Here Comes the Explanation: A Shapley Perspective on Multi-contrast Medical Image Segmentation*

Tianyi Ren[1], Juampablo Heras Rivera[1], Hitender Oswal[2], Yutong Pan[2], Agamdeep Chopra[1], Jacob Ruzevick[3] and Mehmet Kurt[1]

[1]*Department of Mechanical Engineering, University of Washington, 3900 E Stevens Way NE, Seattle, WA 98195*
[2]*Paul G. Allen School of Computer Science, University of Washington, 185 E Stevens Way NE Seattle, WA 98195*
[3]*Department of Neurological Surgery, University of Washington, 1959 NE Pacific Street, Seattle, WA 98195*

Abstract

Deep learning has been successfully applied to medical image segmentation, enabling accurate identification of regions of interest such as organs and lesions. This approach works effectively across diverse datasets, including those with single-image contrast, multi-contrast, and multimodal imaging data. To improve human understanding of these black-box models, there is a growing need for Explainable AI (XAI) techniques for model transparency and accountability. Previous research has primarily focused on post hoc pixel-level explanations, using methods gradient-based and perturbation-based approaches. These methods rely on gradients or perturbations to explain model predictions. However, these pixel-level explanations often struggle with the complexity inherent in multi-contrast magnetic resonance imaging (MRI) segmentation tasks, and the sparsely distributed explanations have limited clinical relevance. In this study, we propose using contrast-level Shapley values to explain state-of-the-art models trained on standard metrics used in brain tumor segmentation. Our results demonstrate that Shapley analysis provides valuable insights into different models' behavior used for tumor segmentation. We demonstrated a bias for U-Net towards over-weighing T1-contrast and FLAIR, while Swin-UNETR provided a cross-contrast understanding with balanced Shapley distribution.

Keywords

Image Segmentation, XAI, Shapley Value, MRI, Brain Tumor

1. Introduction

Segmentation is a fundamental task in medical imaging, involving identifying regions of interest (ROIs) such as organs, lesions, and tissues. By precisely outlining anatomical and pathological structures, segmentation plays a pivotal role in computer-aided diagnosis, ultimately improving diagnostic precision [1]. Typically, segmentations task are carried out using multi-contrast MRI or multi-modal imaging datasets, due to the necessity of identifying unique microstructural features, such as in gliomas [2], that are only apparent in some MRI contrasts, but not others. Many deep learning models, including those used for segmentation, are considered black boxes, offering limited interpretability, resulting in a lack of transparency and accountability [3]. Various Explainable AI (XAI) techniques have been developed in the literature [4] to tackle this problem, primarily categorized into gradient-based and perturbation-based methods.

Gradient-based techniques, such as saliency maps [5] and Grad-CAM [6], visualize deep learning predictions by identifying influential regions in input data, while perturbation-based approaches (Shapley values [7] and LIME [8]) observe model behavior by systematically perturbing inputs and measuring impact. These methods have been applied successfully to explain the classification problem, however, explaining segmentation still presents significant challenges. There is ongoing debate about whether explanations are necessary for segmentation, as the masks themselves may serve as explanations. Furthermore, there remains uncertainty regarding which components should be explained—when using

Late-breaking work, Demos and Doctoral Consortium, colocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

tr1@uw.edu (T. Ren); jehr@uw.edu (J. H. Rivera); hitender@uw.edu (H. Oswal); ypan4@cs.washington.edu (Y. Pan); achopra4@uw.edu (A. Chopra); ruzevick@uw.edu (J. Ruzevick); mkurtr@uw.edu (M. Kurt)
0000-0001-9548-6645 (T. Ren); 0000-0002-0205-6329 (J. H. Rivera); 0000-0002-5618-0296 (M. Kurt)

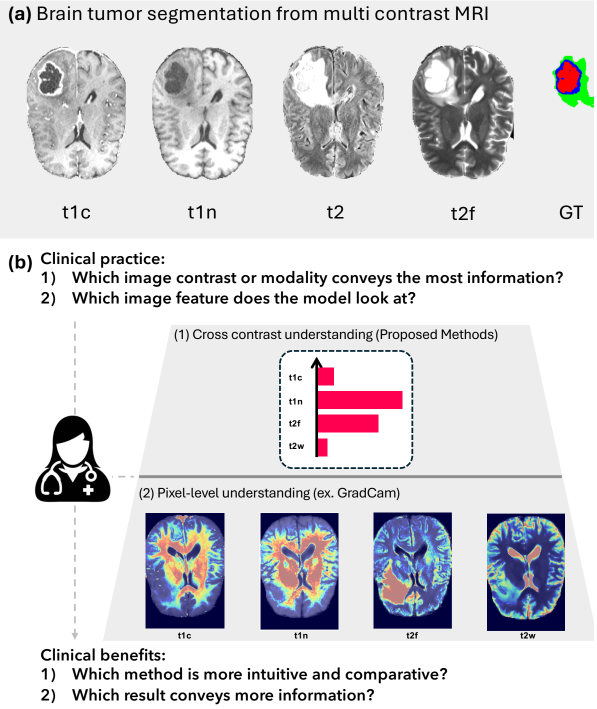

(a) Brain tumor segmentation from multi contrast MRI
t1c
t1n
t2
t2f
GT
(b) Clinical practice:
1) Which image contrast or modality conveys the most information?
2) Which image feature does the model look at?
(1) Cross contrast understanding (Proposed Methods)
t1c
t1n
t2f
t2w
(2) Pixel-level understanding (ex. GradCam)
t1c
t1n
t2f
t2w
Clinical benefits:
1) Which method is more intuitive and comparative?
2) Which result conveys more information?

Figure 1: (a) An example of tumor segmentation from multi-contrast MRI. The decision process is not always intuitive because the model does not explain which contrast contributes to the decision, as redundant information can be observed between image contrasts. (b) Our proposed Contrast-level shapley value aims to provide a cross-contrast level explanation which provides a global understanding of the multi-contrast image segmentation.

gradient-based approaches for models like U-Net, no consensus exists on which layer to target, and in clinical application, which MRI contrasts to explain. Moreover, pixel-level explanations, typically represented as discretized heatmap maps, require further interpretation for grouping analysis [9].

Since in clinical practice radiologists detect lesions by analyzing differences between different MRI contrasts [2], an explainability framework that reveals deep learning model behavior with regards to how they weigh different MRI contrasts in the segmentation process would be immediately clinically relevant. Therefore, the main objective of this paper is to establish a framework for explaining the contributions of different MRI contrasts in the segmentation process with an application in brain tumor segmentation. This method delivers intuitive quantitative model explanations and enables effective comparisons at multiple levels: between contrasts within a subject (see Figure 4), and between model architectures for comprehensive model behavior interpretation (see Section 3). We perform systematic experiments to explain how the state-of-the-art models such as U-Net and Transformer (Swin-UNETR) weigh different MRI contrasts with respect to different evaluation metrics such as Dice and HD95. We conduct statistical analyses to provide an in-depth understanding of how and why different model architectures weigh MRI contrasts differently, even when they achieve similar segmentation performance. In summary, our paper, to the best of our knowledge, is the first study to propose a clinically-relevant explanation framework for brain tumor segmentation in multi-contrast MRI.

2. Methods

2.1. Dataset and Learning Objectives

The training dataset is sourced from the Brain Tumor Segmentation (BraTS) Challenge 2024 GoAT challenge [10], consisting of 1,351 subjects. For each subject, four MRI contrasts were given: Native ($t1n$), Post-contrast T1-weighted ($t1c$), T2-weighted ($t2w$), and T2 Fluid Attenuated Inversion Recovery ($t2f$). The ground truth annotations consist of three disjoint classes: Enhancing tumor (ET), Peritumoral

Table 1
Comparison of Dice Scores and HD95 Metrics for Different Models

Model	Dice Score [-]				HD95 [mm]			
	NCR	ET	ED	Avg	NCR	ET	ED	Avg
U-Net	70.33%	81.26%	84.79%	78.79%	6.99	5.10	4.56	5.55
Segresnet	69.88%	80.30%	84.16%	78.11%	7.57	7.46	5.04	6.69
UNETR	69.45%	80.55%	83.95%	77.98%	7.38	6.24	5.22	6.28
Swin-UNETR	69.32%	81.29%	85.25%	78.62%	7.38	5.60	5.21	6.06

edematous tissue (ED), and Necrotic tumor core (NCR). The detailed preprocessing and training pipeline can be found in our previous research [11, 12].

2.2. Model Architectures and Evaluating Metric

Several state-of-the-art model architectures are tested in this study, including **U-Net** [13], **Seg-Resnet** [14], **UNETR** [15], and **Swin-UNETR** [16]. To evaluate the segmentation quality, we used common metrics, including the Dice coefficient and the 95th percentile Hausdorff distance (HD95).

2.3. Contrast Level Shapley Value

Given a training dataset comprised of the pairs $\{(I, x_0)\}_{i=1}$, where $I \in \mathbb{R}^{4\times D\times W\times H}$ represents the four 3D-MRI contrast as a multi-channel input, $x_0 \in \mathbb{R}^{3\times D\times W\times H}$ represents the associated one-hot encoded segmentation mask, with 3 tumor labels: ED, NCR, and ET as described in Section 2.1. The deep learning models (ω) were trained to predict the tumor labels $\hat{x}_0$ given the input I:

$$\hat{x}_0 = \omega(I). \tag{1}$$

Derived from the Shapley value [7]. The Contrast level Shapley value $\phi_i(M)$ was then evaluated with respect to each specific metric (M) by:

$$\phi_i(M) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} \left(M(S\cup\{i\}) - M(S)\right) \tag{2}$$

where N is the set of all of MRI contrasts; $|N|$ is the total number of contrasts; S is a subset of MRI contrasts excluding certain contrast i ($S \subseteq N \setminus \{i\}$) and $|S|$ is the number of contrasts in S; $M(S)$ is the target metric evaluated on the subset S.

The contrast-level Shapley values are examined to assess whether observed differences(group means and variances) across folds or between models are statistically significant. **Test for equal variance**: Levene's test is applied to assess homogeneity of variance even when the normality assumption cannot be guaranteed. **Test for equal mean**: If the normality assumption cannot be guaranteed, the Kruskal-Wallis test is used instead of ANOVA, and Dunn's test is applied for post-hoc analysis instead of Tukey's test. **Confidence interval of the difference**: If a significant difference in means is observed, we further generate the confidence interval of the mean difference between groups when the normality assumption is not violated.

3. Experiments and Results

Table 1 presents a comparative analysis of model performance. The results demonstrate that all models achieve similar performance in terms of Dice scores and HD95 across all three labels, with U-Net marginally outperforming transformer-based models (Swin-UNETR and UNETR) and the Segresnet model.

Next, contrast-level Shapley values for each metric, averaged over three labels, are computed using four model architectures across five data folds. We define the matrix of contrast-level Shapley values for each combination of metric $M \in \{\text{Dice}, \text{HD95}\}$, model $\omega \in \{\text{U-Net, SegResNet, UNETR, Swin-UNETR}\}$, and fold $f = 1, \ldots, 5$ as:

$$\mathbf{\Phi}^{\omega,f}(M) = \begin{pmatrix} \phi^{\omega,f}_{t1n,1}(M) & \phi^{\omega,f}_{t1n,2}(M) & \cdots & \phi^{\omega,f}_{t1n,J_f}(M) \\ \phi^{\omega,f}_{t1c,1}(M) & \phi^{\omega,f}_{t1c,2}(M) & \cdots & \phi^{\omega,f}_{t1c,J_f}(M) \\ \phi^{\omega,f}_{t2w,1}(M) & \phi^{\omega,f}_{t2w,2}(M) & \cdots & \phi^{\omega,f}_{t2w,J_f}(M) \\ \phi^{\omega,f}_{t2f,1}(M) & \phi^{\omega,f}_{t2f,2}(M) & \cdots & \phi^{\omega,f}_{t2f,J_f}(M) \end{pmatrix}, \mathbf{\Phi}^{\omega,f}(M) \in \mathbb{R}^{4,J_f}, \tag{3}$$

where $\phi^{\omega,f}_{i,j}(M)$ represents the Shapley value for the j-th subject in fold f, given contrast i, model ω, and metric M. We use J_f to denote the total number of subjects in fold f.

For a given combination (M, ω, f), the contrast-wise vector $\mathbf{C}^{\omega,f}_i(M)$ ($i \in \{\text{t1n, t1c, t2w, t2f}\}$) and subject-wise vector $\mathbf{S}^{\omega,f}_j(M)$ ($j = 1, \ldots, J_f$) are defined as follows:

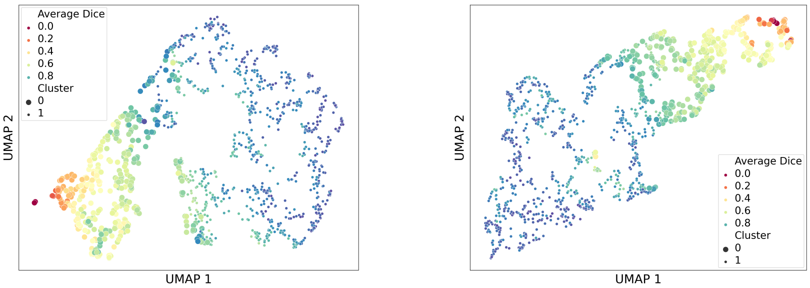
$$\begin{aligned} \mathbf{C}^{\omega,f}_i(M) = \mathbf{\Phi}^{\omega,f}_{i,\cdot}(M) = \left(\phi^{\omega,f}_{i,1}(M),\ \phi^{\omega,f}_{i,2}(M), \cdots,\ \phi^{\omega,f}_{i,J_f}(M)\right), \mathbf{C}^{\omega,f}_i(M) \in \mathbb{R}^{J_f} \\ \mathbf{S}^{\omega,f}_j(M) = \mathbf{\Phi}^{\omega,f}_{\cdot,j}(M) = \left(\phi^{\omega,f}_{t1n,j}(M),\ \phi^{\omega,f}_{t1c,j}(M), \phi^{\omega,f}_{t2w,j}(M), \phi^{\omega,f}_{t2f,j}(M)\right)^T, \mathbf{S}^{\omega,f}_j(M) \in \mathbb{R}^4 \end{aligned} \tag{4}$$

In this study, we utilized four NVIDIA A40 GPUs to train our deep learning model and calculate the Shapley value. The evaluation time for each fold and model is approximately 1–2 minutes per subject.

3.1. Shapley-based prediction insights: a clustering analysis

To analyze how segmentation performance overlaps with model weighting of MRI contrasts via contrast-level Shapley values, we applied k-means clustering. For each model-metric pair (M, ω), clustering was performed on the $\mathbf{S}^{\omega,f}_j(M)$ across five folds, i.e., $\cup_{f=1}^{5} \cup_{j=1}^{J_f} \{\mathbf{S}^{\omega,f}_j(M)\}$.

We then use UMAP to visualize the clusters of Shapley value embeddings. Figure 2 illustrates an example with a significant pattern. For U-Net and Swin-UNETR, Shapley embedding clusters differentiate subjects with higher Dice scores from those with lower Dice scores.



(a) UNETR

(b) SEGRES

Figure 2: Clustering results on (a) $\cup_{f=1}^{5} \cup_{j=1}^{J_f} \{\mathbf{S}^{\text{Unet},f}_j(\text{Dice})\}$ and (b) $\cup_{f=1}^{5} \cup_{j=1}^{J_f} \{\mathbf{S}^{\text{Swin-UNETR},f}_j(\text{Dice})\}$ are visualized using UMAP for dimensionality reduction. The color represents the Dice score; the size of the dot is used to differentiate between cluster labels.

3.2. Shapley-based model prediction consistency: a comparative analysis

3.2.1. Does each model learn consistent explanations?

To assess the consistency of explanations across folds for each model, we analyzed the distribution of $\mathbf{C}^{\omega,f}_i(M)$. The group standard deviation σ and mean μ are key factors for determining distribution

Table 2

Post-hoc tests reveal the pairs of folds where no statistical difference exists.

$H_0(\mu\|_{\text{Dice, t1c, SU, (1,5)}})$	$H_0(\mu\|_{\text{Dice, t1c, SU, (2,3)}})$	$H_0(\mu\|_{\text{Dice, t1c, SU, (2,4)}})$	$H_0(\mu\|_{\text{Dice, t1c, SU, (4,5)}})$	All other tests
$p = 0.0385$	$p = 0.0442$	$p = 0.0687$	$p = 0.0107$	$p < 0.01$

*Note that, we abbreviate Swin-UNETR as SU in this table.

similarity, and statistical tests were applied to these metrics:

$$\begin{aligned} H_0(\sigma|M,i,\omega) &: \sigma(\mathbf{C}_i^{\omega,1}(M)) = \sigma(\mathbf{C}_i^{\omega,2}(M)) = \sigma(\mathbf{C}_i^{\omega,3}(M)) = \sigma(\mathbf{C}_i^{\omega,4}(M)) = \sigma(\mathbf{C}_i^{\omega,5}(M)), \\ H_0(\mu|M,i,\omega) &: \mu(\mathbf{C}_i^{\omega,1}(M)) = \mu(\mathbf{C}_i^{\omega,2}(M)) = \mu(\mathbf{C}_i^{\omega,3}(M)) = \mu(\mathbf{C}_i^{\omega,4}(M)) = \mu(\mathbf{C}_i^{\omega,5}(M)). \end{aligned} \tag{5}$$

If significant differences in mean or standard deviation are found, we conclude that inconsistent explanations are present across folds for a given pair of (M, i, ω).

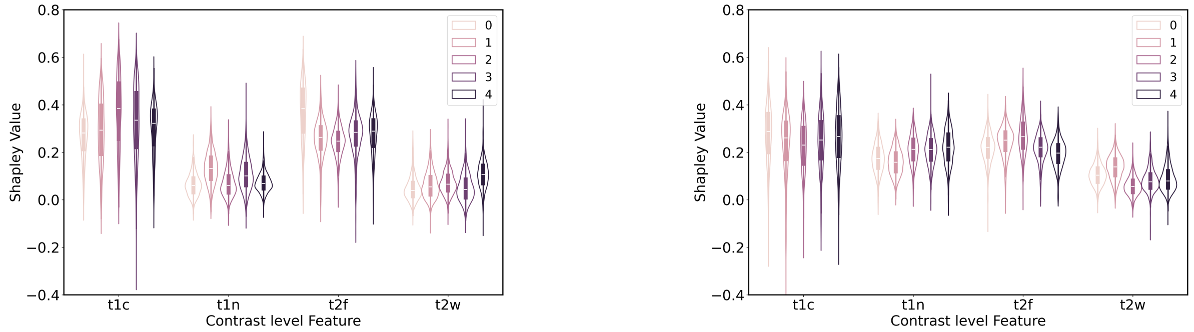
Since the normality assumption for the Shapley value distribution $\mathbf{C}_i^{\omega,f}(M)$ could not be guaranteed for some contrasts i, as indicated by the normality tests and non-zero skewness (Figure 3), Levene's test, Kruskal-Wallis, and Dunn's post-hoc tests were applied.

For all combinations of (M, i, ω), we get $p < 0.01$ in all 32 Levene's tests, rejecting $H_0(\sigma|M, i, \omega)$ and indicating unequal variances across the five folds. Similarly, all 32 Kruskal-Wallis tests yield $p < 0.01$, rejecting $H_0(\mu|M, i, \omega)$ and suggesting unequal means. These results invalidate the assumption that "Model ω learns consistent explanations across all five folds using contrast i for metric M evaluation," indicating significant differences in variance and means for at least one fold pair of each (M, i, ω) combination.

Post-hoc tests are conducted to evaluate which pairs (f_j, f_j') show consistency explanation with the following null hypothesis:

$$H_0(\mu|M, i, \omega, (f_j, f_{j'})) : \mu(\mathbf{C}_i^{\omega,f_j}(M)) = \mu(\mathbf{C}_i^{\omega,f_{j'}}(M)), f_j, f_{j'} \in \{1, 2, \cdots, 5\}; f_j \neq f_{j'}. \tag{6}$$

Dunn's post-hoc tests reveal no significant differences in the $t1c$ explanation between fold pairs 1 & 5, 2 & 3, 2 & 4, and 4 & 5 for Swin-UNETR, while significant differences exist in all other tests (Table 3). For example, in Table 3, $p = 0.038$ in the 1^{st} column, the null hypothesis $\mu(\mathbf{C}_{t1c}^{\text{Swin-UNETR},1}(Dice)) = \mu(\mathbf{C}_{t1c}^{\text{Swin-UNETR},5}(Dice))$ is not rejected, indicating "Swin-UNETR learns consistent $t1c$ contrast-level explanations between the 1st and 5th folds."



(a) U-net

(b) Swin-UNETR

Figure 3: The contrast-level Shapley values for all folds are computed based on the Dice score in each model. Panels (a) and (b) display the case of UNet and Swin-UNETR models, respectively.

3.2.2. Do different models learn consistent explanations?

We first visualize the contrast-level Shapley value across all five folds for U-net, $\mathbf{C}_i^{\text{U-net},f}(\text{Dice})$, and Swin-UNETR, $\mathbf{C}_i^{\text{Swin-UNETR},f}(\text{Dice})$, using violin plot in Figure 3. We could observe that $t1c$ and $t2f$

Table 3

Confidence Interval for Model Difference. The results indicate that Swin-UNETR exhibits significantly higher $t1n$ Shapley values compared to all other models for the Dice score at a 95% confidence level.

	$f = 1$	$f = 2$	$f = 3$	$f = 4$	$f = 5$
$CI_{0.95}(\mu(\mathbf{C}_{t1n}^{\text{(SU, U)},f}(\text{Dice})))$	[0.11,0.12]	[0.02,0.03]	[0.14,0.15]	[0.09,0.10]	[0.15,0.16]
$CI_{0.95}(\mu(\mathbf{C}_{t1n}^{\text{(SU, S)},f}(\text{Dice})))$	[0.05,0.06]	[0.09,0.11]	[0.06,0.07]	[0.01,0.02]	[0.07,0.08]
$CI_{0.95}(\mu(\mathbf{C}_{t1n}^{\text{(SU, UR)},f}(\text{Dice})))$	[0.06,0.07]	[0.00,0.01]	[0.16,0.17]	[0.03,0.06]	[0.11,0.12]

*Note that, we abbreviate Swin-UNETR as SU, U-Net as U, SegResNet as S, and UNETR as UR in this table.

are the most important image contrasts with the highest contrast-level Shapley value, this finding is consistent with the clinical explanation where $t2f$ suppresses cerebrospinal fluid signal, making edema and infiltration more visible, while $t1c$ provides clear delineation of enhancing tumor (see section 2.1). We can also observe from this figure that Swin-UNETR weights $t1n$ significantly higher than U-Net.

To further investigate how model explanations are different within folds, we follow the procedure from Section 3.2.1, with the key difference being that we compare results across multiple models while fixing the fold, unlike the previous tests where the models were fixed:

$$\begin{aligned} H_0(\sigma|M,i,f) &: \sigma(\mathbf{C}_i^{\text{U-Net},f}(M)) = \sigma(\mathbf{C}_i^{\text{Segresnet},f}(M)) = \sigma(\mathbf{C}_i^{\text{UNETR},f}(M)) = \sigma(\mathbf{C}_i^{\text{Swin-UNETR},f}(M)), \\ H_0(\mu|M,i,f) &: \mu(\mathbf{C}_i^{\text{U-Net},f}(M)) = \mu(\mathbf{C}_i^{\text{Segresnet},f}(M)) = \mu(\mathbf{C}_i^{\text{UNETR},f}(M)) = \mu(\mathbf{C}_i^{\text{Swin-UNETR},f}(M)). \end{aligned} \tag{7}$$

For all combinations of (M, i, f), the assumption that "Within each fold f, all models learned consistent explanations when using contrast i for metric M" is invalid [Levene's test ($p < 0.01$), Kruskal-Wallis test ($p < 0.01$) for all tests]. However, the post-hoc tests do not reveal generalizable patterns across the models similar to the conclusion we presented in Table 3. To highlight performance differences, we provide the confidence intervals.

Since the distributions of Shapley values are independent across models, and for each input j, the differences between Shapley values, $\phi_{i,j}^{\omega,f}(M) - \phi_{i,j}^{\omega',f}(M)$ $(\omega \neq \omega')$, passed the normality test, we further assess the difference between models by evaluating the confidence interval $CI_\alpha(\mu(\mathbf{C}_i^{(\omega,\omega'),f}(M)))$ given a desired level α, where we define:

$$\mathbf{C}_i^{(\omega,\omega'),f}(M) = \left(\phi_{i,1}^{\omega,f}(M) - \phi_{i,1}^{\omega',f}(M),\ \cdots,\ \phi_{i,J_f}^{\omega,f}(M) - \phi_{i,J_f}^{\omega',f}(M)\right)^T, \tag{8}$$

with J_f denoting the total number of subjects in fold f from Definition (3).

Here, we focus on the model difference in t1n, to test the hypothesis that Swin-UNETR has a higher contrast shapley value compared to other models, indicating a more balanced shapley value distribution and less basis toward t1c and t2f. The confidence intervals for the mean difference in Shapley values (Swin-UNETR minus the other models) indicate a **significant positive difference** at a confidence level of 0.95, suggesting that Swin-UNETR places more attention on the $t1n$ contrast (Figure 3).

To understand how transformer-based models differ from convolutional neural networks, we analyze cases where the Swin-UNETR model achieves a Dice score at least 20% higher than U-Net and vice versa. Specifically, we examine cases where the Swin-UNETR model achieves a Dice score 25% higher than U-Net (Figure 4), and U-Net achieves a Dice score 23% higher than Swin-UNETR (Figure 4). This comparison highlights the advantages and limitations of each architecture in medical image segmentation tasks.

4. Discussion

In this study, we systematically investigated the Shapley value for model explanation in multi-contrast medical image segmentation. Our proposed contrast-level Shapley explainability framework has three key contributions: (1) It is the first study to use Shapley analysis to explain multi-contrast medical image segmentation; (2) It is the first paper to analyze how different network structures weigh various

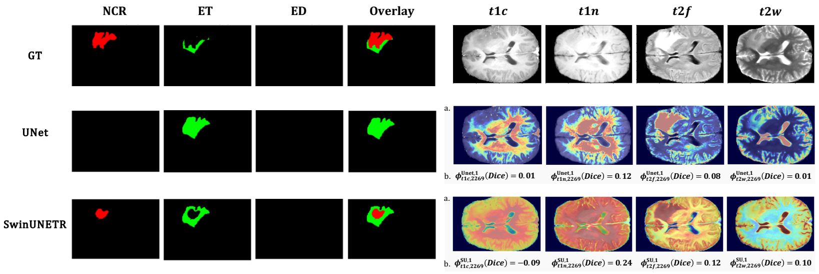

NCR
ET
ED
Overlay
$t1c$
$t1n$
$t2f$
$t2w$
GT
UNet
SwinUNETR
a.
b. $\phi^{\mathrm{Unet},1}_{t1c,2269}(Dice) = 0.01$
$\phi^{\mathrm{Unet},1}_{t1n,2269}(Dice) = 0.12$
$\phi^{\mathrm{Unet},1}_{t2f,2269}(Dice) = 0.08$
$\phi^{\mathrm{Unet},1}_{t2w,2269}(Dice) = 0.01$
a.
b. $\phi^{\mathrm{SU},1}_{t1c,2269}(Dice) = -0.09$
$\phi^{\mathrm{SU},1}_{t1n,2269}(Dice) = 0.24$
$\phi^{\mathrm{SU},1}_{t2f,2269}(Dice) = 0.12$
$\phi^{\mathrm{SU},1}_{t2w,2269}(Dice) = 0.10$

Figure 4: Case comparison where Swin-UNETR outperforms U-Net. For the first four columns, from top to bottom, display: Ground truth, U-Net predictions, and Swin-UNETR predictions. For the last four columns, from top to bottom, display: input images, model explanations for U-Net (explanation (a) and (b)), and Swin-UNETR predictions (explanation (a) and (b)), where (a) shows GradCAM explanation for each contrast and (b) presents the proposed constrast-level Shapley values.

MRI contrasts when making segmentation decisions; (3) It enhances clinical relevance by providing deeper insights into model performance with aggregate contributions of each MRI contrast in the tumor segmentation process, which is inherently interpretable by neuroradiologists, as they detect lesions by analyzing differences between different MRI contrasts in clinical practice.

Specifically, the contrast-level Shapley value reveals the (in)consistency of each model's explanations. The statistics indicate that Swin-UNETR is the most robust among all tested architectures. Despite being trained on different folds, Swin-UNETR consistently learns invariant representations across data subsets, whereas other models show variations in their explanations across folds (Table 1).

Moreover, the contrast-level Shapley value provides insights on the differences among model architectures. As shown in Figure 3, the model explanations indicate that U-Net exhibits a bias toward features from $t1c$ and $t2f$, while Swin-UNETR distributes its explanations more evenly across contrasts. This was further confirmed by comparing $t1n$ Shapley values across different models, which revealed statistically higher Shapley values for Swin-UNETR (Table 3).

We also present a case in Figure 4 to demonstrate how explanations of different models could provide key insights into model failure. As discussed before, the training data includes 3 different tumor subtypes (see section 2.1). The innermost component of the tumor (shown in red in Figure 4) is necrotic tissue in glioblastoma and meningioma, however, in metastasis, the definition of the innermost component is any tumor component that is not enhancing (but not necrotic). This implies that in $t2f$ images, the necrotic core will appear dark but non-enhancing metastatic tumor core and edema will appear bright.

Due to its dependence on contrasts with the highest intensity differences, namely $t1c$ and $t2f$, the U-Net architecture fails to accurately capture the innermost component (NCR). This suggests a potential bias towards $t1c$ and $t2f$, as indicated by the distribution of $C^{\omega,f}_{t1c}(\text{Dice})$ and $C^{\omega,f}_{t2f}(\text{Dice})$ exhibiting a significantly higher central tendency compared to $C^{\omega,f}_{t1n}(\text{Dice})$ and $C^{\omega,f}_{t2w}(\text{Dice})$ across all folds f and models $\omega \in \{\text{UNET, Seg-Resnet, UNETR, Swin-UNETR }\}$, as shown in Figure 2 and supported by statistical tests in Section 3.2. This bias may contribute to confusion with edema prediction, causing over-prediction relying on $t2f$ (edema appears bright as shown in Figure 4). However, swin-UNETR effectively learns both local and global relationships within different contrasts through its self-attention mechanism, and was able to more accurately localize the tumor core in this challenging case.

Finally, for this case, we provide a comparison between GradCAM and our proposed contrast-level Shapley. As seen in Figure 4, pixel-level explanations provided by GradCAM on each MRI contrast show model differences in terms of using pixel-level features. The heatmap of Swin-UNETR is more smooth while the heatmap of U-Net highlights only a few regions, but both of the explanations fail to capture clinically relevant explanations regarding contrast-level importance. For example, in Swin-UNETR, GradCAM exhibits a higher attention to $t1c$ compared to $t2f$. However, Contrast Shapley reveals that t1c negatively impacts the final Dice score, with a lower impact magnitude compared to $t2f$.

5. Conclusion

In this study, we propose Contrast Shapley for multi-contrast glioma segmentation. This method provides a quantitative framework for model explanation, offering insights into the fundamental characteristics of different deep learning architectures.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] M. H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, Journal of digital imaging 32 (2019) 582–596.

[2] L. Yan, C. Wang, F. Zhong, Y. Wang, Clinical inspired mri lesion segmentation, arXiv preprint arXiv:2502.16032 (2025).

[3] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (2019) 206–215.

[4] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, M. A. Viergever, Explainable artificial intelligence (xai) in deep learning-based medical image analysis, Medical Image Analysis 79 (2022) 102470.

[5] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[7] S. Lundberg, A unified approach to interpreting model predictions, arXiv:1705.07874 (2017).

[8] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[9] S. N. Hasany, F. Mériaudeau, C. Petitjean, Misure is all you need to explain your image segmentation, arXiv preprint arXiv:2406.12173 (2024).

[10] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al., The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, arXiv preprint arXiv:2107.02314 (2021).

[11] T. Ren, E. Honey, H. Rebala, A. Sharma, A. Chopra, M. Kurt, An optimization framework for processing and transfer learning for the brain tumor segmentation, arXiv:2402.07008 (2024).

[12] T. Ren, A. Sharma, J. E. H. Rivera, L. H. Rebala, E. Honey, A. Chopra, M. Kurt, Re-diffinet: Modeling discrepancy in tumor segmentation using diffusion models, in: Medical Imaging with Deep Learning, 2024.

[13] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: Learning dense volumetric segmentation from sparse annotation, CoRR abs/1606.06650 (2016). arXiv:1606.06650.

[14] A. Myronenko, 3d mri brain tumor segmentation using autoencoder regularization, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4, Springer, 2019, pp. 311–320.

[15] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, D. Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 272–284.

[16] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, D. Xu, Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 574–584.