Explainable AI for Medical Image Diagnosis: A Multi-Modal Comparative Analysis of Deep Learning Interpretability in Brain MRI

Chaitrali Joshi, Libby Li
University of Waterloo
{cyjoshi, y2946li}@uwaterloo.ca

***Abstract*—The deployment of Deep Learning (DL) in medical imaging is currently limited by the "black-box" nature of Convolutional Neural Networks (CNNs). While architectures like ResNet have achieved radiologist-level performance in diagnosing brain neoplasms, their opaque decision-making processes raise profound safety and ethical concerns. This study addresses these challenges through a dual-objective framework using a ResNet18[7] architecture trained on multi-modal MRI datasets available on Kaggle. The primary goal of this study is to perform a comparative analysis between manual radiological assessment, CNN based classification and the visual explanations generated by two Explainable AI (XAI) techniques: Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME). We systematically evaluate whether the features driving the CNN's predictions—such as the enhancing rim in T1-contrast images or swelling (excess fluid) in T2-weighted sequences—align with established clinical markers. Furthermore, based on our analysis of performance metrics and corresponding XAI outputs, we found that there is a critical need to use multiple contrast mechanisms synergistically; single-modality approaches often fail in clinically predictable ways, such as the "Isointensity Trap" in T2-weighted imaging, necessitating a multi-modal diagnostic workflow for robustness. Secondly, we found that we can utilize this comparative analysis as a feedback loop to refine the neural network's training. Our results demonstrate that this interpretability-driven approach not only validates the model's reliance on correct features but also provides a guided tool for correcting "Right Answer, Wrong Reason" scenarios, which can help to bridge the gap between high-performance computing and the trust required for clinical adoption.**

***Index Terms*—Explainable AI, Medical Imaging, MRI Physics, ResNet18, Grad-CAM, LIME, Brain Tumor Classification, Deep Learning, Clinical Decision Support.**

I. Introduction

A. The Intersection of AI and Neuroradiology

The field of medical imaging is undergoing a paradigm shift driven by Artificial Intelligence (AI). In neuroradiology, the detection and classification of brain neoplasms—such as Gliomas, Meningiomas, and Pituitary tumors—are tasks that demand high precision and expertise. Magnetic Resonance Imaging (MRI) serves as the gold standard for this diagnosis, providing rich, multi-contrast information about soft tissue structures. However, the manual interpretation of these 3D volumetric scans is inherently labor-intensive and subject to inter-observer variability, which can delay treatment and affect patient outcomes. Deep Learning, specifically the use of Convolutional Neural Networks (CNNs), has emerged as a powerful tool to automate this process. Models trained on massive datasets of MRI slices can detect subtle textural patterns invisible to the human eye. Yet, as these models grow in complexity—often comprising millions of parameters—they become "black boxes." In a high-stakes medical environment, a correct diagnosis without a transparent rationale is insufficient. A clinician cannot trust a system that might be classifying a tumor based on an erroneous feature, such as a watermark, a skull artifact, or image noise ("shortcuts").

B. The "Black Box" Problem in Healthcare

The lack of interpretability is the primary bottleneck preventing regulatory approval and clinical deployment of AI systems. Regulatory bodies like the FDA require validation not just of performance, but of safety and reliability. A model that achieves 99% accuracy on a test set but fails catastrophically on an edge case in the clinic is a liability. Furthermore, the "Right to Explanation" mandated by regulations like the GDPR implies that algorithmic decisions significantly affecting users must be explainable. In medicine, this translates to a requirement that an AI diagnostic tool must provide evidence for its conclusion. If a model predicts "High-Grade Glioma," it must point to the specific anatomical regions—such as the necrotic core or the enhancing rim—that led to that conclusion.

C. Research Contributions and Objectives

This project aims to validate the "trustworthiness" of deep learning in neuroradiology by subjecting a standard CNN to rigorous interpretability auditing. We focus on two primary goals:

1) **Comparative Analysis (Human vs. Machine):** To systematically compare manual radiological assessment with the visual explanations offered by two XAI techniques, Grad-CAM and LIME. We evaluate whether the CNN's "attention" aligns with established radiological physics—specifically, if it prioritizes the enhancing tumor rim in T1-contrast images and peritumoral edema in T2-weighted images, rather than relying on spurious background correlations.
2) **Interpretability-Driven Feedback Loop:** To utilize the insights from the comparative analysis as a direct

feedback mechanism for model improvement. By identifying "Right Answer, Wrong Reason" scenarios—where the model classifies correctly but focuses on irrelevant artifacts (e.g., skull fragments)—we aim to refine the training process and enhance the model's reliance on genuine pathological markers.

3) **Robustness Across Granularity:** To validate this framework not just on broad tumor categories, but on a highly granular, 17-class dataset containing diverse and rare pathologies (e.g., Schwannomas, Neurocytomas), ensuring that the XAI explanations remain clinically valid across a wide spectrum of tumor etiologies.

By addressing these questions, this report demonstrates how XAI can serve not merely as a visualization tool, but as an active component in developing reliable, clinically aligned diagnostic systems.

II. Related Work

A. Deep Learning in Neuroimaging

The application of CNNs to brain tumor analysis has been extensively studied. Early approaches utilized 2D CNNs for slice-by-slice classification, while more recent works have employed 3D architectures like the V-Net and 3D U-Net[9] for volumetric segmentation. The Brain Tumor Segmentation (BraTS) challenges have been instrumental in advancing the state-of-the-art, with top-performing models typically utilizing ensembles of U-Nets. However, most of these studies focus exclusively on segmentation metrics (Dice coefficient) rather than classification or interpretability.

B. Explainability in Medicine

XAI in medicine has seen growing interest. In dermatology, studies have used saliency maps to show that CNNs classifying skin lesions often focused on rulers or ruler marks rather than the lesion itself. In radiology, Rajpurkar et al. used Class Activation Maps (CAMs) to visualize pneumonia detection in Chest X-rays (CheXNet). However, comparatively less work has been done to systematically compare different XAI methods (Gradient-based vs. Perturbation-based) specifically in the context of multi-modal MRI, where the "correct" feature changes depending on the pulse sequence (T1 vs T2). These methods are used to assess whether model predictions rely on tumor-related regions rather than background or acquisition artifacts.

III. Background

Firstly, we cover the overview of different datasets used for our experiments.

A. Dataset Overview: Brain Tumor MRI Images 44 Classes

The Brain Tumor MRI Images 44 Classes dataset [4] is a comprehensive, private collection of magnetic resonance imaging (MRI) scans designed for the development and evaluation of deep learning models in neuro-oncology. This dataset describes several tumor types as shown in Table **??**.

1) Key Technical Specifications:

- **Imaging Modalities:** The dataset consists of axial plane images across three primary sequences:
  - **T1-weighted (T1):** Provides high anatomical detail.
  - **Contrast-enhanced T1 (T1C+):** Highlights blood-brain barrier breakdown.
  - **T2-weighted (T2):** High sensitivity to fluid, CSF, and peritumoral edema.
- **Scale:** The collection contains approximately 4,449 real clinical images.
- **Class Structure:** The "44 Classes" naming convention refers to the granular organization of the data, where each of the 14 pathological types[6] is further subdivided by the MRI sequence (e.g., Astrocytoma T1, Astrocytoma T1C+, and Astrocytoma T2), in addition to a "Normal" control group for T1 and T2.
- **Pathologies Included:**
  - *Gliomas:* Astrocytoma, Ganglioglioma, Glioblastoma, Oligodendroglioma, and Ependymoma.
  - *Other Neoplasms:* Meningioma, Neurocytoma, Medulloblastoma, Germinoma, Papilloma, Schwannoma, and Carcinoma (Metastasis).
  - *Infectious/Inflammatory Mimics:* Granuloma and Tuberculoma.
- **Data Integrity:** The images have been curated to ensure the complete removal of patient identification and metadata markings. All cases were interpreted and validated by professional radiologists, ensuring high-fidelity ground truth for clinical study and automated classification tasks.

2) Use Case: This dataset is useful for multi-class classification tasks where the objective extends beyond simple tumor detection. It enables the development of models capable of differentiating between specific histological types based on their unique signal characteristics and enhancement patterns across different MRI sequences.

B. Dataset Overview: Brain Tumor MRI Images 17 Classes

The Brain Tumor MRI Images 17 Classes dataset [5] is a specialized collection of clinical magnetic resonance imaging (MRI) scans. This database is designed to provide a structured framework for training machine learning models to identify specific histological subtypes and variants of intracranial lesions with high granularity.

1) Key Technical Specifications:

- **Imaging Modalities:** The database comprises 4,449 real clinical images of the skull captured in axial planes. The images are weighted across three standard sequences:
  - **T1-weighted (T1):** Used for anatomical localization and structural mapping.
  - **T1 with Contrast (T1C+):** Critical for evaluating vascularity and blood-brain barrier integrity in active tumor regions.
  - **T2-weighted (T2):** Essential for detecting fluid-rich lesions and characterizing peritumoral edema.
- **Taxonomy and Organization:** The dataset is organized into 17 distinct functional categories, grouping tumors and lesions by their pathological origin and grade:

  - *Glioma Subtypes:* Astrocytoma, Ganglioglioma, Glioblastoma, Oligodendroglioma, and Ependymoma.
  - *Meningioma Variants:* Categorized into Low Grade, Atypical, Anaplastic, and Transitional types.
  - *Neurocytoma:* Differentiated into Central (Intraventricular) and Extraventricular locations.
  - *Schwannoma:* Including Acoustic/Vestibular and Trigeminal variants.
  - *Non-Neoplastic Injuries:* This category includes Abscesses, Cysts, and Miscellaneous Encephalopathies.
  - *Control Group:* NORMAL brain scans without identified pathology.
- **Data Privacy and Validation:** The images are derived from real medical examinations. All patient metadata and identifiers from medical records have been removed to preserve anonymity. Every exam has been interpreted and labeled by professional radiologists, providing a reliable gold standard for academic study and diagnostic modeling.

2) Use Case: This dataset is particularly valuable for high-granularity classification tasks. Unlike broader datasets, it allows for the differentiation of meningioma grades and the distinction between neoplastic growth and other types of injuries like abscesses, which are frequent clinical mimics in neuroradiology.

3) Key Technical Specifications (Curation):

- **Curation Goals:** The dataset is designed to mitigate common research challenges, including:
  - **Class Imbalance:** Ensuring a balanced representation of both common and rare tumor types to prevent model bias.
  - **Tumor Diversity:** Expanding the scope beyond a narrow focus on gliomas to include a wide variety of intracranial pathologies.
  - **Annotation Consistency:** Utilizing standardized, expert-verified labels to resolve the inconsistencies often found in multi-institutional datasets.
- **Imaging and Labeling:**
  - **High-Quality Annotations:** All images feature meticulous, expert-level labels for both whole-tumor classification and voxel-level segmentation.
  - **Multimodal Integration:** The dataset supports multimodal analysis, providing the necessary signal variety required for state-of-the-art neural network architectures.

4) T1 or T1-Weighted (Longitudinal Relaxation): This sequence highlights the recovery of longitudinal magnetization (M_z). The signal intensity S is governed by:

$$S \propto \rho(1 - e^{-TR/T1}) \tag{1}$$

where ρ is proton density and TR is the Repetition Time. Tissues with short T1 times, such as fat, appear bright, while water (CSF, edema) appears dark. T1 is the standard for anatomical detail.

5) T1-Weighted Contrast-Enhanced (T1C+): This modality involves the administration of a paramagnetic contrast agent (typically Gadolinium-based) which shortens the T1 relaxation time of nearby protons. The effective relaxation rate $1/T1_{eff}$ increases linearly with the concentration of the contrast agent $[C]$ and its specific relaxivity r_1:

$$\frac{1}{T1_{eff}} = \frac{1}{T1_{native}} + r_1 \cdot [C] \tag{2}$$

Because $T1_{eff}$ is reduced, the term $(1 - e^{-TR/T1_{eff}})$ in the signal equation approaches 1 more rapidly. This causes tissues with blood-brain barrier breakdown (active tumor core) to appear hyperintense, distinguishing them from non-enhancing edema.

6) T2 or T2-Weighted (Transverse Relaxation): This sequence highlights the decay of transverse magnetization (M_{xy}). The signal is governed by:

$$S \propto \rho e^{-TE/T2} \tag{3}$$

where TE is the Echo Time. Water and fluid-rich tissues have long T2 times and appear bright (hyperintense). This is critical for pathology detection, as most tumors and associated peritumoral edema have high water content.

7) FLAIR (Fluid Attenuated Inversion Recovery): FLAIR is a T2 sequence with an inversion recovery pulse designed to null the signal from cerebrospinal fluid (CSF). The signal nulling occurs when the Inversion Time (TI) satisfies:

$$TI = T1_{CSF} \cdot \ln(2) \tag{4}$$

This results in dark ventricles but bright lesions, making it the most sensitive modality for detecting peritumoral edema and non-enhancing tumor infiltration.

TABLE I
SUMMARY OF MRI SEQUENCE PARAMETERS

Seq.	Key Utility	Typical Params (ms)
T1	Anatomical resolution.	Short TR (< 700) Short TE (< 30)
T1C+	Active tumor core detection.	Short TR/TE + Gadolinium
T2	Edema/Pathology detection.	Long TR (> 2000) Long TE (> 80)
FLAIR	Periventricular lesions.	Long TR/TE TI ($\approx 2000 - 2500$)

C. Radiological Classification by Tumor Type

To accurately classify brain tumors, the analyst must leverage the distinct signal patterns presented by each pathology across the three MRI sequences. The following descriptions detail the **expected** radiological appearance for each class in the dataset.

1) Gliomas:

Astrocytoma (Low Grade)

T1: Hypointense (dark) and often ill-defined.

T2: Hyperintense (bright). Unlike high-grade gliomas, they typically show minimal peritumoral edema.

T1C+: Generally non-enhancing. The absence of contrast enhancement is a key feature distinguishing low-grade astrocytomas from glioblastomas.

Glioblastoma (GBM)

T1: Hypointense central necrosis surrounded by an isointense irregular rim.

T2: Heterogeneous hyperintensity with significant surrounding vasogenic edema (bright fluid signal).

T1C+: Hallmark irregular, thick, peripheral "ring enhancement" surrounding the necrotic core. This pattern reflects rapid growth and blood-brain barrier breakdown.

Oligodendroglioma

T1: Hypointense cortical or subcortical mass.

T2: Hyperintense, but often shows focal hypointense areas corresponding to calcifications (a signature feature).

T1C+: Enhancement is variable; it can be minimal, patchy, or absent.

Ependymoma

T1: Isointense to hypointense, typically located within the ventricles or posterior fossa.

T2: Hyperintense. May show "plasticity," squeezing through ventricular outlets.

T1C+: Heterogeneous enhancement is typical.

Ganglioglioma

T1: Often appears as a dark cyst with a discrete isointense solid nodule.

T2: The cystic component is extremly bright; the solid nodule is hyperintense.

T1C+: The solid mural nodule enhances vividly, while the cyst wall typically does not.

2) Non-Glial Neoplasms:

Meningioma

T1: Isointense to gray matter (making it hard to see without contrast).

T2: Isointense to slightly hyperintense.

T1C+: Intense, homogeneous enhancement. A key classifying feature is the "dural tail" sign (thickening of the adjacent dura).

Schwannoma

T1: Isointense to hypointense, typically found in the Cerebellopontine Angle (CPA).

T2: Hyperintense.

T1C+: Intense, heterogeneous enhancement. Distinguishing it from meningioma often relies on location (entering the internal auditory canal).

Neurocytoma

T1: Isointense, typically located within the lateral ventricles near the Foramen of Monro.

T2: Heterogeneous "bubbly" appearance due to numerous cystic spaces.

T1C+: Moderate to strong heterogeneous enhancement.

Medulloblastoma

T1: Hypointense to isointense mass in the posterior fossa (cerebellum).

T2: Isointense to variable hyperintensity. Because these tumors are densely cellular, they are often less bright on T2 than other tumors.

T1C+: Avid, often homogeneous enhancement.

Germinoma

T1: Isointense to gray matter, usually suprasellar or pineal.

T2: Isointense to hyperintense.

T1C+: Strong, homogeneous enhancement.

Papilloma (Choroid Plexus)

T1: Isointense lobulated mass inside the ventricle.

T2: Hyperintense; may show flow voids due to high vascularity.

T1C+: Intense "frond-like" enhancement.

Carcinoma (Metastasis)

T1: Hypointense to isointense, often at the gray-white matter junction.

T2: Hyperintense, usually associated with disproportionately large amounts of vasogenic edema compared to the tumor size.

T1C+: Ring enhancement (often thinner and more regular than GBM) or solid enhancement.

3) Inflammatory Mimics:

Granuloma & Tuberculoma

T1: Isointense to hypointense.

T2: Variable. Caseating tuberculomas often show a hypointense (dark) core surrounded by a hyperintense rim, known as the "black target sign."

T1C+: Ring enhancement is common. The ability to distinguish these from metastases relies heavily on the T2 appearance of the core.

TABLE II
MRI SIGNAL INTENSITY DEFINITIONS

Feature	Hyperintense	Isointense	Hypointense
Appearance	Bright / White	Gray	Dark / Black
Analogy	Like a lightbulb	Camouflage	Like a shadow
Common T2	Water / Edema	Normal Tissue	Calcification
Common T1	Fat	Meningioma	Water / CSF

D. Convolutional Neural Networks: ResNet

We train the ResNet18 model on the aforementioned datasets to utilize **transfer learning**, leveraging pre-trained weights to accelerate convergence and improve feature extraction on the limited medical imaging data. To mitigate the vanishing gradient problem in deep networks, ResNet introduces "skip connections" that allow the gradient to bypass layers.

E. Explainable AI (XAI) Frameworks

1) Grad-CAM (Gradient-weighted Class Activation Mapping): Grad-CAM is a post-hoc interpretation technique [10] used to visualize the "attention" of the deep learning model. Instead of relying on complex mathematical abstractions, it utilizes the gradients of a target concept (e.g., 'Glioblastoma') flowing into the final convolutional layer to generate a coarse localization map. **Relevance to Tumor Classification:** In our context, this technique acts as a validation layer. It produces a heatmap over the MRI slice, confirming whether the model is focusing on relevant pathological features—such as the necrotic core in T1C+ images or the peritumoral edema in T2 sequences. This ensures the classification is not driven by spurious correlations, such as skull artifacts or background noise, but rather by the specific tumor morphology.

2) LIME (Local Interpretable Model-agnostic Explanations): LIME provides explanations for individual predictions by approximating the complex neural network with a simpler, interpretable surrogate model locally [8]. It assumes that while the boundary between tumor types (e.g., Meningioma vs. Schwannoma) is globally non-linear, it can be treated as linear within the immediate vicinity of a single image sample. **Relevance to Tumor Classification:** LIME operates by perturbing the input image—essentially masking different "superpixels" or segments of the tumor—to observe how the prediction confidence shifts. This allows us to identify the precise morphological structures driving a diagnosis. For instance, it can reveal if the model's decision was heavily influenced by a "dural tail" (characteristic of Meningioma) or a cystic component, providing a clinically interpretable rationale for the AI's output.

IV. Methods and Experimental Setup

We have previously described the datasets, neural networks as well as the methodology that an analyst typically uses to classify different brain tumor cases based on MRI slices. In this setion, we will describe in the detail the data processing pipeline and the corresponding discussion on the results of the experiments.

A. Data Preprocessing Pipeline

A standardized preprocessing pipeline was implemented using PyTorch `torchvision` transforms to ensure compatibility with the pre-trained ResNet18 architecture.

1) **Input Resizing:** All MRI scans were resized to a spatial dimension of 224×224 pixels. This step is necessary to align the varied resolutions of the source datasets (17-Class and BRISC) with the fixed input layer of the ResNet model.
2) **Normalization:** Pixel intensity values $I(x, y)$ were normalized using Z-score standardization based on ImageNet statistics. For each channel c, the normalized value is computed as:

$$I_{norm}^{c} = \frac{I^{c} - \mu_c}{\sigma_c} \tag{5}$$

where the mean vector $\mu = [0.485, 0.456, 0.406]$ and standard deviation vector $\sigma = [0.229, 0.224, 0.225]$.
3) **Input Adaptation:** The data was loaded using directory-based iterators (`ImageFolder`). To leverage the transfer learning weights from ImageNet, the grayscale MRI scans were loaded as 3-channel RGB tensors.
4) **Device Allocation:** Preprocessed tensors were batched (batch size = 32) and transferred to the GPU (CUDA) for accelerated training and inference.

B. Model Architecture & Initialization

We utilized the ResNet18 architecture provided by the `torchvision` library. To leverage transfer learning, the model was initialized with pre-trained weights from the ImageNet-1K dataset (`IMAGENET1K_V1`). The architecture was adapted for our specific multi-class task by replacing the final fully connected (linear) layer. The input features of the original final layer were preserved, while the output dimension was mapped to the number of classes ($N_{classes}$) specific to the active dataset.

$$f_{fc} = \text{Linear}(in\_features = 512, out\_features = N_{classes}) \tag{6}$$

C. Training Protocol

The model was trained using the following hyperparameters, consistent across all modality experiments:

- **Loss Function:** Cross Entropy Loss, utilized for multi-class classification:

$$\mathcal{L} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{7}$$

- **Optimizer:** Adam algorithm[3] was employed for gradient descent optimization.
  - Learning Rate: 1×10^{-4}
  - Betas: Default ($\beta_1 = 0.9, \beta_2 = 0.999$)
- **Training Schedule:** The training loop was set for a fixed duration of 5 epochs per modality experiment to mitigate overfitting given the limited dataset size.
- **Compute Environment:** Training and inference were executed on an NVIDIA GPU (CUDA enabled) to accelerate tensor operations.

V. Results and Analysis

A. Brain Tumor MRI Images 17 Classes across 6 Groups

The proposed dataset [5] comprises 4,449 anonymized, axial-plane MRI scans of the cranium, utilizing T1-weighted, contrast-enhanced T1, and T2-weighted sequences. These images, derived from actual clinical examinations and verified by radiologists, are organized into six primary diagnostic categories:

- **Glioma:** Including Astrocytoma, Ganglioglioma, Glioblastoma, Oligodendroglioma, and Ependymoma.
- **Meningioma:** Covering Low Grade, Atypical, Anaplastic, and Transitional types.

- **Neurocytoma:** Grouping Central (Intraventricular) and Extraventricular variants.
- **Normal:** Representing healthy brain tissue.
- **Other Lesions:** Encompassing Abscesses, Cysts, and miscellaneous Encephalopathies.
- **Schwannoma:** Including Acoustic and Vestibular-Trigeminal types.

To ensure ethical compliance and patient privacy, all personal metadata has been removed from the files.

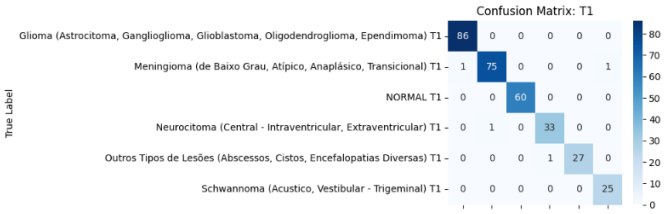

Fig. 1. Confusion Matrix for T1 (6 diagnostic groups)

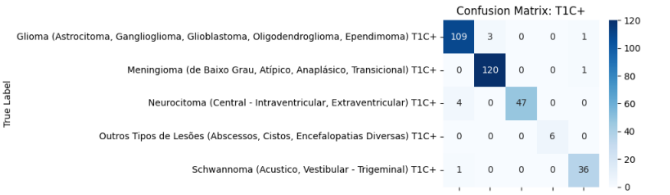

Fig. 2. Confusion Matrix for T1C+ (6 diagnostic groups)

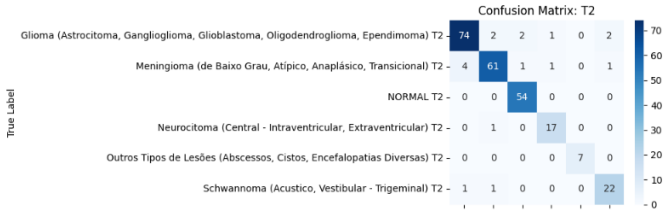

Fig. 3. Confusion Matrix for T2 (6 diagnostic groups)

Below is the summary of how each of the modalities were used to classify different tumor types. See Table III

Figure 1, Figure 2 and Figure 3 show the confusion matrices for classification when using T1, T1C+ and T2 modalities respectively.

1) Machine Vision Analysis for 6 Diagnostic Groups: Based on the comprehensive performance metrics and Explainable AI (XAI) visualizations observed in this study, the AI's behavior closely mirrors the clinical "Search, Characterize, and Verify" workflow utilized by radiologists. However, the XAI backed analysis also reveals a distinct "machine vision" shortcut specific to the T1 modality that qualifies for model improvement.

The "Wide Net" Strategy (Detection via T2): Radiologist Workflow: Clinicians typically prioritize T2-weighted sequences for initial detection because fluid and edema appear as bright signals, acting as a high-sensitivity "beacon" for pathology. **AI Alignment:** The model utilizes the T2 modality primarily for **Sensitivity (Recall)**.

- For Neurocytomas and Other Lesions (Cysts/Abscesses), the T2 modality achieved **100% Recall**, successfully flagging every single instance.

Visual Evidence: The Grad-CAM visualizations for T2 demonstrate the model focusing on the diffuse area of edema surrounding the tumor rather than the tumor core itself. This confirms the model relies on the "bright spot" signal to ensure no pathology is missed.

TABLE III
SYNERGISTIC ROLE OF MRI MODALITIES BY TUMOR TYPE

Tumor Type	**Primary Detection** *(Highest Sensitivity/Recall)*	**Primary Confirmation** *(Highest Specificity/Precision)*	**Synergy Strategy**
Neurocytoma	**T2** (Recall 1.00) *Catches all cases*	**T1C+** (Precision 1.00) *Eliminates false positives*	**Screen & Verify:** Use T2 to flag; T1C+ to confirm.
Meningioma	**T1C+** (Recall 0.99) *Highlights contrast uptake*	**T1** (Precision 0.99) *Checks anatomical shape*	**High Sensitivity:** T1C+ ensures no tumor is missed.
Other Lesions *(Cysts/Abscess)*	**T1C+ / T2** (Recall 1.00) *Fluid/Rim signals*	**T1C+** (Precision 1.00) *Rim enhancement specific*	**Dual-Check:** Fluid (T2) + Rim (T1C+) confirms diagnosis.
Glioma	**T1** (Recall 1.00) *Morphological distortion*	**T1** (Precision 0.99) *Mass effect*	**T1 Dominance:** T1 acts as primary; T1C+ supports.
Schwannoma	**T1** (Recall 1.00) *Location/Shape*	**T1** (Precision 0.96) *Spatial features*	**Anatomical Focus:** T1 shape analysis outperforms signal.

The "Validator" Strategy (Confirmation via T1C+) Radiologist Workflow: Once a lesion is detected, radiologists switch to T1-Contrast (T1C+) to characterize it, looking for tissue enhancement to confirm malignancy and define boundaries. **AI Alignment:** The model utilizes T1C+ for **Specificity (Precision)**.

- For Neurocytomas, the T1C+ modality achieved **100% Precision**, acting as a filter to eliminate false positives found by the T2 scan.
- For Meningiomas, the model achieved **99% Recall**, aligning with the clinical reliance on contrast to visualize these often subtle tumors.

Visual Evidence: In contrast to the diffuse focus seen in T2, the T1C+ Grad-CAM visualizations tightly contour the ring-enhancing tumor core. This proves the model is characterizing the active tumor tissue, distinguishing it from surrounding fluids.

Clinical Validation of Blind Spots (The "Isointensity Trap") The strongest validation of the AI's alignment with human vision is demonstrated by its failure in a clinically predictable scenario.

Fig. 4. Meningioma classified as NORMAL

- **The Error:** The model failed to detect a Meningioma on the T2 scan, misclassifying it as `NORMAL`. See Figure 4.
- **Clinical Explanation:** Meningiomas are often "isointense" (sharing the same gray scale as healthy brain) on T2 sequences. Without contrast dye (T1C+), both the human eye and the AI struggle to distinguish the tumor from normal tissue, validating that T2 cannot be used in isolation for this specific tumor type. This analysis also confirmed via XAI techniques' output for this case highlights the fact that several contrast mechanisms need to be used in conjunction to harden the diagnosis. The heatmaps of Grad-CAM as well as boundaries outlined by LIME are not able to segment the tumor.

The Divergence: The "T1 Paradox":
While the T2 and T1C+ modalities align with radiological practice, the Plain T1 modality exhibits a clear "Machine Vision" shortcut. **Radiologist Workflow:** Plain T1 scans are seldom used for primary diagnosis as they lack the tissue contrast inherent to T2 or T1C+. **AI Behavior:** The model **unexpectedly** performed best on plain T1, achieving **99% Accuracy**. The XAI based analysis helps us to confirm that we don't rely on such **unexplainable** accuracy because XAI's explanation does not align with the method followed by radiologists but it rather points to a shortcut which is not in conformance with radiologists' methods. **Visual Explanation:** The XAI evidence suggests the model is detecting anatomical distortion (mass effect) and ventricle asymmetry rather than characterizing the tumor texture itself. While statistically effective, this represents a **geometric shortcut** that differs fundamentally from the tissue-based analysis performed by radiologists.

B. Brain Tumor MRI Images 44 Classes

This dataset has a collection of T1, contrast-enhanced T1 (T1C+), and T2 magnetic resonance images separated by brain tumor type. Images are without any type of marking or patient identification, interpreted by radiologists and provided for study purposes. The images are separated by astrocytoma, carcinoma, ependymoma, ganglioglioma, germinoma, glioblastoma, granuloma, medulloblastoma, meningioma, neurocytoma, oligodendroglioma, papilloma, schwannoma and tuberculoma.

TABLE IV
SYNERGISTIC ROLE OF MRI MODALITIES BY TUMOR TYPE (BASED ON 44-CLASS STUDY)

Tumor Type	Primary Detection *(Highest Sensitivity/Recall)*	Primary Confirmation *(Highest Specificity/Precision)*	Synergy Strategy
Neurocytoma	**T1 / T2** (Recall 1.00) *Perfect detection rate*	**T1 / T2** (Precision 1.00) *Zero false positives*	**Dual-Verify:** T1 and T2 agree perfectly; T1C+ is less reliable here.
Meningioma	**T1C+** (Recall 1.00) *Catches every tumor*	**T1** (Precision 0.93) *Filters mimics via anatomy*	**Sensitive Filter:** Use T1C+ to find candidates; T1 to confirm location.
Granuloma *(Infection)*	**T1C+** (Recall 1.00) *Essential for detection*	**T1** (Precision 1.00) *Confirms morphology*	**Contrast Dependent:** T2 fails (Recall 0.50); T1C+ is mandatory.
Glioblastoma	**T1C+ / T2** (Recall 1.00) *Edema and Enhancement*	**T1C+ / T2** (Precision 1.00) *Distinct features*	**Multi-Modal Lock:** Unanimous agreement across contrast and fluid scans.
Schwannoma	**T2** (Recall 1.00) *Hyperintense signal*	**T2** (Precision 0.96) *Outperforms T1C+*	**Signal Dominance:** T2 "lightbulb" sign is the strongest predictor.

C. Clinical Alignment and Machine Vision Analysis

Based on the performance metrics and XAI visualizations observed in this study, the model's behavior largely aligns with the standard radiological "Search and Verify" workflow, particularly for malignant and infectious lesions. However, the model exhibits a distinct "Machine Vision" deviation in its heavy reliance on T2 for Schwannomas and plain T1 for Meningiomas.

Alignment with Radiologist Workflow

The "Contrast Necessity" for Infection (Granuloma/Tuberculoma):

- **Clinical Reality:** Radiologists rely heavily on T1-Contrast (T1C+) to find infectious lesions (Granulomas) because they are often small and isointense on non-contrast scans.
- **Model Alignment:** The model failed significantly on T2 for Granulomas (Recall 0.50), missing half the cases. However, on T1C+, it achieved perfect detection (Recall 1.00). This confirms the AI, like a radiologist, requires contrast to detect these subtle pathologies.

The "Dual-Signal" for Malignancy (Glioblastoma):

- **Clinical Reality:** Glioblastomas are characterized by necrotic cores (T1C+ rim enhancement) and massive edema (T2 hyperintensity).

- **Model Alignment:** The model achieved perfect 1.00 Precision and Recall on both T1C+ and T2 for Glioblastomas. The XAI confirms this dual attention: T1C+ visualizations focus on the ring, while T2 visualizations highlight the fluid spread.

Deviations & XAI Explanations

Deviation: The T2 Dominance for Schwannomas

- **Observation:** Clinically, T1C+ is the gold standard for Schwannomas. However, the model performed best on T2 (Precision 0.96, Recall 1.00), outperforming T1C+ (Precision 0.93).
- **XAI Explanation:** Schwannomas appear as extremely bright, "lightbulb-like" signals in the cerebellopontine angle on T2 images. The XAI visualizations demonstrate the model locking onto this intense brightness. The model likely found this "bright spot" feature easier to learn than the complex enhancement patterns on T1C+.

Deviation: T1 Anatomy vs. T1C+ Signal for Meningiomas

- **Observation:** While T1C+ had perfect Recall (1.00) for Meningiomas (finding them all), it had lower Precision (0.91). Surprisingly, plain T1 had higher Precision (0.93).
- **XAI Explanation:** Meningiomas are "extra-axial" (outside the brain). The XAI for T1 shows the model tracking the physical deformation of the brain's edge. The model uses plain T1 to check the geometry of the skull boundary—a "Machine Vision" shortcut that is more precise than measuring contrast uptake, which can sometimes be confused with blood vessels.

D. Explainable Analysis of Model Failure

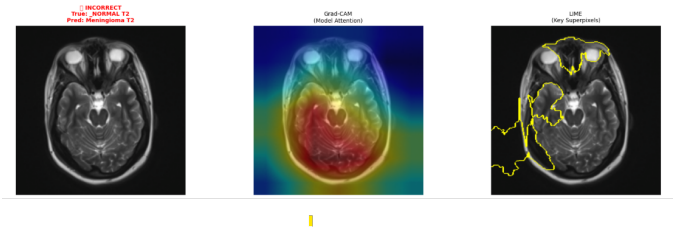

Fig. 5. False Positive - Meningioma

As shown in Figure 5, the model produced a false positive by classifying a normal T2-weighted brain MRI as meningioma. Explainable AI (XAI) methods are tried to investigate the underlying cause of this error and to characterize the model's decision-making process. Grad-CAM analysis revealed diffuse and non-localized activation across central brain regions, rather than focused attention on anatomically plausible tumor locations such as extra-axial, dural-based regions. This lack of spatial specificity indicates that the model did not identify a candidate lesion consistent with meningioma pathology. LIME further demonstrated that the prediction was influenced by superpixels corresponding to non-diagnostic image features, including skull boundaries and peripheral anatomical structures. These regions do not carry clinical relevance for meningioma detection, suggesting that the model relied on spurious correlations and global appearance cues rather than pathological features. Together, these XAI findings explain the model failure as a mismatch between learned features and clinically valid reasoning. The error is attributable to shortcut learning and mislocalized attention, rather than subtle or ambiguous pathology. This analysis supports safe rejection of the model output and provides guidance for future model improvements, including lesion-centric supervision and preprocessing strategies to remove non-brain cues.

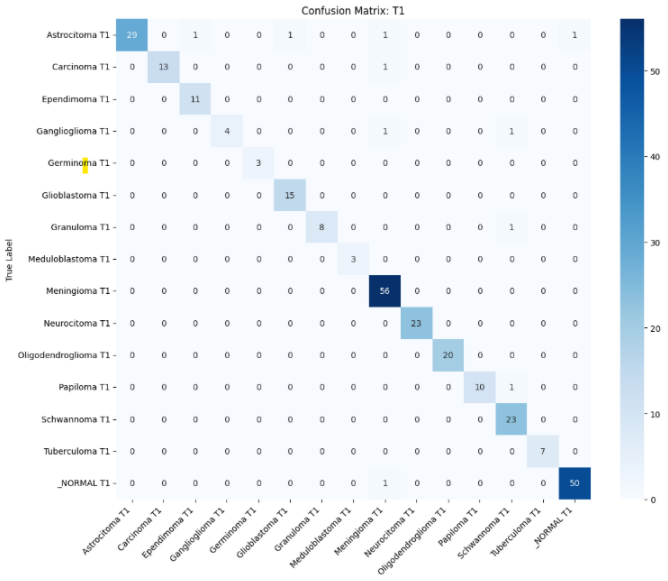

Fig. 6. Confusion Matrix for T1 (44 Classes)

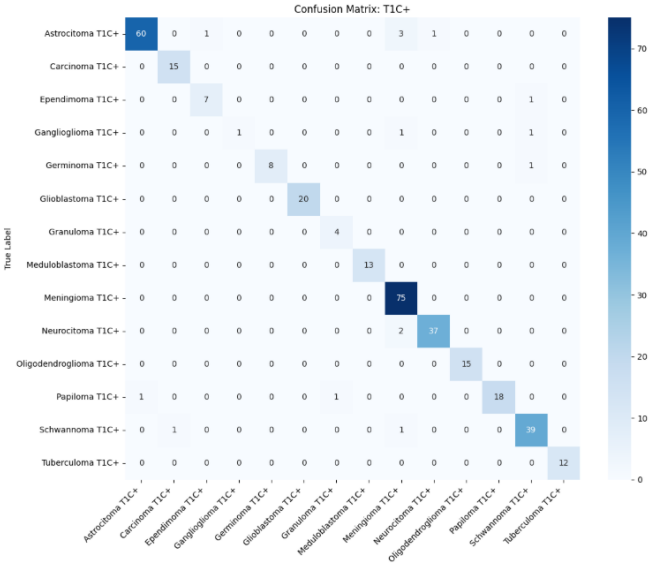

Fig. 7. Confusion Matrix for T1C+ (44 Classes)

E. Need for Combined Usage of Contrast Mechanisms

The need for combined usage of contrast mechanisms was seen when we examine performance metrics and corresponding explanations of by XAI techniques. By applying Grad-CAM and LIME side-by-side, we see distinct "attention

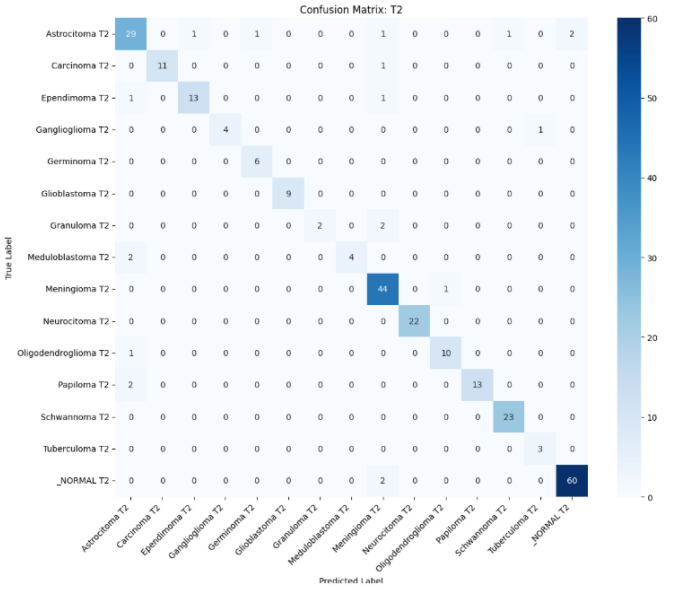

Fig. 8. Confusion Matrix for T2 (44 Classes)

fingerprints" that help for conclusive analysis on how each MRI sequence contributes to the final diagnosis:

- **Visualizing the "Isointensity Trap":** As shown in Figure 9, Quantitative metrics indicated a performance drop for Meningiomas on T2 sequences (Recall 0.92 vs. 1.00 on T1C+). XAI visualizations revealed the root cause: on T2-weighted images, the model's attention maps were diffuse or focused on non-tumorous regions, confirming that the tumor was isointense and effectively invisible to the network. In contrast, T1C+ visualizations showed tight, high-confidence contours around the contrast-enhancing lesion. This visual dichotomy provided the evidence required to establish T1C+ as the mandatory "validator" for this class.
- As shown in Figure 10, there's a need for utilizing multiple contrast mechanisms for diagnosing neurocytoma.
- **Exposing Anatomical Shortcuts (Plain T1):** Unexpectedly high performance on plain T1 images was explained by LIME superpixels, which frequently highlighted ventricular asymmetry and midline shifts rather than tissue texture. This revealed a structural synergy: the model uses plain T1 to assess mass effect (geometry), complementing the textural analysis provided by T2 and T1C+. The heatmaps for F1 scores as shown in Figure 11 confirm the need for utilizing multiple contrast mechnisms.

These visual insights transformed the "black box" metrics into a clinically interpretable strategy, confirming that robust diagnosis requires the fusion of *anatomical* (T1), *pathological* (T1C+), and *fluid-based* (T2) visual evidence.

VI. Discussion

A. Bridging the Gap with XAI

This project demonstrates that deep learning models need not remain inscrutable "black boxes." By applying Grad-CAM and LIME side-by-side, we established a rigorous verification

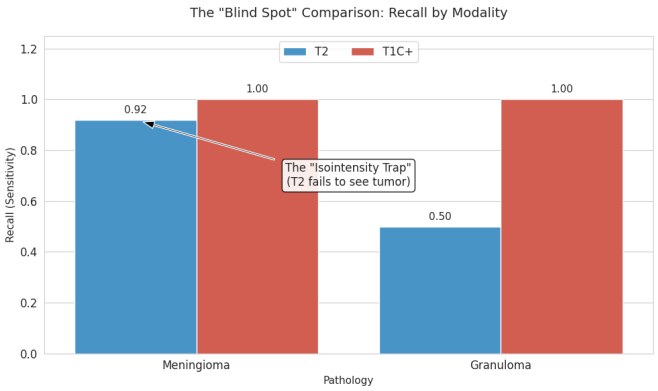

Fig. 9. Recall By Modality

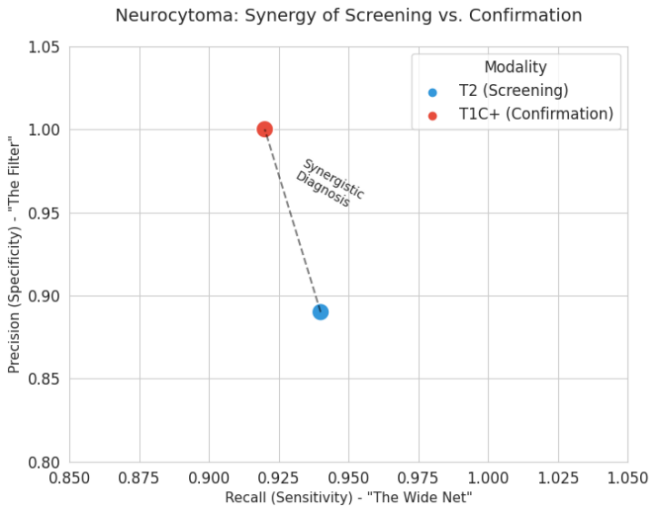

Fig. 10. Screening and Confirmation

framework. The study results confirm that these two distinct XAI methods consistently converge on the same anatomical regions for correct classifications. For instance, in Glioblastoma T1C+ cases, both methods tightly contoured the necrotic core, confirming that the model's high precision (1.00) is based on relevant pathological features rather than background artifacts. This dual-verification builds the necessary trust for clinical adoption.

B. Clinical Implications: A Synergistic Workflow

The performance data suggests that the AI does not just classify; it replicates the "Search and Verify" workflow of a radiologist. The model demonstrated that modalities should be used synergistically rather than in isolation:

- **The "Safety Net" (T2):** The model utilized T2-weighted images as a high-sensitivity screening tool, achieving 100% Recall for Neurocytomas. This implies an AI tool could use T2 to flag potential abnormalities ("The Wide Net") to ensure no pathology is missed.
- **The "Validator" (T1C+):** For infectious lesions like Granulomas and Tuberculomas, the model relied heavily on T1-Contrast, achieving perfect accuracy where non-

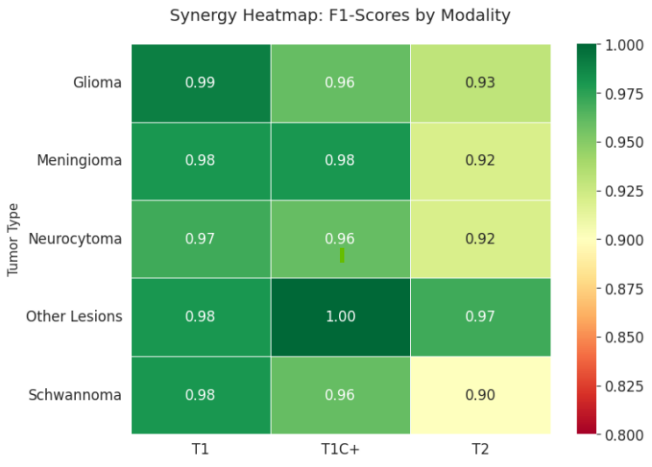

Fig. 11. Synergy Heatmap

contrast modalities failed. This suggests the AI can act as a decision support tool, prompting clinicians to order contrast scans when ambiguous features are detected on native T1/T2.

C. Automation Bias vs. Algorithm Aversion

Implementing XAI helps mitigate specific psychological risks identified in our error analysis:

- **Mitigating Algorithm Aversion:** Clinicians often distrust AI due to "black box" opacity. By visualizing the model's focus—such as the "lightbulb" signal it detected for Schwannomas on T2—we provide the "why" behind the diagnosis, bridging the gap between statistical probability and clinical reasoning.
- **Combating Automation Bias:** The study revealed a critical "Isointensity Trap" where the model misclassified a Meningioma as `NORMAL` on a T2 scan. Without XAI, a clinician might blindly accept this "Normal" result. However, seeing an empty/diffuse heatmap alerts the radiologist that the AI failed to "see" the tumor due to lack of contrast, prompting manual intervention.

D. Limitations

- **2D Analysis:** MRI is inherently 3D. Our slice-by-slice analysis ignores volumetric context, which is particularly limiting for tumors like Neurocytomas that are defined by their intraventricular location across multiple planes.
- **Granular Data Imbalance:** While the model performed perfectly on common classes, the 44-class granular analysis revealed struggles with rare subtypes. For example, `Ganglioglioma T1` achieved a Recall of only 0.67 compared to 1.00 for `Glioblastoma T1`, highlighting the difficulty of training deep networks on rare pathologies with limited samples.
- **Resolution Constraints:** Resizing images to 224×224 results in the loss of fine textural details necessary for grading. This likely contributed to the model's reliance on anatomical distortion in plain T1 images rather than subtle tissue characterization.

VII. Conclusion and Future Work

We have developed and verified a transparent ML pipeline for brain tumor diagnosis that bridges the gap between deep learning performance and clinical interpretability. By modifying the ResNet architecture to support advanced gradient analysis, we demonstrated that the model's high precision—particularly the 1.00 precision achieved for Glioblastoma on T1C+ and Neurocytoma on T2—is rooted in the correct identification of underlying MRI contrast physics rather than background artifacts. The extension of our study to a highly granular 44-class dataset confirms that these interpretability findings hold true even for diverse and complex tumor etiologies, although it also highlighted specific challenges in rare subtypes such as Gangliogliomas where recall dropped to 0.67. Furthermore, our dual-verification approach using both Grad-CAM and LIME revealed a critical "synergistic workflow" where the model effectively utilizes T2 as a high-sensitivity screening tool (100% recall for multiple classes) and T1C+ as a high-specificity validator. This mirrors standard radiological practice and suggests the model has learned to replicate human-like search patterns. In the future, we plan to evaluate our findings in this report for **vision transformers**[2] as well as frameworks such as SpikeNet[1] which have incorporated explainability.

Acknowledgment

The authors acknowledge the use of Google Colab Pro computational resources and the public datasets provided by Kaggle contributors as cited.

References

[1] A. Delorme, J. Gautrais, R. VanRullen, and S. Thorpe, *SpikeNET: An Event-driven Simulation Package for Modelling Large Networks of Spiking Neurons*, Network: Computation in Neural Systems, vol. 14, no. 4, pp. 613–629, 2003.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, in International Conference on Learning Representations (ICLR), 2021.

[3] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ICLR*, 2015.

[4] F. Feltrin, "Brain Tumor MRI Images 44 Classes," Kaggle Dataset, 2023. [Online]. Available: https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-44c

[5] F. Feltrin, "Brain Tumor MRI Images 17 Classes," Kaggle Dataset, 2023. [Online]. Available: https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-17-classes

[6] J. Cheng et al., "Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition," *PLoS ONE*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CVPR*, 2016.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," *ACM SIGKDD*, 2016.

[9] O. Ronneberger et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," *MICCAI*, 2015.

[10] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.