effect of treatment), MRI (anatomical context), and metadata (specific context for a patient).
- **Independent Subnetworks:** To avoid cross-modal interference and to enable specialization, each modality is handled by a separate subnetwork.
- **Hierarchical Feature Learning:** Every branch can identify both local and global patterns in the input volumes thanks to the deep convolutional stacks.
- **Robust Aggregation:** Dual global pooling (avg + max) enhances feature robustness and generalizability.
- **Dense Classifier:** Following combination, the fully connected layers allow the model to identify high-level patterns for ultimate classification.

**Training Observations**

This model demonstrated the highest classification performance across all metrics, particularly in correctly identifying recurrent lesions, which are often underrepresented in the dataset. It benefited from having access to multiple modalities and showed better generalization on the test set.

- The fusion of RTDose improved sensitivity to recurrence.
- Clinical features helped the model capture individual treatment histories and biological variations.

**Limitations and Role in the Study**

While this model achieved the best results, it is also the most computationally expensive and depends on the availability of all three input modalities. In clinical scenarios where RTDose or metadata are missing, its full potential may not be realized.

Nonetheless, this architecture best reflects real-world decision-making, where radiologists and oncologists consider imaging, treatment, and patient-specific context simultaneously.

### 4.3.4 Focal Loss

To handle the pronounced class imbalance present in the dataset, we adopted the Focal Loss function, originally introduced by Lin et al. in their work on dense object detection (RetinaNet) [28]. While their approach was developed for object detection in computer vision, the underlying motivation applies directly to medical classification tasks: when one class is overwhelmingly more frequent than the other, standard loss functions (such as binary cross-entropy) tend to bias model optimization toward the majority class. This bias results in poor sensitivity for the minority class in this case, recurrent brain lesions which are precisely the instances of greatest clinical interest.

Focal Loss modifies the standard cross entropy by introducing a modulating factor that down weights the loss contribution from well classified examples and focuses learning on hard, misclassified examples. The mathematical formulation is:

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

Where:

- $p_t$ is the predicted probability of the true class label,
- $\alpha_t$ is a weighting factor that adjusts for class imbalance,
- $\gamma$ is a focusing parameter that adjusts the rate at which easy examples are down-weighted[45].

As $\gamma$ increases, the loss focuses more on misclassified examples. Lin et al.[28] demonstrated that this modification significantly improves detection performance in scenarios with high imbalance, and their results strongly motivated its use in this project.

According to lesion classification, 221 of the 244 lesions in the dataset are classified as stable, and only 23 are classified as recurrent. Because of this huge imbalance, focus loss was used to direct the model away from the obvious solution of consistently predicting the majority class.

In my training pipeline, we implemented Focal Loss as a custom Keras-compatible loss class. we used the following parameters:

- $\gamma = 2.0$: to focus the model more heavily on misclassified or ambiguous examples
- $\alpha = 0.5$: to increase the weight of the minority (recurrent) class

This focal loss replaced binary cross-entropy and was applied consistently across all models. It helped improve the model's ability to detect recurrence, leading to significantly higher recall and F1-scores on recurrent cases. This was especially important for the third model, where focal loss worked in synergy with the multi-input architecture to boost sensitivity.

## 4.4 Training Setup

To train and evaluate the proposed models for lesion-level recurrence classification, we implemented a carefully designed training pipeline with specific strategies to handle class imbalance, improve generalization, and ensure reproducibility. This section outlines the key components of the training procedure, including optimizer settings, data balancing strategy, and regularization methods.

### 4.4.1 Optimizer and Learning Strategy

All three models were trained using the AdamW optimizer, a variant of the standard Adam optimizer that incorporates decoupled weight decay for better regularization. AdamW

has been shown to improve generalization in deep networks, especially when training with limited data.

The specific hyperparameters used were:

- Learning rate: $1 \times 10^{-4}$
- Weight decay: 0.01
- Gradient clipping: applied with a maximum norm of 1.0 to prevent gradient explosion in deep layers.

To help the model settle into a good minimum, we watch the validation loss during training and cut the learning rate in half whenever it hasn't improved for seven straight epochs. Once it drops below $1 \times 10^{-6}$ we hold it there so gradients don't vanish. Although we allow up to 100 epochs, we also use early stopping: if validation performance stalls for a predefined stretch, training halts and the best weights are restored.

**Early Stopping**

we employed early stopping during training to avoid overfitting and pointless computation. When the model no longer improves, this method stops training and tracks the validation loss over epochs. In particular, if the validation loss did not decrease for 15 consecutive epochs, training was terminated. In order to guarantee that the final model reflected the best-performing state on the validation set, the model weights were additionally restored to the epoch with the lowest recorded validation loss.

This strategy helped prevent overtraining, especially on deeper architectures or when using augmented recurrent lesions, where the risk of memorizing patterns is higher. By preserving the best model checkpoint and avoiding wasted epochs, early stopping also contributed to more efficient and stable training.

### 4.4.2 Balanced Sampling

Due to the highly imbalanced nature of the dataset where only 9% of lesions are recurrent, we implemented a balanced sampling strategy to ensure that each training batch included a representative mix of classes.

Specifically:

- Each batch was constructed to include 50% stable and 50% recurrent lesions.
- Since stable lesions greatly outnumber recurrent ones, this was achieved by:
  - Using each stable lesion at most once per epoch.
  - Generating multiple augmented variants of recurrent lesions to match the batch quota.

This approach allowed the model to see both classes equally often during training, avoiding bias toward the dominant class and improving sensitivity to recurrence.

### 4.4.3 Regularization and Logging

To improve generalization and prevent overfitting, particularly given the limited sample size and high class imbalance, we applied a set of regularization techniques across all models. These include **Dropout**, **Batch Normalization**, Gradient Clipping, and Early Stopping. We also used Weights & Biases (**WandB**) for logging and model checkpointing to retain the best-performing weights during training.

#### Dropout

Dropout is a regularization technique where, during training, a randomly selected fraction of neurons is temporarily deactivated or "dropped" from the network. This forces the model to not rely too heavily on any particular neuron and encourages the development of more robust, distributed feature representations.

we applied dropout in the fully connected layers of each model. The dropout rate varied between 0.3 and 0.5, depending on the size and depth of the layer:

- For smaller dense layers, we used 0.3.
- For larger ones, such as the 512 unit dense layer in the third model, we used 0.5.

This helped prevent the model from overfitting to specific patterns in the training data, particularly important given the small number of recurrence samples.

#### Batch Normalization

The technique known as batch normalization (BN) normalizes the inputs to a layer so that, over each mini-batch, their mean is zero and their standard deviation is one. This lessens internal covariate shift, which can lead to instability and slow learning by altering layer input distributions during training.

we positioned BN before the activation function and after each Conv3D and Dense layer in my models. This was beneficial:

- Stabilize training, allowing for faster convergence.
- Reduce dependence on careful weight initialization.
- Serve as a form of regularization, because the use of batch statistics introduces slight noise that prevents overfitting.

#### Gradient Clipping

Exploding gradients, where large updates destabilize learning, can affect deep learning models with a lot of parameters, particularly 3D CNNs. we used gradient clipping, which restricts the gradients' magnitude during backpropagation, to fix this.

In order to prevent excessive shifts in weight values from any one gradient update, we clipped the gradient norm to a maximum of 1.0. Training became more stable as a result, particularly in deeper networks' later phases.

### Logging with Weights & Biases (WandB)

Weights & Biases (WandB) served as our experiment tracking and logging platform for the duration of this project in order to track model performance, visualize training dynamics, and make debugging easier. An interactive tool called WandB was created to oversee machine learning research. Its smooth integration with TensorFlow and Keras enables the visualization of important metrics in real time and offers a centralized dashboard for comparing various training runs.

For each training session, we configured WandB to automatically log and visualize a wide range of relevant information, including:

- **Training and validation loss curves**
  logged the model's training and validation loss at the end of every epoch. This allowed us to observe how well the model was fitting the data and whether overfitting was occurring. For example, a consistent decrease in training loss but a plateau or increase in validation loss is a strong indicator of overfitting, something we could catch and act on quickly.
- **Classification metrics**
  At each training epoch, we tracked key metrics: accuracy, precision, recall, F1-score and AUC, to see not only how the model performed overall but also how well it detected the rare yet clinically crucial recurrent lesions. Because the data were imbalanced, we paid special attention to recall and F1-score for those recurrent cases; accuracy by itself really didn't capture the model's true effectiveness.
- **Learning rate behavior**
  As we used a dynamic learning rate schedule (ReduceLROnPlateau), WandB allowed us to track how the learning rate evolved over time. This was especially helpful to confirm that the scheduler was triggering as expected when validation loss plateaued, and to evaluate how learning rate reductions impacted convergence.
- **Hardware usage and system metrics**
  WandB also tracked GPU memory consumption, CPU usage, and epoch runtime duration. This was helpful for identifying performance bottlenecks, optimizing batch sizes, and determining which models were more computationally efficient, important considerations when working with 3D volumetric data, which is memory intensive.

WandB made it simple for us to share training results, examine hyperparameter selections, and compare various model architectures side by side. we could precisely monitor the effects of modifications, like deleting metadata or switching to a different loss function, on performance, which was helpful during ablation studies.

WandB ensured complete reproducibility by acting as a versioned experiment history in addition to visualization. A fixed code snapshot, model configuration, and dataset version were linked to each training run. This made it possible for me to go back to any earlier run, replicate the outcomes precisely, and improve upon them.

In conclusion, WandB was a crucial component of my entire deep learning workflow since

it was necessary for robust experiment tracking, debugging, and model interpretability in addition to real-time monitoring.

In addition to tracking training metrics, we kept an eye on GPU resource usage, such as memory and temperature, to make sure training held steady over time. It is computationally demanding to train 3D convolutional neural networks on volumetric medical data, and if extended training sessions are not closely monitored, they may cause hardware strain and throttling or performance degradation.

we recorded the GPU temperature during several training runs of the third model to ensure that hardware performance stayed constant. All of the experiments' temperature variations fell within the typical operating range, which is 45°C to 60°C.

Monitoring GPU temperature helped us ensure that:

- All training runs were executed under stable thermal conditions, without throttling or GPU, induced slowdowns.
- Hardware performance was not a confounding variable in comparing results across models and runs.

This level of resource monitoring supports the reproducibility and reliability of the training process and confirms that performance differences were due to model behavior rather than hardware inconsistency.

To evaluate how well the model balanced sensitivity and precision, particularly for detecting recurrent lesions. we tracked the F1-score during training. F1-score is the harmonic mean of precision and recall, and is especially important in settings where both false positives and false negatives carry clinical significance.

## 4.5 Evaluation Metrics

To evaluate the performance of the trained models, we used a set of classification metrics that provide insight into both overall accuracy and class specific behavior. Given the dataset's severe class imbalance and the clinical we importance of correctly identifying recurrent lesions, we focused on metrics that go beyond simple accuracy. This section explains the core metrics we used: accuracy, sensitivity (recall), specificity, and F1-score

### 4.5.1 Accuracy

Accuracy is the proportion of total correct predictions made by the model. It is calculated by summing all correctly classified samples (true positives and true negatives) and dividing by the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Where:

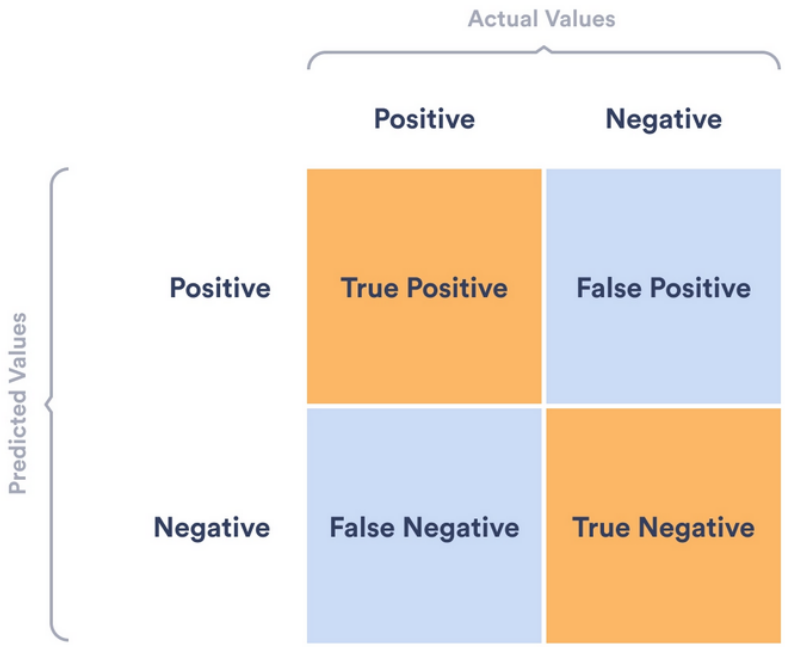



Figure 4.4: Confusion Matrix[3]

- **TP (True Positive):** Model correctly predicts recurrence.
- **TN (True Negative):** Model correctly predicts stability.
- **FP (False Positive):** Model wrongly predicts recurrence.
- **FN (False Negative):** Model wrongly predicts stability.

Accuracy gives a general sense of overall performance.

In imbalanced datasets like ours (with 90% stable lesions), a model could predict "stable" for all cases and still achieve high accuracy, despite completely failing to detect any recurrence.

### 4.5.2 Sensitivity (Recall)

Also known as recall, sensitivity measures how well the model identifies actual positive cases(recurrent lesions).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.2}$$

It can be clinically frustrating to miss a recurrence in a medical diagnosis (a false negative), which delays treatment or follow up. A high recall guarantees that the majority of real

recurrences are reported.

we used balanced batch sampling and focal loss to give sensitivity top priority. When compared to the other two, my best model (the third architecture) had the highest recall.

### 4.5.3 Specificity

Specificity evaluates how well the model avoids false alarms, it measures how accurately stable (non-recurrent) lesions are identified.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.3}$$

Many stable lesions will be incorrectly classified as recurrent by a model with low specificity, resulting in needless scans, worry, and follow-up procedures. On the other hand, recurrence alerts with high specificity are guaranteed to be significant and clinically actionable.

we were able to determine whether my model was calling recurrence too aggressively thanks to specificity, particularly when optimizing for higher recall.

### 4.5.4 F1-Score

The F1-score balances precision and recall. It's especially useful when both false positives and false negatives are harmful, and the dataset is imbalanced.

$$\text{F1-Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.4}$$

Where:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.6}$$

The trade-off between completely missing recurrence (low recall) and overpredicting it (low precision) is captured by the F1-score. Models that excel at one but not the other are penalized.

Particularly during validation and model selection, the F1-score served as the main evaluation metric. Compared to accuracy alone, it provided a more significant performance metric.

In the context of this project, classifying brain lesions as stable or recurrent, these evaluation metrics were selected to reflect the clinical relevance and practical challenges of the task. While accuracy provides a general measure of correctness, it is insufficient on its own

due to the severe class imbalance in the dataset. Therefore, we placed greater emphasis on sensitivity, specificity, and F1-score, which offer a more nuanced view of model behavior.

Sensitivity (recall) is especially critical, as missing a recurrent lesion (false negative) could delay necessary treatment and compromise patient outcomes. At the same time, specificity is important to avoid falsely labeling stable lesions as recurrent, which could lead to unnecessary anxiety, follow-up imaging, and intervention. The F1-score captures the balance between these two concerns, providing a single, interpretable metric that reflects the model's ability to detect recurrence without overwhelming the system with false alarms.

By combining these metrics, we were able to evaluate each model not only by its raw performance but by how well it would function in a real clinical scenario, where both correct detection and false alarms carry meaningful consequences.

# Chapter 5

# Results and Conclusion

## 5.1 Experimental Setup

This section outlines the computational environment used to develop, train, and evaluate the deep learning models presented in this study. Given the volumetric nature of the input data (3D MRI patches) and the complexity of the model architectures, particularly the multi-input model integrating imaging, dose distribution, and clinical metadata, it was essential to conduct experiments on a system capable of handling high memory demands and prolonged training sessions. Below, we describe the hardware configuration, software environment, training settings, and reproducibility strategy.

### 5.1.1 Hardware Configuration

All experiments were conducted on a high-performance local workstation optimized for deep learning workloads. The hardware setup included:

- **Primary GPU:** NVIDIA GeForce RTX 5090 with 32 GB of dedicated VRAM. This GPU was used to perform all training and inference tasks, particularly for the third model which required extensive memory due to multiple input branches and large kernel stacks.
- **Secondary GPU:** NVIDIA GeForce RTX 3060 with 12 GB VRAM. This GPU was used for auxiliary tasks such as preprocessing, data loading, and parallel evaluation.
- **CPU**: AMD Ryzen 9 5950X (16 cores / 32 threads). This processor provided substantial parallel processing power, facilitating fast data preprocessing and multi-threaded execution for batch preparation.
- **RAM:** 128 GB of DDR4 memory. we were able to load and process several 3D volumes in memory without running into paging delays or I/O bottlenecks thanks to the large memory capacity.

- **Storage:** 2 TB NVMe SSD. High-speed solid-state storage significantly reduced data loading times for volumetric MRI and RTDose files, which are large and I/O-intensive.
- **Operating System:** All experiments were run inside GPU, passthrough containers managed by Proxmox (64-bit). The use of virtualized containers allowed clean environment isolation, system monitoring, and reproducibility.

With this setup, we could run multimodal deep learning models, manage massive amounts of 3D data, and keep an eye on performance even during lengthy training sessions.

### 5.1.2 Software Environment and Training Configuration

All deep learning models were implemented in Python 3.10 using the TensorFlow 2.13 framework with GPU support via CUDA 11.8 and cuDNN 8.6. The environment was managed using conda, ensuring package consistency and isolated dependencies across experiments.

The primary libraries and tools used include:

- **TensorFlow:** for model building, training, and inference.
- **NumPy and pandas:** for data processing and handling clinical tabular data.
- **Scikit-learn:** for evaluation metrics, confusion matrices, and statistical tools.
- **SimpleITK, pydicom, and nibabel:** for reading and preprocessing DICOM and NRRD files.
- **Matplotlib and seaborn:** for generating plots and visual summaries.
- **Weights & Biases (wandb):** for experiment tracking, metric logging, and model checkpointing.

## 5.2 Performance Evaluation

In this section, we present the quantitative evaluation of my final deep learning model, a multi-input 3D convolutional neural network (CNN) that incorporates imaging (MRI), dosimetric (RTDOSE), and clinical information to classify brain lesions as either stable or recurrent. The primary aim of this evaluation is not only to demonstrate the predictive capacity of the model but also to analyze its behavior under the constraints of significant class imbalance.

Out of 244 lesions, only 23 (9.4%) were labeled as recurrent, whereas 221 were labeled as stable. This skew introduces a strong bias toward the majority class in typical learning algorithms, which, if not explicitly addressed, can lead to misleadingly high accuracy but poor clinical utility.

**Overview of Test Results**

On the test set, the final model yielded the following results:

- **Accuracy: 75.6%**
- **Sensitivity (Recall) – Recurrent: 50.0%**
- **Specificity – Stable: 78.4%**
- **F1-score – Recurrent: 28.6%**

**Metric Interpretations**

- Despite its apparent strength, accuracy is insufficient on its own because of the imbalance in class labels. By consistently predicting "stable," a naive model could attain over 90% accuracy, but this performance would have no clinical significance.
- In this situation, sensitivity (recall) is especially crucial. A 50% recall for the recurrent class indicates that half of the recurrence cases were detected by the model, which is a notable improvement over majority class prediction or random guessing.
- For recurrence, the F1-score, which balances recall and precision, is 28.6%. The impact of false positives and the small number of recurrence cases account for this comparatively low value. It is noticeably higher than the benchmark model's F1-score, though.
- Specificity of 78.4% shows that the model was also reasonably good at correctly identifying stable lesions, which adds confidence in its reliability for negative predictions.

These findings point to a model that, even in the presence of highly skewed data, actually learns to differentiate between stable and recurrent lesions rather than just falling back to the majority class.

**Comparison with Published Benchmark**

To contextualize the performance of my model, I compared it with the benchmark proposed by Wang et al. (2023), who first published the Brain-TR-GammaKnife dataset. Their final model was a multi-input 3D convolutional neural network that incorporated three data streams: MRI volumes, RTDOSE maps, and a small set of clinical metadata (diagnosis, age, gender). All three inputs were processed through independent branches and then concatenated before prediction, as illustrated in their published architecture(Figure 5.2).

Although they used data augmentation during training and testing, their pipeline lacked a dedicated validation set. This meant that important hyperparameters, including the decision threshold, were not optimized with respect to generalization, and early stopping or calibration techniques were not applied. This potentially contributed to their model's overfitting on the training distribution, and underperformance on rare recurrence cases.

Their reported test performance was:

- **Accuracy: 90.1%**
- **Sensitivity (Recall) – Recurrent: 10.0%**

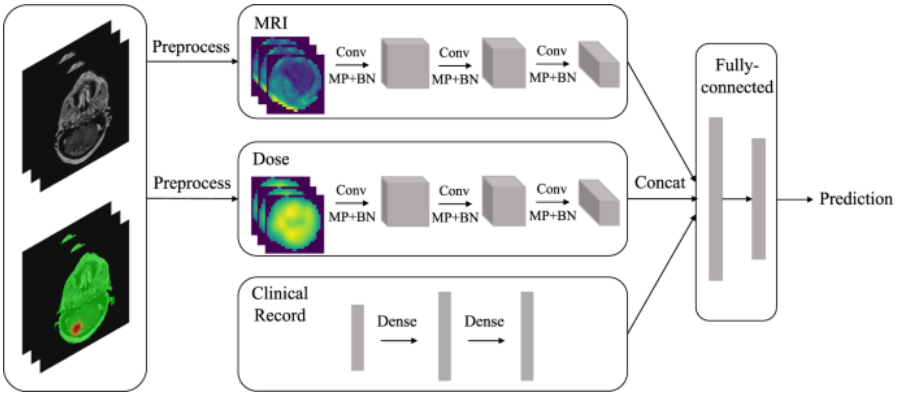



Figure 5.1: model[4]

- **Specificity – Stable: 89.0%**
- **F1-score – Recurrent: 18.2%**

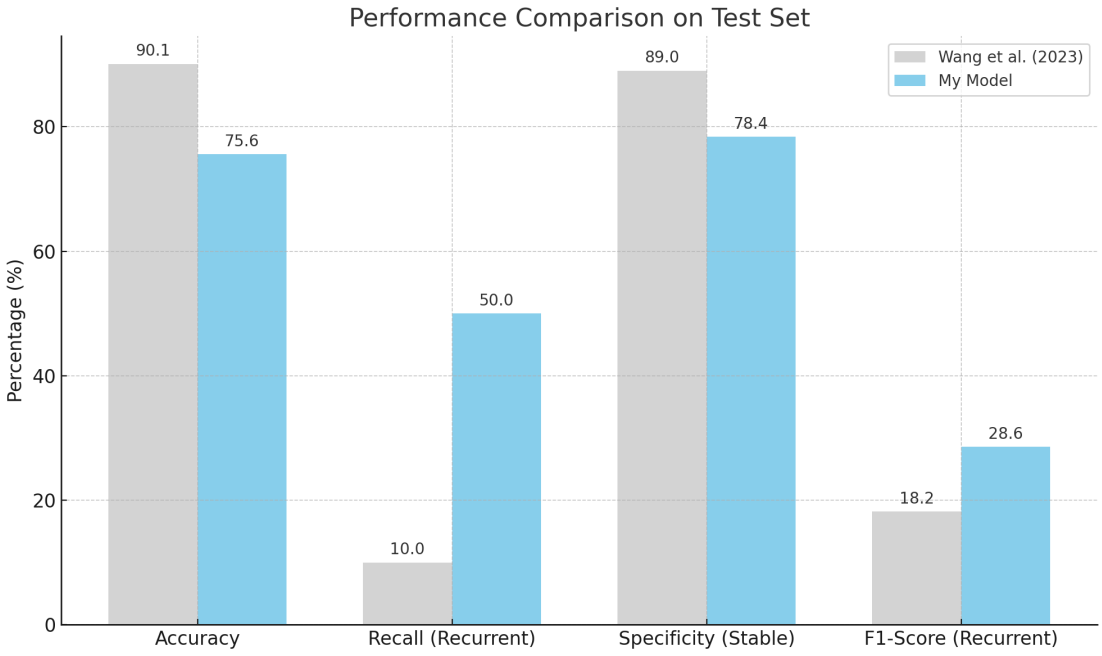



Figure 5.2: Performance Comparison on Test Set

While these numbers indicate strong performance on stable cases, their model failed to generalize to recurrence, correctly identifying only 1 in 10 recurrence lesions. This is a clear symptom of class imbalance, which was not explicitly addressed through loss functions or batch balancing.

As a result,shown in Figure 5.2, our model achieved a **fivefold improvement in recurrence recall** (50% vs. 10%) and a substantially better **F1-score** (28.6% vs. 18.2%), while maintaining clinically acceptable accuracy (75.6%) and specificity (78.4%).

## 5.3 Conclusion

This section reviews how the model performed and provides a deeper look into the results beyond just the numbers. It explores what the model did well, where it faced difficulties, and how those outcomes relate to the challenges of applying deep learning in real clinical situations. Predicting whether a brain lesion is stable or recurrent after treatment is a complex task, not only because of the limited number of recurrence cases in the dataset, but also because of the clinical uncertainty that often surrounds recurrence itself. For this reason, the discussion doesn't only focus on metrics like accuracy or recall, but also considers how the model's behavior might support or complicate clinical decisions. In addition to evaluating performance, this section discusses the clinical relevance of the model's predictions, highlights technical limitations, and suggests directions for future improvement.

### 5.3.1 Model Strengths

The proposed model demonstrated several significant benefits, especially in addressing the common class imbalance problem that has historically hindered recurrenc detection. The model achieved 50% recall for recurrent lesions, compared to 10% for the baseline model, by using batch-level balancing, selective data augmentation, and focal loss. This fivefold improvement is particularly significant given that recurrent lesions account for less than 10% of the entire dataset.

Furthermore, the model used a multi-input architecture to concurrently integrate structured clinical features, RTDOSE, and MRI. Because of this design, each modality was able to be processed separately and benefit from its distinct contributions: clinical metadata provided patient-level prognostic information, RTDOSE added treatment distribution context, and MRI provided structural insight.

Moreover, by increasing the spatial input resolution to 64×64×64, the model preserved more lesion detail, which may have contributed to better learning of recurrence patterns , particularly subtle ones that could be lost at lower resolutions (40×40×40, as used in Wang et al.).

### 5.3.2 Model Challenges and Weaknesses

Despite these advancements, the model still faced important limitations. While recall improved, the precision for recurrence remained low, suggesting that the model also flagged a notable number of stable lesions as potentially recurrent (false positives). This result reflects a common trade-off: in optimizing for sensitivity, specificity and precision may

be sacrificed. In the clinical setting, this could lead to unnecessary concern or follow-up testing, especially if used as a stand-alone diagnostic tool.

The inconsistent lesion labeling across various data sources presented another difficulty. Mismatches between RTSTRUCT ROI names and clinical annotations remained even after a location mapping scheme was used to standardize terminology. These discrepancies might have affected the accuracy of training and evaluation by adding noise to the label generation process.

Furthermore, while the multi-input architecture was effective, the clinical branch used relatively simple dense layers, which may not have fully captured complex interactions between clinical features. A more advanced representation (e.g., using attention or learned embeddings) might further boost performance, especially when clinical data is sparse or heterogeneous.

## 5.4 Future Work

While the model developed in this thesis achieved meaningful results, particularly in detecting recurrent brain lesions, several areas remain open for further improvement and exploration.

One key direction is the integration of temporal information. Currently, each lesion is treated as an independent static sample, even when the patient has multiple follow-up visits. Incorporating temporal context, such as changes in size, shape, dose overlap, or imaging intensity across time, could help distinguish between stable post-treatment changes and early signs of recurrence. Time-aware architectures, such as recurrent neural networks or attention-based models, could be explored for this purpose.

Another opportunity lies in enhancing the clinical metadata pathway. In the current implementation, clinical variables are processed through basic dense layers. This approach may not capture more nuanced relationships between features such as treatment course, diagnosis timing, or patient demographics. Future versions of the model could apply graph-based representations, embeddings, or attention mechanisms to learn deeper patterns from clinical context.

Additionally, the model's interpretability remains an essential area for development. While the model demonstrates useful performance, it operates as a black box. Implementing visualization techniques such as Grad-CAM, SHAP, or saliency maps would allow researchers and clinicians to see which regions or features contributed most to a prediction, enhancing both scientific understanding and clinical trust.

A further extension would be to perform external validation on independent datasets. While this thesis focused exclusively on the Brain-TR-GammaKnife dataset, testing the model on additional patients or multi-center data would be necessary to assess generalizability and robustness. Furthermore, clinical deployment would require regulatory validation, usability studies, and integration with hospital systems.

Finally, exploring multi-task learning, for example, jointly predicting recurrence status and

lesion growth, could strengthen the model's feature representations and improve overall performance.

# Bibliography


[1] K. R. Fink and J. R. Fink, "Imaging of brain metastases," *Surgical neurology international*, vol. 4, no. Suppl 4, p. S209, 2013.

[2] J. Cho, Y. J. Kim, L. Sunwoo, G. P. Lee, T. Q. Nguyen, S. J. Cho, S. H. Baik, Y. J. Bae, B. S. Choi, C. Jung *et al.*, "Deep learning-based computer-aided detection system for automated treatment response assessment of brain metastases on 3d mri," *Frontiers in Oncology*, vol. 11, p. 739639, 2021.

[3] L. Seraydarian, "What is a confusion matrix and how to read it?"

[4] Y. Wang, W. N. Duggar, D. M. Caballero, T. V. Thomas, N. Adari, E. K. Mundra, and H. Wang, "A brain mri dataset and baseline evaluations for tumor recurrence prediction after gamma knife radiotherapy," *Scientific Data*, vol. 10, no. 1, p. 785, 2023.

[5] H. Mehrabian, J. Detsky, H. Soliman, A. Sahgal, and G. J. Stanisz, "Advanced magnetic resonance imaging techniques in management of brain metastases," *Frontiers in oncology*, vol. 9, p. 440, 2019.

[6] V. Jairam, V. L. Chiang, J. B. Yu, and J. P. Knisely, "Role of stereotactic radiosurgery in patients with more than four brain metastases," *CNS oncology*, vol. 2, no. 2, pp. 181–193, 2013.

[7] S. Ghaderi, S. Mohammadi, M. Mohammadi, Z. N. A. Pashaki, M. Heidari, R. Khatyal, and R. Zafari, "A systematic review of brain metastases from lung cancer using magnetic resonance neuroimaging: clinical and technical aspects," *Journal of Medical Radiation Sciences*, vol. 71, no. 2, pp. 269–289, 2024.

[8] R. Soffietti, R. Rudā, and R. Mutani, "Management of brain metastases," *Journal of neurology*, vol. 249, pp. 1357–1369, 2002.

[9] J. Y. Delattre, G. Krol, H. T. Thaler, and J. B. Posner, "Distribution of brain metastases," *Archives of neurology*, vol. 45, no. 7, pp. 741–744, 1988.

[10] H. Soliman, S. Das, D. A. Larson, and A. Sahgal, "Stereotactic radiosurgery (srs) in the modern management of patients with brain metastases," *Oncotarget*, vol. 7, no. 11, p. 12318, 2016.

[11] P. Reinhardt, U. Ahmadli, E. Uysal, B. K. Shrestha, P. Schucht, A. Hakim, and E. Ermiş, "Single versus multiple fraction stereotactic radiosurgery for medium-sized brain metastases (4-14 cc in volume): reducing or fractionating the radiosurgery dose?" *Frontiers in Oncology*, vol. 14, p. 1333245, 2024.

[12] F. Guo, "3-d treatment planning system—leksell gamma knife treatment planning system," *Medical Dosimetry*, vol. 43, no. 2, pp. 177–183, 2018.

[13] R. Mazzola, S. Corradini, F. Gregucci, V. Figlia, A. Fiorentino, and F. Alongi, "Role of radiosurgery/stereotactic radiotherapy in oligometastatic disease: brain oligometastases," *Frontiers in Oncology*, vol. 9, p. 206, 2019.

[14] Z. Zhang, J. Yang, A. Ho, W. Jiang, J. Logan, X. Wang, P. D. Brown, S. L. McGovern, N. Guha-Thakurta, S. D. Ferguson *et al.*, "A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from mr images," *European radiology*, vol. 28, pp. 2255–2263, 2018.

[15] L. Peng, V. Parekh, P. Huang, D. D. Lin, K. Sheikh, B. Baker, T. Kirschbaum, F. Silvestri, J. Son, A. Robinson *et al.*, "Distinguishing true progression from radionecrosis after stereotactic radiation therapy for brain metastases with machine learning and radiomics," *International Journal of Radiation Oncology* Biology* Physics*, vol. 102, no. 4, pp. 1236–1243, 2018.

[16] S. Mohammadi and S. Ghaderi, "The role of mri in detecting and characterizing brain metastases from breast cancer," *Journal of Cellular and Molecular Immunology*, vol. 2, no. 1, pp. 25–26, 2023.

[17] Q. Fu, Q.-G. Cheng, X.-C. Kong, D.-X. Liu, Y.-H. Guo, J. Grinstead, X.-Y. Zhang, Z.-Q. Lei, and C.-S. Zheng, "Comparison of contrast-enhanced t1-weighted imaging using dante-space, petra, and mprage: a clinical evaluation of brain tumors at 3 tesla," *Quantitative Imaging in Medicine and Surgery*, vol. 12, no. 1, p. 592, 2022.

[18] H. Baek, Y. J. Heo, D. Kim, S. Yun, J. Baek, H. Jeong, H. Choo, J. Lee, and S.-I. Oh, "Usefulness of wave-caipi for postcontrast 3d t1-space in the evaluation of brain metastases," *American Journal of Neuroradiology*, vol. 43, no. 6, pp. 857–863, 2022.

[19] S. Ahn, H. Kwon, J. Yang, M. Park, Y. Cha, S. Suh, and J. Lee, "Contrast-enhanced t1-weighted image radiomics of brain metastases may predict egfr mutation status in primary lung cancer. sci rep. 2020; 10 (1): 8905," *European Journal of Radiology*, vol. 155, p. 110499, 2022.

[20] J. R. F. Kathleen R. Fink, "Imaging of brain metastases."

[21] B. Jeong, D. S. Choi, H. S. Shin, H. Y. Choi, M. J. Park, K. N. Jeon, J. B. Na, and S. H. Chung, "T1-weighted flair mr imaging for the evaluation of enhancing brain tumors: comparison with spin echo imaging," *Investigative Magnetic Resonance Imaging*, vol. 18, no. 2, pp. 151–156, 2014.

[22] N. Tomura, K. Narita, S. Takahashi, T. Otani, I. Sakuma, K. Yasuda, T. Nishii, and J. Watarai, "Contrast-enhanced multi-shot echo-planar flair in the depiction of

metastatic tumors of the brain: comparison with contrast-enhanced spin-echo t1-weighted imaging," *Acta Radiologica*, vol. 48, no. 9, pp. 1032–1037, 2007.

[23] O. L. Wong, J. Yuan, D. M. Poon, S. T. Chiu, B. Yang, G. Chiu, S. K. Yu, and K. Y. Cheung, "Prostate diffusion-weighted imaging (dwi) in mr-guided radiotherapy: reproducibility assessment on 1.5 t mr-linac and 1.5 t mr-simulator," *Magnetic Resonance Imaging*, vol. 111, pp. 47–56, 2024.

[24] S. S. Mannam, C. D. Nwagwu, C. Sumner, B. D. Weinberg, and K. B. Hoang, "Perfusion-weighted imaging: The use of a novel perfusion scoring criteria to improve the assessment of brain tumor recurrence versus treatment effects," *Tomography*, vol. 9, no. 3, pp. 1062–1070, 2023.

[25] M. F. Chernov, M. Hayashi, M. Izawa, Y. Ono, and T. Hori, "Proton magnetic resonance spectroscopy (mrs) of metastatic brain tumors: variations of metabolic profile," *International journal of clinical oncology*, vol. 11, pp. 375–384, 2006.

[26] W. H. Chae, K. Niesel, M. Schulz, F. Klemm, J. A. Joyce, M. Prümmer, B. Brill, J. Bergs, F. Rödel, U. Pilatus *et al.*, "Evaluating magnetic resonance spectroscopy as a tool for monitoring therapeutic response of whole brain radiotherapy in a mouse model for breast-to-brain metastasis," *Frontiers in oncology*, vol. 9, p. 1324, 2019.

[27] M. Hutchings and S. F. Barrington, "Pet/ct for therapy response assessment in lymphoma," *Journal of nuclear medicine*, vol. 50, no. Suppl 1, pp. 21S–30S, 2009.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[29] S. Tiwari, G. Jain, D. K. Shetty, M. Sudhi, J. M. Balakrishnan, and S. R. Bhatta, "A comprehensive review on the application of 3d convolutional neural networks in medical imaging," *Engineering Proceedings*, vol. 59, no. 1, p. 3, 2023.

[30] M. Ibrahim, "The basics of resnet50."

[31] H. Sahli, A. Ben Slama, and S. Labidi, "U-net: A valuable encoder-decoder architecture for liver tumors segmentation in ct images," *Journal of X-ray science and technology*, vol. 30, no. 1, pp. 45–56, 2022.

[32] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," in *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*. Springer, 2021, pp. 267–276.

[33] L. Zhang, C. Xu, Y. Li, T. Liu, and J. Sun, "Mcse-u-net: multi-convolution blocks and squeeze and excitation blocks for vessel segmentation," *Quantitative Imaging in Medicine and Surgery*, vol. 14, no. 3, p. 2426, 2024.

[34] M. Motiur Rahman, S. Shokouhmand, S. Bhatt, and M. Faezipour, "Mist: Medical

image segmentation transformer with convolutional attention mixing (cam) decoder," *arXiv e-prints*, pp. arXiv–2310, 2023.

[35] S. A. Jalalifar and A. Sadeghi-Naini, "Data-efficient training of pure vision transformers for the task of chest x-ray abnormality detection using knowledge distillation," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 1444–1447.

[36] B. B. Ozkara, M. M. Chen, C. Federau, M. Karabacak, T. M. Briere, J. Li, and M. Wintermark, "Deep learning for detecting brain metastases on mri: a systematic review and meta-analysis," *Cancers*, vol. 15, no. 2, p. 334, 2023.

[37] E. Chukwujindu, H. Faiz, A.-D. Sara, K. Faiz, and A. De Sequeira, "Role of artificial intelligence in brain tumour imaging," *European Journal of Radiology*, p. 111509, 2024.

[38] M. J. Amsbaugh and C. S. Kim, "Brain metastasis," 2017.

[39] S. K. Yoo, T. H. Kim, J. Chun, B. S. Choi, H. Kim, S. Yang, H. I. Yoon, and J. S. Kim, "Deep-learning-based automatic detection and segmentation of brain metastases with small volume for stereotactic ablative radiotherapy," *Cancers*, vol. 14, no. 10, p. 2555, 2022.

[40] I. Shin, H. Kim, S. Ahn, B. Sohn, S. Bae, J. Park, H. Kim, and S.-K. Lee, "Development and validation of a deep learning–based model to distinguish glioblastoma from solitary brain metastasis using conventional mr images," *American Journal of Neuroradiology*, vol. 42, no. 5, pp. 838–844, 2021.

[41] J. D. Rudie, R. Saluja, D. A. Weiss, P. Nedelec, E. Calabrese, J. B. Colby, B. Laguna, J. Mongan, S. Braunstein, C. P. Hess *et al.*, "The university of california san francisco brain metastases stereotactic radiosurgery (ucsf-bmsr) mri dataset," *Radiology: Artificial Intelligence*, vol. 6, no. 2, p. e230126, 2024.

[42] B. Ocaña-Tienda, J. Pérez-Beteta, J. D. Villanueva-García, J. A. Romero-Rosales, D. Molina-García, Y. Suter, B. Asenjo, D. Albillo, A. Ortiz de Mendivil, L. A. Pérez-Romasanta *et al.*, "A comprehensive dataset of annotated brain metastasis mr images with clinical and radiomic data," *Scientific data*, vol. 10, no. 1, p. 208, 2023.

[43] L. Nayak, E. Q. Lee, and P. Y. Wen, "Epidemiology of brain metastases," *Current oncology reports*, vol. 14, pp. 48–54, 2012.

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[45] mljourney, "Optimizing loss functions for imbalanced datasets."

[46] MayoClinic, "Brain stereotactic radiosurgery."

[47] L. V. Le, "Brain stereotactic radiosurgery."

[48] S. G. Sana Mohammadi1, "The role of mri in detecting and characterizing brain metastases from breast cancer."