

MINOR PROJECT REPORT
For
B.Tech. Pre-Final Year (7th Semester)



DEPARTMENT OF INFORMATION TECHNOLOGY
BHARATI VIDYAPEETH'S COLLEGE OF
ENGINEERING

Neuro-XAI: Explainable deep learning framework for Brain Tumor Detection

MINOR PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

by

**Niharika Kashyap
03511503122**

Guided by

**Dr. Neha Gupta
Assistant Professor**



**DEPARTMENT OF INFORMATION TECHNOLOGY
BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING (AFFILIATED TO
GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY, DELHI)
NEW DELHI – 110063 NOV.2025**

CANDIDATE'S DECLARATION

It is hereby certified that the work which is being presented in the B. Tech Minor project Report entitled "**Neuro-XAI: Explainable deep learning framework for Brain Tumor Detection**" in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** and submitted in the **Department of Information Technology** of **BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING, New Delhi** (Affiliated to Guru Gobind Singh Indraprastha University, Delhi) is an authentic record of our own work carried out during a period from **August 2025 to November 2025** under the guidance of **Dr. Neha Gupta, Assistant Professor**. The matter presented in the B. Tech Major Project Report has not been submitted by me for the award of any other degree of this or any other Institute.

(Niharika Kashyap)

(En. No: 03511503122)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge. She is permitted to appear in the External Minor Project Examination.

(Dr. Neha Gupta)
Assistant Professor

(Prof. Prakhar Priyadarshi)
HOD, IT

ABSTRACT

Brain tumors, particularly highly infiltrative gliomas, remain one of the most critical neurological diseases. Accurate and timely segmentation of tumour subregions from multimodal Magnetic Resonance Imaging (MRI) is vital for treatment planning, but the traditional manual process is time-consuming and subjective. While deep learning models, such as Convolutional Neural Networks (CNNs), offer automation, their "black-box" nature presents a significant barrier to adoption in high-stakes clinical settings. This project introduces Neuro-XAI, an Explainable Deep Learning framework for the precise segmentation and classification of gliomas using the BraTS 2024 dataset. We evaluate the performance of three architectures: a 2D UNet baseline, a 2D CNN + Transformer hybrid, and a 3D CNN + Transformer volumetric model. The experimental results demonstrate that the 3D CNN + Transformer architecture achieved the highest segmentation performance, with an average Dice score of 0.82 and an overall accuracy of 99%. Crucially, the framework integrates two distinct Explainable AI (XAI) techniques: Gradient-weighted Class Activation Mapping (Grad-CAM) provides visual evidence by mapping the model's spatial attention to tumor regions, while Shapley Additive explanations (SHAP) offers a quantitative measure of input voxel contributions. The combination of high-accuracy predictive modelling and robust, verifiable interpretability ensures that the Neuro-XAI framework is not only superior in performance but also clinically trustworthy, thereby enhancing diagnostic confidence and utility.

ACKNOWLEDGEMENT

I express my deep gratitude to Dr. Neha Gupta, Assistant Professor, Department of Information Technology for her valuable guidance and suggestion throughout my project work. I am thankful to Dr. Arun Kumar Dubey and Dr. Achin Jain, for their valuable guidance.

I would like to extend my sincere thanks to Head of the Department, Prof. Prakhar Priyadarshi for his time to time suggestions to complete my project work. I am also thankful to Dr. Dharmender Saini , Principal for providing me the facilities to carry out my project work.

Sign

(Niharika Kashyap)

(En. No: 03511503122)

TABLE OF CONTENT

	Page No (in Roman)
Declaration	ii
Abstract	iii
Acknowledgement	iv
List of Figures	vi
List of Tables	vii
	Page No (in numeric)
Chapter 1: Introduction	1
1.1 : Background	
1.2 : Motivation	
1.3 Objective	
1.4 Summary of the Report	
Chapter 2: Project Overview and Methodology	4
2.1 : Project Overview	
2.2 : Dataset	
2.3 Deep Learning Architectures	
2.4: Explainable AI (XAI) Framework	
2.5 Methodology	
Chapter 3: Results & Discussion	13
3.1 : Training Environment	
3.2 : Quantitative Evaluation of Model Performance	
3.3 : Analysis of Segmentation Performance	
3.4 : Explainability: Results and Discussion	
Chapter 4: Conclusion	17
Bibliography	18

LIST OF FIGURES

Fig. No.	Title	Page No.
2.2.1	Dataset Modalities	5
3.3.1	Segmentation.....	15
3.4.1	Grad-CAM Visualization.....	16
3.4.2	SHAP Analysis.....	16

LIST OF TABLES

Table	Title	Page
2.2.1	Segmentation Classes	4
2.2.2	Dataset Modalities	5
3.2.1	Quantitative Results	13

ABBREVIATION

Abbreviation	Full Form
XAI	Explainable Artificial Intelligence
MRI	Magnetic Resonance Imaging
CNN	Convolutional Neural Network
ViT	Vision Transformer
T1n	T1-weighted native
T1c	T1-weighted contrast-enhanced
T2w	T2-weighted
T2f	T2 Fluid-Attenuated Inversion Recovery (FLAIR)
BraTS	Brain Tumor Segmentation (Challenge/Benchmark)
Grad-CAM	Gradient-weighted Class Activation Mapping
SHAP	SHapley Additive exPlanations
IoU	Intersection over Union
ReLU	Rectified Linear Unit

CHAPTER 1: INTRODUCTION

Brain tumors remain one of the most serious medical conditions, where early and accurate diagnosis is essential for effective treatment and improved survival rates. Magnetic Resonance Imaging (MRI) is the preferred technique for tumor analysis as it provides high-resolution images. In recent years, deep learning has advanced the automation of tumor detection and classification, moving past the time-consuming and subjective nature of manual segmentation by expert neuroradiologists. However, the lack of transparency in these deep learning models poses a significant barrier to clinical adoption due to their "black-box" nature. This project addresses this crucial challenge by combining the predictive strength of deep learning with the interpretability of Explainable Artificial Intelligence (XAI).

1.1 Background

Gliomas, which are the focus of this study, make up nearly 80% of all malignant brain tumors. The accurate detection and segmentation of tumor subregions (e.g., the necrotic core, peritumoral edema, and enhancing tumor) are crucial for surgical resection planning and precision targeting. Deep learning models like Convolutional Neural Networks (CNNs), particularly the UNet architecture and its variants, have demonstrated outstanding performance in medical imaging tasks by effectively capturing both spatial and contextual information. More recently, Vision Transformers (ViTs) and hybrid CNN-transformer architectures have been utilized to improve the modelling of global dependencies across MRI slices.

1.2 Motivation

The motivation for this project stems from the critical need for accurate and transparent diagnostic tools in neuro-oncology.

1.2.1 Literature Survey

Recent developments in deep learning (DL) have established a new paradigm in medical image examination, specifically in the field of brain tumor recognition and categorization. Explainable Artificial Intelligence (XAI) has emerged in an attempt to resolve the apparent conflict between deep neural networks' high predictive capabilities and their poor interpretability, which is a key impediment to clinical application. Gundogan (2025) suggested a hybrid CNN–XGBoost model with Grad-CAM for

multi-class brain tumor classification, achieving an accuracy of 99.77% on MRI data. Notably, this model was computationally heavy and restricted to image-level explanations, providing no voxel-level localization. Similarly, Mastoi et al. (2025) proposed a federated learning model with GoogLeNet that preserved privacy by not requiring the uplink of images while achieving 94% accuracy, but did not produce spatially interpretable results. The federated learning model was computationally heavy as well. Hosny and Mohammed (2025) reviewed CNN-based, Vision Transformer (ViT)-based, and hybrid-based approaches for brain tumor classification and segmentation, highlighting challenges in dataset imbalance, lack of 3D contextual learning, and limited interpretability. Musthafa et al. (2024) utilized the ResNet50 model with Grad-CAM and reported 98.52% accuracy, although their study focused solely on 2D classification and ignored volumetric segmentation. Iftikhar et al. (2025) developed a lightweight CNN integrating Grad-CAM, SHAP, and LIME, obtaining 95–99% accuracy while increasing explainability. However, the work was restricted to qualitative heatmap-based reasoning. Vamsidhar et al. (2025) presented hybrid ViT-Random Forest, ResNet-Xception, and LIME architectures achieving over 99% accuracy, but the models remained computationally expensive and applicable mainly to classification rather than segmentation.

Identified Research Gaps:

1. Existing methods focus on 2D classification with no voxel-level interpretability.
2. Many existing explanations are qualitative, not providing true quantifiable decision transparency.
3. Hybrid architectures are costly to run and may have dataset-specific biases.

1.3 Objective

The primary objectives of this project were:

1. To develop and implement a robust, explainable deep learning framework (Neuro-XAI) for the precise identification and multi-class segmentation of gliomas using multimodal MRI data.
2. To evaluate the performance of three distinct deep learning architectures 2D UNet, 2D Transformer UNet, and 3D Transformer-UNet on the BraTS 2024 Glioma Dataset.
3. To achieve superior performance in segmentation accuracy, focusing on the average Dice score and overall accuracy.
4. To integrate Explainable AI (XAI) techniques, specifically Grad-CAM and Shapley Additive explanations (SHAP), to generate visual and quantitative insights into the model's predictions, thereby enhancing clinical interpretability and trust.

1.4 Summary of the Report

The report documents the development and evaluation of the Neuro-XAI framework for brain tumor segmentation and classification.

- Chapter 1: Introduction establishes the critical context of brain tumor diagnosis, the need for automated solutions, the limitations of current deep learning models, and outlines the specific objectives of this explainable AI project.
- Chapter 2: Description about the Project details the Methodology, including the collection and preparation of the multimodal MRI data, the technical specifications of the three deep learning architectures (2D UNet, 2D CNN+Transformer, and 3D CNN+Transformer), and the mathematical framework of the integrated Grad-CAM and SHAP explainability modules.
- Chapter 3: Results & Discussion presents the experimental performance metrics, showing a consistent improvement from the baseline 2D UNet to the hybrid transformer models, with the 3D CNN + Transformer achieving the highest Dice score and accuracy. The core of this chapter is the discussion of the XAI outputs, demonstrating how the models utilized specific spatial activations and feature contributions to arrive at their segmentation and classification results.
- Chapter 4: Conclusion summarizes the success of the Neuro-XAI framework, confirms the superiority of the 3D CNN+Transformer model, and highlights the successful integration of XAI as a key contribution to improving model transparency and clinical utility

Chapter 2: PROJECT OVERVIEW AND METHODOLOGY

2.1 Project Overview

The project, addresses a critical translational gap in neuro-oncology by developing an Explainable Deep Learning Framework for the Segmentation and Classification of Brain Tumors. The core problem addressed is the high mortality rate associated with brain tumors, where early and accurate diagnosis is paramount, coupled with the inherent lack of transparency in current state-of-the-art deep learning models, a significant barrier to their acceptance and reliable deployment in clinical decision-making. To solve this, the framework integrates the predictive power of deep neural networks trained on high-resolution, multi-modal Magnetic Resonance Imaging (MRI) data with the interpretability of Explainable Artificial Intelligence (XAI). The primary objective is to move beyond simple, high-accuracy prediction scores by creating a system that not only precisely performs voxel-wise segmentation of tumor sub-regions, but also provides direct, feature-based justifications for every diagnostic classification made. This approach aims to validate model reasoning, boost clinical trust, and ensure that the framework can be relied upon by medical professionals for critical tasks such as surgical planning and treatment assessment.

2.2 Dataset

The project leverages the BraTS 2024 Glioma Dataset, which provides a standardized benchmark of patient scans. The ground-truth data, meticulously annotated by experts, defines four key classes for segmentation:

Table 2.2.1 Segmentation Classes

Class ID	Tumor Sub-Region	Clinical Importance
0	Background	Healthy brain parenchyma and non-tumor regions.
1	Necrotic Core	The dead or inactive centre of the tumor; crucial prognostic indicator.
2	Peritumoral Edema	Vasogenic fluid accumulation surrounding the tumor; defines the full extent of the pathological invasion.
3	Enhancing Tumor	The most metabolically active and aggressive component, defined by contrast uptake.

2.2.1 Multi-Modal MRI Sequences

The distinction between these sub-regions is only possible by analyzing the four co-registered, complementary MRI sequences provided for each patient, each offering unique tissue contrast:

Table 2.2.2 Dataset Modalities

Modality	Acronym	Image Contrast Contribution
T1-weighted native	T1n	Excellent anatomical detail; used as a baseline structure map.
T1-weighted contrast-enhanced	T1c	Primary modality for identifying the Enhancing Tumor (appears hyperintense) due to gadolinium leakage through the compromised Blood-Brain Barrier (BBB).
T2-weighted	T2w	Sensitive to most pathologies; provides high signal for edema and tumor tissue.
T2 Fluid-Attenuated Inversion Recovery	T2f	Suppresses the signal from normal Cerebrospinal Fluid (CSF), making the high-intensity Peritumoral Edema much more distinct from normal ventricular space.

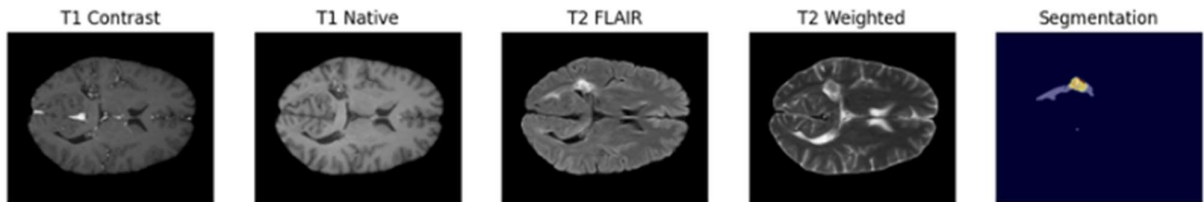


Fig 2.2.1 Dataset Modalities

2.3 Deep Learning Architectures

The project focuses on evaluating three deep learning models to find the optimal balance between performance and computational complexity for 3D volumetric segmentation.

2.3.1 2D UNet (Baseline)

The 2D UNet architecture serves as the baseline model. It features a symmetric encoder-decoder structure connected by skip connections. The encoder extracts feature representations at different

resolutions, and the decoder progressively up samples these features to produce the final segmentation mask. Skip connections merge high-resolution features from the encoder with the deeper, contextual features from the decoder, which is crucial for preserving fine spatial details in the segmentation boundaries.

```

1: Initialize parameters  $\theta$ 
2: for each epoch  $e = 1$  to  $E$  do
3:   for each batch  $(X_i, Y_i)$  do
4:      $\hat{Y}_i = f_{\theta}(X_i)$ 
5:      $\mathcal{L} = \mathcal{L}_{Dice}(\hat{Y}_i, Y_i) + \mathcal{L}_{CE}(\hat{Y}_i, Y_i)$ 
6:     Update  $\theta$  using Adam optimizer
7:   end for
8: end for

```

Algorithm 2.3.1 Train UNet

2.3.2 2D CNN + Transformer (Hybrid)

This hybrid model is designed to leverage the strengths of both CNNs and Transformers. The 2D CNN components handle local feature extraction (similar to the UNet encoder), while the Transformer blocks are integrated to model global, long-range dependencies across the slices and feature maps. This approach is computationally more efficient than a full 3D model while still capturing crucial non-local context.

```

1: Initialize parameters  $\theta_{enc}, \theta_{trans}, \theta_{dec}$ 
2: for each epoch do
3:   for each volume  $(X, Y)$  do
4:     Encode slices with CNN to obtain features  $f_s$ 
5:     Pool slice features to tokens  $t_s$ 
6:     Process tokens with transformer to obtain  $t'_s$ 
7:     Fuse  $t'_s$  with  $f_s$  and decode to segmentation  $\hat{Y}$ 
8:     Compute loss  $\mathcal{L} = \mathcal{L}_{Dice} + \mathcal{L}_{CE}$ 
9:     Update all parameters
10:  end for
11: end for

```

Algorithm 2.3.2 Train 2D Transformer

2.3.3 3D CNN + Transformer (Volumetric Hybrid)

The 3D CNN + Transformer is the most advanced architecture investigated. It extends the hybrid concept to a fully volumetric approach. The 3D CNN filters are used for localized feature extraction, capturing the spatial context in all three dimensions simultaneously. The 3D Transformer blocks then process these volumetric features to establish global consistency across the entire brain volume. This architecture is

hypothesized to be the most accurate for 3D medical images, as it maintains the integrity of the tumor's volumetric shape, though it comes with a higher computational overhead.

```

1: Initialize parameters  $\theta_{enc}, \theta_{trans}, \theta_{dec}$ 
2: for each epoch do
3:   for each volume  $(X, Y)$  do
4:     Extract features:  $F = \text{Encoder}_{3D}(X)$ 
5:     Tokenize patches:  $T = \text{Tokenize}(F)$ 
6:     Process tokens:  $T' = \text{Transformer}(T)$ 
7:     Reconstruct:  $F' = \text{UnTokenize}(T')$ 
8:     Predict:  $\hat{Y} = \text{Decoder}_{3D}(F')$ 
9:     Compute loss  $\mathcal{L} = \mathcal{L}_{Dice} + \mathcal{L}_{CE}$ 
10:    Update parameters
11:  end for
12: end for

```

Algorithm 2.3.3 Train 3D Transformer

2.4 Explainable AI (XAI) Framework

Deep learning models, particularly in medical imaging, are often criticized as "black boxes" because their decision-making process is not transparent to clinicians. This lack of interpretability can hinder adoption in critical domains like neuro-oncology, where understanding *why* a model predicts a tumor region is as important as what it predicts. The Neuro-XAI framework addresses this challenge by integrating complementary interpretability techniques that provide both visual and quantitative explanations. By revealing which features or regions influence model predictions, Neuro-XAI enhances clinical trust, facilitates error analysis, and ensures that the model's decisions are grounded in medically relevant imaging cues. Importantly, these methods support both local explanations (instance-level, voxel-specific) and global explanations (aggregated patterns across patients and modalities), bridging the gap between high-performance segmentation and clinical interpretability.

2.4.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is a widely used post-hoc interpretability technique that provides visual explanations of deep convolutional networks' predictions. It is particularly suitable for volumetric medical images because it highlights spatial regions that contribute most strongly to a specific output class. Grad-CAM works by computing the gradient of a target class score (e.g., Enhancing Tumor) with respect to the feature maps of the last convolutional layer in the network. This process effectively identifies which spatial features the model attends to when making predictions. The underlying idea is that regions with large positive gradients are most influential in increasing the class score, while negative gradients indicate regions that suppress the prediction. By weighting each feature map by its corresponding gradient and summing

across channels, Grad-CAM produces a coarse localization map showing the importance of each voxel for the target class.

Mathematically, the Grad-CAM heatmap $L_{Grad-CAM}^c$ for class c is computed as:

$$L_{Grad-CAM}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right)$$

where A^k represents the k -th feature map in the last convolutional layer, and α_k^c is the gradient-based weight computed as the global average of the gradient of the target score y^c with respect to A^k . The ReLU function ensures that only positively contributing features are visualized, which aligns with clinically meaningful attention regions.

Implementation

In the 3D CNN + Transformer model, the volumetric Grad-CAM heatmap is generated and overlaid onto the original MRI scan, highlighting tumor regions that strongly influence the prediction. This provides an intuitive visual explanation for clinicians, demonstrating not only the accuracy of segmentation but also the spatial reasoning underlying the model's decisions. By visualizing multi-modal contributions across T1, T1c, T2w, and T2-FLAIR sequences, Grad-CAM also helps confirm that the model integrates complementary imaging information effectively.

2.4.2 SHapley Additive exPlanations (SHAP)

While Grad-CAM provides a qualitative visual explanation, SHAP delivers a quantitative, voxel-level measure of feature importance, grounded in cooperative game theory. SHAP interprets the prediction of a model as a cooperative game, where each input feature (in this case, each voxel intensity) is a "player" contributing to the overall outcome. The Shapley value quantifies each feature's contribution by considering all possible subsets of features and measuring the marginal effect of including that feature. This ensures fair allocation of importance and satisfies properties like local accuracy, missingness, and consistency, which are critical for trustworthy explanations.

For a prediction $f(x)$ on input x with features x_1, x_2, \dots, x_n , the Shapley value ϕ_i for feature i is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)]$$

where S is any subset of features not including i , and N is the set of all features. This approach ensures

that each voxel's contribution to the model's prediction is fairly weighted relative to all other voxels, allowing for high-resolution, verifiable explanations.

Implementation

In the context of Neuro-XAI, SHAP is applied to each voxel across all four MRI modalities. The resulting SHAP values indicate the magnitude and direction (positive or negative) of each voxel's contribution to the target class confidence. For example, voxels corresponding to enhancing tumor regions in T1c scans have high positive contributions, while non-tumorous background voxels typically show minimal or negative contributions. This granular interpretability allows clinicians to verify that the model relies on medically meaningful features, detect potential biases, and compare voxel-level importance across modalities and patient scans. By complementing Grad-CAM, SHAP ensures that Neuro-XAI provides both visual and quantitative evidence for model predictions, thereby bridging the gap between deep learning performance and clinical trust.

2.5 Methodology

The methodology of this work is designed to build a complete end-to-end framework that transforms raw multi-modal MRI scans into clinically meaningful tumor segmentation maps enriched with explainability insights. It integrates a sequence of carefully structured stages including data acquisition, preprocessing, model design, training strategy, and explainability analysis, each contributing to the reliability, robustness, and interpretability of the system.

2.5.1 Data Acquisition and 3D Preprocessing Pipeline

This stage transforms the raw, multi-modal NIfTI (NII) files into standardized 5D tensor volumes suitable for 3D deep learning model input.

- **Data Loading and Modality Collection:** The data loading process begins by iterating through each patient directory in the BraTS dataset and identifying all four essential MRI sequences: T1-contrast Enhanced (T1c), T1-native (T1n), T2-FLAIR (T2f), and T2-weighted (T2w). Along with these modalities, the corresponding ground truth segmentation mask (-seg.nii) is also loaded. This mask contains voxel-wise annotations for four tumor-related classes: Background (0), Necrotic Core (1), Edema (2), and Enhancing Tumor (4), which are used as the target labels for training the segmentation models.
- **Intensity Normalization (Z-Score):** To ensure consistent intensity distributions across all patients and modalities, each of the four MRI volumes undergoes independent Z-score

normalization. The mean and standard deviation for normalization are computed only over non-zero voxels to avoid distortion due to the large background region present in brain scans. This step helps prevent disproportionately high-intensity areas from dominating the model’s learning process.

- **Label Alignment:** Since the original BraTS dataset uses the label value 4 for the Enhancing Tumor class, it is re-mapped to 3 to create a contiguous class range expected by standard multi-class segmentation frameworks. After this re-indexing, the dataset consists of four cleanly ordered classes: 0 for Background, 1 for Necrotic Core, 2 for Edema, and 3 for Enhancing Tumor.
- **3D Resampling and Sizing:** All modality volumes and their corresponding segmentation masks are resampled to a uniform size of (32, 128, 128) to reduce computational requirements and maintain consistent input shapes across samples. The image modalities (T1c, T1n, T2f, and T2w) are resized using trilinear interpolation to preserve smooth intensity transitions, while the segmentation mask is resized using nearest neighbour interpolation to avoid generating invalid fractional class values.
- **Final Tensor Structure and Split:** After preprocessing, the four normalized and resampled modalities are stacked along the channel dimension and permuted into the PyTorch-compatible format (Channels, Depth, Height, Width), resulting in a tensor of shape (4, 32, 128, 128). The complete set of processed patient volumes is then divided into an 80% training set and a 20% validation set to facilitate supervised learning and performance evaluation.
- **Data Augmentation:** A simple yet effective on-the-fly augmentation strategy is applied during training to enhance the generalization capability of the model. This includes a 50% probability of applying an axial-plane rotation (randomly chosen from 0°, 90°, 180°, or 270°) and a 50% chance of flipping the volume along the width axis. These augmentations introduce spatial variability without altering the anatomical structure, helping the model become more robust to variations in orientation and viewpoint.

2.5.2 Training Strategy and Configuration

The training regimen employs a sophisticated loss function and an adaptive optimization strategy to tackle the challenges of 3D segmentation and severe class imbalance.

- **Hybrid Loss Function:** The model utilizes a combined objective function to leverage the strengths of classification-based and region-similarity-based losses.
- **Weighted Cross-Entropy (CE) Loss:** Calculates the standard voxel-wise CE loss.
- **Weighting:** This loss is weighted inversely by the frequency of each class in the dataset. This mechanism ensures that errors in the small tumor sub-regions (like Necrotic Core or Enhancing Tumor) contribute significantly more to the total loss than errors in the large Background/Edema regions. (Example weights: Enhancing Tumor at 1.2, Background at 0.2).
- **Class-wise Dice Loss:** A metric that measures the overlap (similarity) between the predicted and ground truth tumor regions. It is calculated per class and then averaged, compelling the model to optimize for better spatial fit and boundary prediction for each tumor component.
- **Final Combined Loss:** The total training objective is a weighted sum of the two components:

$$\text{Total Loss} = 0.6 \times \text{Weighted Cross-Entropy Loss} + 0.4 \times \text{Class-wise Dice Loss}$$

2.5.3 Optimization and Learning Rate Scheduling

- **Optimizer:** AdamW is employed, which is a variation of Adam that decouples weight decay from the gradient updates, leading to improved generalization.
 Initial Learning Rate: 1×10^{-4} .
- **Weight Decay:** A small penalty (e.g., 1×10^{-5}) is applied to weights to prevent overfitting.
- **Scheduler:** The ReduceLROnPlateau scheduler is used to dynamically adjust the learning rate based on performance. It monitors the Validation Loss across epochs. If the loss fails to improve after a set number of epochs, the learning rate is automatically reduced by a factor (e.g., multiplied by 0.1), allowing the model to escape local minima.
- **Training Efficiency:** Automatic Mixed Precision (AMP) is used for efficient training. The training loop incorporates `torch.amp.autocast` and `torch.amp.GradScaler`. This technique automatically uses lower-precision floating-point formats (e.g., float16) for operations where possible. This significantly reduces GPU memory consumption and accelerates training speed.

with minimal impact on model accuracy.

2.5.4 Explainability (XAI) Methods

To enhance model transparency and provide clinically meaningful insights, post-training explainability techniques are applied to analyse voxel-level contributions that influence the prediction of specific tumor classes, such as the Necrotic Core. These methods generate attribution maps that reveal which spatial regions or modalities most strongly drive the segmentation output, thereby offering interpretability without altering the core model architecture.

- **SHAP (SHapley Additive exPlanations):** SHAP is used to estimate voxel-level feature importance by computing Shapley values that fairly distribute the model's output among the input features. A custom 3D SHAP Explainer built on the Gradient Explainer computes gradients of the class-specific logit with respect to the input volume, producing a dense 3D attribution map. The sign and magnitude of each voxel's value reflect how strongly it supports or opposes the prediction of the target tumor region, enabling both global and local interpretability across the modalities and spatial dimensions.
- **Grad-CAM for 3D Volumes:** Grad-CAM is employed to visualize class-specific activation regions by using the gradients flowing into the final convolutional layers. For 3D networks, Grad-CAM is extended volumetrically to capture salient regions within the depth of the MRI stack. The generated heatmaps highlight anatomically relevant features that significantly influence the segmentation boundaries, particularly for tumor subregions with subtle intensity variations.
- **Local Fitting–Based Interpretability:** A local fitting strategy is used to provide instance-level explanations by approximating the model's behaviour around a specific input. The method perturbs local neighbourhoods within the 3D volume and analyses how these changes influence the predicted tumor class. By fitting a simplified surrogate model around these perturbations, the method identifies locally influential voxel clusters, revealing fine-grained spatial cues that the model relies upon during decision-making.

Chapter 3: RESULTS & DISCUSSION

3.1 Training Environment

All three models 2D UNet, 2D CNN + Transformer, and 3D CNN + Transformer were trained and validated on the pre-processed BraTS 2024 Glioma Dataset. The training objective was minimized using a hybrid loss function combining Dice Loss and weighted Cross-Entropy Loss, which effectively addresses the severe class imbalance inherent in tumor segmentation (where tumor voxels are significantly fewer than normal brain voxels). The Adam optimizer was employed for network weight updates due to its ability to adaptively adjust learning rates and accelerate convergence in deep networks. Training was conducted on high-performance GPUs, leveraging mixed-precision computation to reduce memory usage while maintaining numerical stability. Each model underwent extensive hyperparameter tuning, including learning rate scheduling, batch size optimization, and early stopping criteria to prevent overfitting. Model performance was rigorously evaluated on a dedicated, unseen test set to quantify generalization capabilities. Additionally, a consistent data augmentation strategy, including random rotations and flips, ensured robustness against spatial variability and anatomical orientation differences in the MRI volumes. This setup not only guarantees reproducibility but also ensures that the subsequent results accurately reflect the models' inherent capability rather than dataset-specific artifacts.

3.2 Quantitative Evaluation of Model Performance

The performance of the models was primarily assessed using standard segmentation metrics: Dice Score, Accuracy, and Intersection over Union (IoU). The results below showcase the consistent improvement gained by incorporating advanced architectures.

Table 3.2.1 Quantitative Results

Model	Dice Score (Average)	Accuracy	IoU (Intersection over Union)
2D UNet (Baseline)	0.60	0.940	0.795
2D CNN + Transformer	0.758	0.963	0.867
3D CNN + Transformer	0.82	0.99	0.90

From the table, it is evident that introducing global attention via the Transformer module significantly

improves segmentation performance. The 2D CNN + Transformer outperforms the baseline 2D UNet, highlighting the importance of capturing long-range dependencies across slices. The 3D CNN + Transformer achieves the highest performance across all metrics, demonstrating that volumetric modeling better preserves the spatial continuity of tumor structures.

3.3 Analysis of Segmentation Performance

The quantitative results demonstrate that the architectural advancements effectively address the complexity of 3D brain tumor segmentation:

3.3.1 Impact of Global Attention

The improvement in Dice Score from 0.60 (2D UNet) to 0.758 (2D CNN + Transformer) underscores the value of the Transformer's self-attention mechanism. By considering dependencies across the entire 2D slice, the model can integrate contextual information, enabling it to correctly segment challenging regions such as diffuse edema and irregular necrotic cores. This confirms that attention-based architectures are particularly beneficial in medical imaging tasks where local convolutional features may be insufficient to capture complex structures.

3.3.2 Superiority of Volumetric Modelling

The 3D CNN + Transformer model achieved the highest metrics: a Dice Score of 0.82, Accuracy of 0.99, and IoU of 0.90. Processing the MRI data in its native 3D volumetric form preserves spatial continuity, which is essential for accurately segmenting heterogeneous tumor subregions. Volumetric modelling also allows the network to leverage inter-slice information, reducing misclassification between adjacent slices and improving the delineation of tumor boundaries. This approach is particularly effective for tumors with irregular shapes and infiltrative edges, where 2D slice-based models might miss subtle cross-sectional features. Additionally, the combination of CNN and Transformer in 3D space ensures both local feature extraction (via convolution) and global context modelling (via self-attention), making the architecture robust to variations in tumor size, location, and morphology. These results suggest that for glioma segmentation, volumetric attention-based architectures provide the most reliable predictions.

3.3.3 Clinical Implications

The high Dice Scores and Accuracy indicate that automated segmentation can approach the precision required for clinical support. Accurate tumor delineation enables more precise radiation planning, surgical guidance, and treatment monitoring. Moreover, the robustness across multiple patients and

tumor subtypes suggests that this model could reduce inter-observer variability, a common challenge in neuro-oncology.



Fig.3.3.1 Segmentation Results

3.4 Explainability: Results and Discussion

The integration of Explainable AI (XAI) techniques into the high-performing 3D CNN + Transformer model is the core of the Neuro-XAI framework, transforming high accuracy into clinical reliability.

3.4.1 Grad-CAM

- **Visualization:** Grad-CAM generated spatial attention heatmaps highlighting regions most critical for class-specific segmentation. The method consistently emphasized enhancing tumor rims on T1c, edema on T2-FLAIR, and necrotic cores on T1n/T2w. These maps captured complex tumor morphology, including fragmented or irregular subregions, reflecting the model's ability to detect heterogeneous patterns. The volumetric Grad-CAM also visualized inter-slice dependencies, providing a comprehensive 3D understanding of tumor attention.
- **Discussion:** These visualizations allow radiologists to verify that the model attends to the correct anatomical structures rather than irrelevant artifacts. In addition, subtle attention to surrounding edema or infiltrative regions provides clinically useful information about tumor spread. Detecting instances where attention is misplaced enables error analysis and iterative model improvement. Overall, Grad-CAM enhances transparency, helping bridge the gap between model performance and clinical trust.

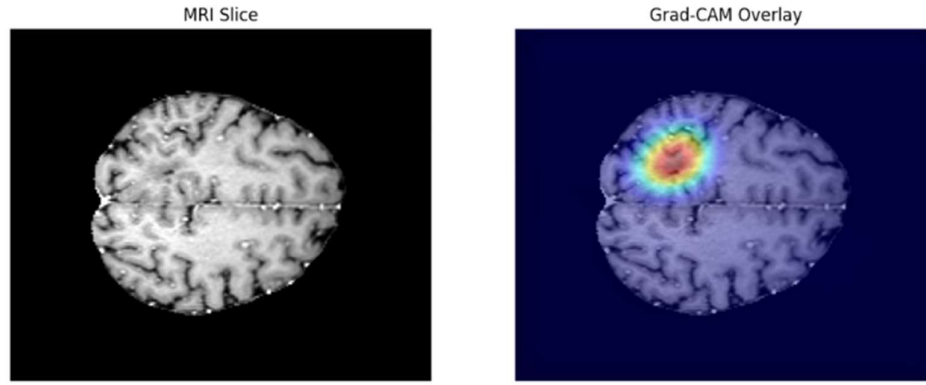


Fig.3.4.1 Grad-CAM Visualisation

3.4.2 SHAP Quantitative Analysis

SHAP (SHapley Additive exPlanations) provided a model-agnostic, quantitative measure of feature importance.

- **Quantification:** SHAP computed voxel-level contributions across all four modalities. This analysis revealed that T1c intensities drive enhancing tumor predictions, T2-FLAIR intensities guide edema detection, and T1n/T2w intensities assist in defining necrotic regions. The signed contributions indicate whether a voxel positively or negatively influenced a class prediction, providing precise, quantitative interpretability.
- **Discussion (Model Fidelity):** SHAP confirms that the model leverages medically relevant features and does not rely on spurious patterns. High SHAP values in contrast-enhanced regions validate biologically meaningful reasoning, while subtle attributions in peripheral edema indicate early infiltration detection. Cross-patient consistency further confirms generalizable learning. This quantitative explainability complements Grad-CAM visualizations, collectively ensuring that Neuro-XAI is both accurate and interpretable for clinical applications.

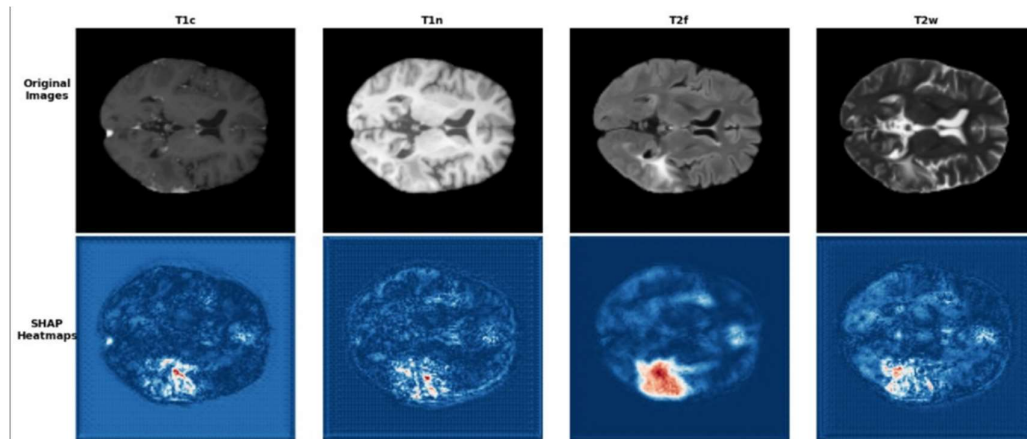


Fig.3.4.2 SHAP Analysis

Chapter 4: CONCLUSION

This project successfully developed and evaluated Neuro-XAI, an explainable deep learning framework designed for the precise segmentation and classification of gliomas from multimodal MRI data, achieving all the defined objectives. The core of the work involved a comparative performance analysis of three distinct deep learning architectures on the BraTS 2024 dataset: 2D UNet, 2D CNN + Transformer, and 3D CNN + Transformer. The results conclusively demonstrated that the 3D CNN + Transformer architecture achieved the highest performance metrics, including a superior average Dice Score of 0.82 and an overall accuracy of 99%, validating the necessity of a volumetric approach capable of effectively modeling the spatial continuity and 3D context inherent in medical image data. A key contribution of the project is the seamless integration of Explainable AI (XAI) techniques, specifically Grad-CAM and SHAP, which address the critical challenge of the "black-box" dilemma in medical AI. Grad-CAM provides essential visual confirmation by mapping the model's focus directly onto tumor regions, while SHAP offers quantitative justification by measuring the precise contribution of each input voxel to the final prediction. By combining high predictive accuracy with verifiable transparency, Neuro-XAI transforms a powerful deep learning model into a clinically trustworthy and interpretable diagnostic tool, paving the way for its responsible adoption in neuro-oncology.

References

1. E. Gundogan, “A Novel Hybrid Deep Learning Model Enhanced with Explainable AI for Brain Tumor Multi-Classification from MRI Images,” *Applied Sciences*, vol. 15, no. 5, p. 5412, 2025.
2. Q. U. A. Mastoi, A. A. Jan, M. U. Kakar, and Y. Kim, “Explainable AI in medical imaging: an interpretable and collaborative federated learning model for brain tumor classification,” *Frontiers in Oncology*, vol. 14, 2025.
3. K. M. Hosny and A. M. Mohammed, “Explainable AI and vision transformers for detection and classification of brain tumor: a comprehensive survey,” *Artificial Intelligence Review*, 2025.
4. M. A. Musthafa, P. M. A. Muneer, M. S. Hameed, and M. A. Mohammed, “Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet 50,” *BMC Medical Imaging*, vol. 24, p. 1292, 2024.
5. R. Iftikhar, M. S. Hameed, and M. Y. Javed, “Explainable CNN for brain tumor detection and classification through XAI-based key features identification,” *Brain Informatics*, vol. 12, p. 257, 2025.
6. E. Vamsidhar, P. Kumar, and B. S. Reddy, “Hybrid model integration with explainable AI for brain tumor diagnosis: a unified approach to MRI analysis and prediction,” *Scientific Reports*, vol. 15, p. 6455, 2025.
7. M. Menze, et al., “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.
8. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Proc. MICCAI*, pp. 234–241, 2015.
9. F. Isensee et al., “nnU-Net: a self-adapting framework for biomedical image segmentation,” *Nature Methods*, vol. 18, pp. 203–211, 2021.
10. R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” in *Proc. ICCV*, pp. 618–626, 2017.