COMMUNICATIONS IN SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

Decoding brain tumor insights: Evaluating CAM variants with 3D U-Net for segmentation

Dian Nova Kusuma Hardani[a,b], Igi Ardiyanto[a], Hanung Adi Nugroho*[a]

[a]*Department of Electrical and Information Engineering, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia*
[b]*Department of Electrical Engineering, Universitas Muhammadiyah Purwokerto, Purwokerto 53182, Indonesia*

Article history:
Received: 10 July 2024 / Received in revised form: 10 November 2024 / Accepted: 22 November 2024

Abstract

Brain tumor segmentation is critical for effective diagnosis and treatment planning. While, conventional manual segmentation techniques are seen inefficient and variable, highlighting the need for automated methods. This study enhances medical image analysis, particularly in brain tumor segmentation by improving the explainability and accuracy of deep learning models, which are essential for clinical trust. Using the 3D U-Net architecture with the BraTS 2020 dataset, the study achieved precise localization and detailed segmentation with the mean recall values of 0.8939 for Whole Tumor (WT), 0.7941 for Enhancing Tumor (ET), and 0.7846 for Tumor Core (TC). The Dice coefficients were 0.9065 for WT, 0.8180 for TC, and 0.7715 for ET. By integrating explainable AI techniques, such as Class Activation Mapping (CAM) and its variants (Grad-CAM, Grad-CAM++, and Score-CAM), the study ensures high segmentation accuracy and transparency. Grad-CAM, in this case, provided the most reliable and detailed visual explanations, significantly enhancing model interpretability for clinical applications. This approach not only enhances the accuracy of brain tumor segmentation but also builds clinical trust by making model decisions more transparent and understandable. Finally, the combination of 3D U-Net and XAI techniques supports more effective diagnosis, treatment planning, and patient care in brain tumor management.

Keywords: brain tumor segmentation; 3D U-Net; explainable AI; class activation mapping; deep learning; medical imaging

1. Introduction

Brain tumors present significant medical challenges due to their intricate structures and the critical functions of the brain regions they affect [1,2]. In medical imaging, the precise and reliable segmentation of brain tumors is paramount for accurate diagnosis, effective treatment planning, and continuous monitoring of disease progression [3]. Segmentation helps in accurately identifying and isolating relevant anatomical structures, which is crucial for determining disease severity and progression [4]. By enabling the precise identification of critical structures, segmentation aids in the early detection of diseases, allowing for timely intervention and treatment, potentially preventing severe outcomes. Traditional manual segmentation methods, despite their usage, are labor-intensive and susceptible to inter-observer variability, highlighting the necessity for automated and precise segmentation techniques [5,6].

Recent advancements in medical imaging and artificial intelligence (AI) have significantly enhanced tumor segmentation capabilities. The Brain Tumor Segmentation (BraTS) benchmark [7], a widely recognized dataset, has been instrumental in this progress, and drives the development of diverse machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction. Of these advancements, the 3D U-Net model, an extension of the U-Net architecture, has proven particularly influential in medical image segmentation [8,9]. This architecture, which integrates an encoder-decoder structure with skip connections, is a powerful tool for volumetric image segmentation, allowing for precise localization and detailed segmentation [7,8,10,11]. It is well-suited for capturing spatial hierarchies and fine details in volumetric data [12,13]. The U-Net architecture exemplifies a fully convolutional network capable of achieving precise segmentation in biomedical imaging applications [14]. Despite the success of these models, their "black-box" nature presents substantial challenges in clinical settings, where comprehending the decision-making process is essential for establishing trust and ensuring safety [15,16].

*Corresponding author. Telp.: +62-274-552305
Email: adinugroho@ugm.ac.id
https://doi.org10.21924/cst.9.2.2024.1477

To address this issue, explainable AI (XAI) techniques have been developed to provide insights into the model's decision-making process. Among these, Class Activation Mapping (CAM) and its variants including Grad-CAM, Grad-CAM++, and Score-CAM, are the prominent XAI methods that generate visual explanations highlighting the regions of an image that are important for the model's predictions [17-20].

CAM, introduced by Zhou et al. [17], uses global average pooling to produce class-specific activation maps, offering a straightforward yet effective visualization technique. Grad-CAM, developed by Selvaraju et al. [18], extends CAM by using gradients to produce finer and more localized explanations. Grad-CAM++, proposed by Chattopadhay et al. [19], further refines Grad-CAM by considering pixel importance, thereby improving explanation accuracy. Score-CAM, introduced by Wang et al. [20], takes a different approach by generating visual explanations without using gradients; it thus addresses some limitations of gradient-based methods.

CAM, Grad-CAM, and their modifications, as implemented by Natekar et al. [21] and Saleem et al. [22], create visual heatmaps to clarify the specific influence of image areas on model predictions. This helps medical practitioners to understand which parts of the image are most significant in the model. Liu et al. [23] used CAM, and Zhu et al. [24] used Grad-CAM to provide additional insights into the decision-making process of the system.

This study aims to compare these four XAI techniques including CAM, Grad-CAM, Grad-CAM++, and Score-CAM within the context of brain tumor segmentation using a 3D U-Net model. By evaluating their performances and interpretabilities, this study seeks to provide insights into their applicability in clinical practice and contribute to the development of more transparent AI systems in medical imaging. The integration of explainable AI (XAI) techniques into these models has been explored to enhance interpretability. These techniques have been specifically adapted for three-dimensional data to preserve the spatial relationships within volumetric images, ensuring accurate and meaningful analysis. The combination of advanced deep learning models with XAI techniques marks a significant advancement toward more transparent and reliable AI systems in medical imaging.

This study significantly contributes to medical image analysis, particularly in brain tumor segmentation, by enhancing the explainability of deep learning models, which is crucial for clinical trust and adoption. Firstly, it advances deep learning applications in medical imaging by improving both accuracy and interpretability of segmentation. Different from traditional methods that prioritize performance, this study emphasizes the understanding of model decisions through the integration of explainable AI (XAI) techniques such as Class Activation Mapping (CAM) and its variants (Grad-CAM, Grad-CAM++, and Score-CAM). Secondly, it employs the 3D U-Net architecture, a leading model for volumetric image segmentation, to segment brain tumors using the BraTS 2020 dataset. The 3D U-Net facilitates precise localization and detailed segmentation, capturing intricate spatial details. By combining this robust architecture with XAI techniques, the study ensures both high segmentation accuracy and model transparency. The comparative analysis of CAMs and their variants is most effective in producing reliable and detailed visual explanations, thereby enhancing the interpretability of the 3D U-Net model.

The study's rigorous evaluation using metrics such as Dice coefficients and area under the curve (AUC) analyses provides a quantitative assessment of both segmentation performance and explanation quality. This dual focus sets a new standard for future research, promoting the integration of XAI techniques in medical imaging.

This paper is organized into four sections. The initial section introduces the fundamental research context by reviewing prior studies on brain tumor segmentation and the application of deep learning models, particularly with a focus on the challenges of interpretability. Section 2 details the BraTS 2020 dataset, the 3D U-Net model architecture, CAM variants, evaluation metrics, and experimental procedures. Section 3 presents the findings and evaluates segmentation performance and explainability using metrics such as recall, Dice coefficient, Jaccard index, specificity, DAUC, and IAUC. It compares CAM techniques and discusses clinical applicability and limitations. The last section summarizes the findings and suggests future research directions.

2. Materials and Methods

This section outlines the materials and methods used in the study. It describes the BraTS 2020 dataset and the 3D U-Net architecture employed for brain tumor segmentation. The implementation details of four explainable AI techniques, including CAM, Grad-CAM, Grad-CAM++, and Score-CAM, were provided along with their adaptation for 3D images. The evaluation metrics for segmentation performance and explainability are presented. Lastly, this section also details the experimental procedures, including training, validation, and testing phases.

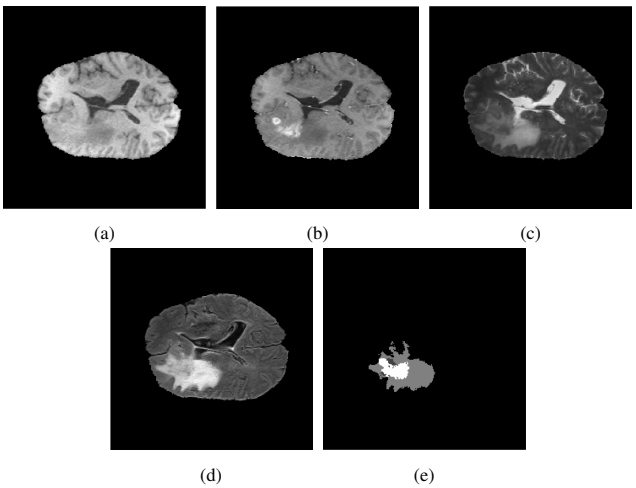

Fig. 1. Example images from the BraTS 2020 dataset: (a) T1, (b) T1ce, (c) T2, (d) FLAIR, and (e) the associated segmentation mask.

2.1. Data

This research employed the BraTS 2020 dataset, a highly regarded standard for brain tumor segmentation [6,25]. The dataset comprised 3D MRI scans from 369 glioma patients, including 76 with lower-grade glioma (LGG) and 293 with high-grade glioma (HGG). These MRI scans feature multimodal sequences such as T1-weighted (T1), T1-weighted contrast-enhanced (T1ce), T2-weighted (T2), and Fluid Attenuated In-

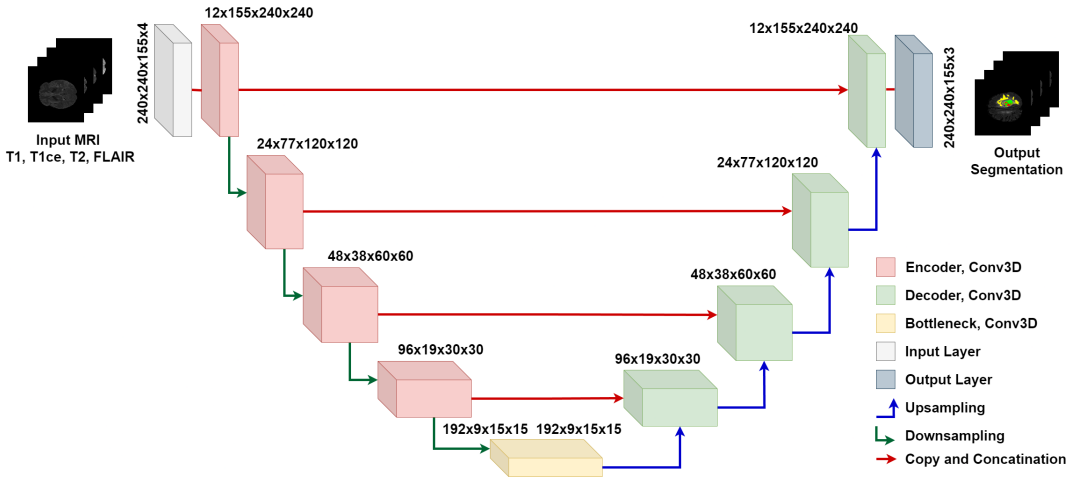

Fig. 2. The proposed 3D U-Net model's architecture for segmentation tasks.

version Recovery (FLAIR) sequences, encompassing patients diagnosed with glioblastoma (GBM) and LGG. Each 3D scan measures $240 \times 240 \times 155$ voxels. To maintain consistency, preprocessing steps included skull stripping, alignment to a standardized anatomical space, and resampling to a 1 mm^3 resolution. Expert annotations identify three tumor regions: enhancing tumor (ET), tumor core (TC), and whole tumor (WT). Fig. 1 illustrates a sample image from the 2020 dataset.

2.2. *3D U-Net Model Architecture*

For the segmentation model, this study employed a 3D U-Net, featuring an encoder-decoder structure with skip connections that facilitate precise localization and detailed segmentation. The architecture is designed to effectively capture spatial hierarchies and fine details, making it highly suitable for volumetric image segmentation tasks. Fig. 2 shows the 3D U-Net architecture used in this study. Below is a detailed description of the key components of the architecture.

Encoder Path. The encoder path of the 3D U-Net consists of five levels, each of which comprises sequential convolutional operations followed by 3D max-pooling layers. Each convolutional block within these levels contains a 3D convolutional layer (Conv3D) applying 3D filters to the input volume, followed by Group Normalization to stabilize and accelerate training. This is succeeded by a Leaky Rectified Linear Unit (LeakyReLU) activation function to introduce non-linearity, and another Conv3D layer to further process the features.

At each subsequent level of the encoder, the number of filters doubles, started from 24 filters at the first level. As the spatial dimensions of the feature maps decrease due to pooling, the gradual increase in the number of filters enables the network to capture more complex features.

Bottleneck. The bottleneck of the network, situated between the encoder and decoder paths, has a depth of 192 filters. This section serves as the deepest layer of the network, capturing the most abstract and high-level features of the input volume.

Decoder Path. The decoder path is designed to reconstruct the segmentation map from the high-level features captured by the encoder and bottleneck. It features transposed convolutional layers that perform upsampling, increasing the spatial resolution of the feature maps. Each upsampling operation is followed by convolutional blocks that incorporate skip connections from the corresponding encoder layers. These skip connections allow the decoder to utilize fine-grained information from the encoder, leading to more accurate and detailed segmentation results.

Output Layer. The final output layer consists of a Conv3D layer, which produces the segmentation map. This layer is followed by a softmax activation function that generates a probabilistic segmentation map with three channels. Each channel corresponds to different tumor regions, enabling the model to effectively segment and classify various parts of the tumor.

By integrating these components, the 3D U-Net architecture achieves a robust and precise segmentation performance, making it well-suited for medical imaging applications where accurate and detailed segmentation is crucial.

2.3. *Implementation of CAM Variants*

To generate visual explanations for the 3D U-Net model, this study implemented and evaluated different Class Activation Mapping (CAM) variants: CAM, Grad-CAM, Grad-CAM++, and Score-CAM. These techniques are crucial for enhancing the interpretability of deep learning models in medical imaging by providing visual explanations that can help to identify which parts of the input data contributed most to the model's predictions.

CAM (Class Activation Mapping). It uses global average pooling to generate class-specific activation maps, thus allowing straightforward visualizations. This technique highlights regions determining the model's prediction, providing basic insights into model decisions though it requires the modification of the network's architecture. For 3D images, the method is extended to consider volumetric data [17].

CAM identifies important regions in an image for a specific class prediction. The process starts by loading the trained model M and identifying the target layer L_{target}. A forward pass with the input image I_{input} captures the feature maps A and the class-specific weights $W_{C_{target}}$.

The CAM map is computed by taking a weighted sum of the feature maps, $CAM_{raw} = \sum_k W^k_{C_{target}} \cdot A^k$. Applying ReLU to CAM_{raw} produces CAM_{ReLU}, which is then resized to match the input image dimensions, yielding the final CAM output CAM_{out} as shown in (1).

$$CAM_{out} = Resize\left(ReLU\left(\sum_k W^k_{C_{target}} \cdot A^k\right), SizeOf\left(I_{input}\right)\right) \quad (1)$$

This model explains how CAM uses feature maps and class-specific weights to highlight important regions in the input image.

Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM extends CAM by utilizing gradient information to create more localized visual explanations. It is versatile for being applicable to a wider range of CNN architectures without required modifications, providing more refined explanations, especially useful for medical images. The Grad-CAM technique calculates the gradients for the target class relative to the output of the final convolutional layer, multiplies these gradients by the activation maps, and averages the results to generate the heatmap [18].

The process from start to forward pass is similar with the CAM model for data load. The gradients are globally averaged to obtain weights W. Each feature map A^k is weighted by W^k, and these weighted maps are summed to form $H_{weighted}$. Applying ReLU to $H_{weighted}$ produces H_{ReLU}, which is then resized to match the input image dimensions. Equation (2) indicates the output results of $H_{Grad-CAM}$.

$$H_{Grad-CAM} = Resize\left(ReLU\left(\sum_k W^k \cdot A^k\right), SizeOf\left(I_{input}\right)\right) \quad (2)$$

This model illustrates how Grad-CAM utilizes gradient information to emphasize significant regions in the input image, offering more detailed visual explanations compared to CAM.

Grad-CAM++. The Grad-CAM++ builds on Grad-CAM by incorporating higher-order gradients, allowing more accurate and detailed visual explanations [19]. This technique is particularly beneficial when handling complex or overlapping features in medical images, improving precision over Grad-CAM. The process begins by loading the trained model M and identifying the target layer L_{target}. A forward pass with the input image I_{input} is performed, capturing the feature maps A and computing the gradients α for the target class C_{target}. Additionally, positive partial derivatives β are calculated.

The weights W are computed using both α and β through the Grad-CAM++ formula. Each feature map A^k is then weighted by W^k, and these weighted maps are summed to produce $M_{weighted}$. Applying ReLU to $M_{weighted}$ results in M_{ReLU} with resized to match the input image dimensions, yielding the final Grad-CAM++ output in (3).

$$H_{Grad-CAM++} = Resize\left(ReLU\left(\sum_k (A^k \times W^k)\right), SizeOf\left(I_{input}\right)\right) \quad (3)$$

This model elucidates how Grad-CAM++ employs higher-order derivatives to generate the precise and detailed visual representations of significant regions within the input image.

Score-CAM. Score-CAM generates visual explanations by perturbing the input image, observing the changes in the model's output scores, and combining these changes to produce a comprehensive heatmap [20]. Different from other CAM variants, Score-CAM avoids reliance on gradients, which helps to reduce gradient noise and often results in clearer visualizations. However, this approach can be more computationally intensive due to the need for repeated perturbations and score evaluations. The process starts by performing a forward pass to capture the feature maps A. Each feature map A^k is normalized to A^k_{norm}, and the input image is perturbed by multiplying it with the normalized feature map, resulting in a perturbed image I'_{input}.

A forward pass is then performed with the perturbed image to obtain the output S^k, and the target class score $score^k$ is recorded. These scores serve as weights for the feature maps. The feature maps are weighted by these scores, and a weighted sum is computed. This sum is passed through a ReLU activation function, resulting in H_{ReLU}. The activated map is resized to match the input image dimensions, resulting in the Score-CAM output as shown in (4)

$$H_{Score-CAM} = Resize\left(ReLU\left(\sum_k \left(score^k \cdot A^k\right)\right), SizeOf(I_{input})\right) \quad (4)$$

These methods have been specifically adapted for three-dimensional data to maintain the integrity of spatial relationships within volumetric images. By doing so, they ensure that the intricate spatial dependencies and structures inherent in the volumetric data are preserved, enabling more accurate and meaningful analysis.

Table 1 highlights the key differences between the CAM variants, focused on their underlying mechanisms, advantages, limitations, and clinical applicability. The addition aims to improve the accessibility and understanding of these techniques for readers who may not be well-versed in deep learning or explainable AI.

2.4. Evaluation Metrics

In medical imaging, the evaluation of segmentation performance and explainability is critical to ensure the effectiveness and transparency of the 3D U-Net model. The following metrics are employed to comprehensively assess the model's performance.

Segmentation Performance. To evaluate the segmentation performance, several metrics are used. Recall measures the model's ability to correctly identify tumor regions. High recall indicates that the model successfully detects a significant portion of the true positive cases, thereby reducing the number of false negatives. It is defined in (5). The Dice Coefficient evaluates the degree of overlap between the predicted segmentation and the actual ground truth. It is calculated as twice the area of overlap divided by the total number of voxels in both the predicted and the ground truth segmentations. A higher Dice Coefficient indicates better performance and more accurate segmentation. Equation (6) shows the formula of the dice. Similar to the Dice Coefficient, the Jaccard Index is an additional overlap metric used to evaluate the similarity between the predicted segmentations and the actual ground truth. It is the ratio of the intersection over the union of the predicted and ground truth areas as shown in (7). Higher Jaccard Index values reflect better segmentation accuracy. Equation (8) is specificity, which evalu-

Table 1. Summary of differences between CAM, Grad-CAM, Grad-CAM++, and Score-CAM

CAM Variant	Underlying Mechanism	Advantages	Limitations	Clinical Applicability
CAM	Global average pooling	Simple, direct visual explanation	Architecture modification is required	Suitable for general model visualization but limited to specific architectures
Grad-CAM	Gradients with respect to feature maps	No architecture modification required, more refined than CAM	Gradient noise can affect results	Highly applicable in clinical settings for models with no architectural modifications required
Grad-CAM++	Higher-order gradients	Improved localization accuracy	Higher computational cost	Effective for complex medical images with overlapping features
Score-CAM	Model response to perturbed input	Gradient-independent, clearer explanations	Computationally intensive	Useful in clinical scenarios where gradient noise impacts reliability

ates the model's capability to accurately distinguish voxels that are not part of the segmentation of interest.

$$Recall = \frac{|X \bigcap Y|}{|Y|} \tag{5}$$

$$Dice = \frac{2 \times |X \bigcap Y|}{|X| + |Y|} \tag{6}$$

$$Jaccard = \frac{|X \bigcap Y|}{|X \bigcup Y|} \tag{7}$$

$$Specificity = \frac{|N \bigcap M|}{|N|} \tag{8}$$

In the realm of 3D segmentation evaluation, X denotes the set of voxels classified by the model as part of the segmentation. Conversely, Y represents the set of voxels corresponding to the ground truth segmentation. On the other hand, N signifies the set of voxels excluded from the ground truth segmentation, while M identifies the voxels accurately recognized by the model as not belonging to the segmentation.

Explainability Evaluation. To assess the quality of visual explanations provided by the model, two specific metrics are used including Deletion Area Under the Curve (DAUC) and Insertion Area Under the Curve (IAUC). These metrics were introduced by Petsiuk et al. [26]. The DAUC measures the change in model performance when the most important features, as identified by the Explainable Artificial Intelligence (XAI) method, are perturbed. A significant drop in performance indicates that the identified features are indeed crucial for the model's decision-making process. The DAUC is defined in (9). The model's prediction score for the correct class, $f(x_i)$, is given the modified input x_i after removing important pixels up to the i-th percentile based on the explanation. N represents the total number of steps in the deletion process, typically set to 100 for a smooth curve.

$$DAUC = \frac{1}{N} \sum_{i=1}^{N} f(x_i) \tag{9}$$

Conversely, the IAUC metric measures the change in model performance when the least important features are perturbed. A minimal change or improvement in performance suggests that these features have little to no impact on the model's decisions, validating the effectiveness of the XAI method in feature importance ranking. The IAUC is mathematically defined in (10). The model's prediction score for the correct class, $f(x_i')$, is given the modified input x_i', after inserting the most important pixels up to the i-th percentile into the baseline image based on the explanation. N is the total number of insertion steps, often set to 100 for detailed evaluation.

$$IAUC = \frac{1}{N} \sum_{i=1}^{N} f(x_i') \tag{10}$$

Together, these metrics provide a comprehensive evaluation of both the segmentation accuracy and the explainability of the 3D U-Net model, ensuring its reliability and transparency in medical imaging applications.

2.5. Experimental Procedures

This section details the training, validation, and testing processes used in the study. The training of the 3D U-Net model on the BraTS 2020 dataset is outlined, including data preprocessing and training parameters. The validation approach using ground truth segmentations is described, followed by the testing phase, where segmentation performance and explainability of the XAI techniques are evaluated using specified metrics.

Training. For this study, experiments were conducted on the BraTS2020 challenge dataset,which has been preprocessed by the organizers. Due to intensity variations from different MRI scanners, preprocessing was crucial. To mitigate these variations, intensity normalization was performed using Z-score normalization, entailing subtracting the mean of each voxel and dividing by its standard deviation. This process standardizes each brain image to have a mean of zero and a variance of one [27].

The data distribution was done randomly, allocating 64% of the data for training, 27% for validation, and 9% for testing. This distribution ensured a robust training process, allowing the model to generalize well while providing sufficient data for validation and testing phases. The comprehensive preprocessing and balanced data distribution contribute to the accuracy and reliability of the segmentation model, enhancing its applicability in clinical practice.

The model's hyperparameter settings included an input size of $4 \times 240 \times 240 \times 155$ for four-channel volumetric data and an output size of $3 \times 240 \times 240 \times 155$ for a three-class segmentation task. A batch size of 2 processed two samples per iteration. Group Normalization helped to stabilize training despite the small batch size by normalizing features within groups of channels. The Leaky ReLU activation function prevents the dying neuron problem by allowing small gradients for inactive neurons.

The Adam optimizer, with a learning rate of 0.001, combines

the strengths of AdaGrad and RMSProp for efficient and stable training. The hybrid loss function, combining Binary Cross-Entropy (BCE) and Dice loss, is ideal for segmentation tasks. BCE ensures pixel-wise accuracy, while Dice loss ensures the predicted segments match the ground truth in shape and size. Training for 100 epochs allows the model to learn and improve its performance incrementally. This setup aims to achieve high accuracy and robustness in multi-class segmentation tasks, reflecting a well-designed training strategy.

Validation. The validation phase refines model parameters and prevents overfitting by using different data subsets for testing and fitting. This ensures the model generalizes well to new data. Loss functions are crucial, as their scores indicate the model's accuracy and robustness. Performance is validated using ground truth segmentations, which are essential for tuning hyperparameters and evaluating generalization. Monitoring this phase helps to maintain high accuracy across diverse samples.

Testing. The final evaluation of the model was conducted on a test set consisting of 34 samples of test data. During this phase, various segmentation performance metrics were computed to assess the accuracy of the model's predictions. Additionally, the quality of visual explanations was evaluated to ensure the model's decisions that can be interpreted and trusted. This comprehensive evaluation covers both the segmentation accuracy and the explainability of the model, providing a thorough assessment of its overall performance.

3. Results and Discussion

This section evaluates the 3D U-Net model's performance and interpretability, enhanced with various Class Activation Mapping (CAM) techniques. We used metrics such as recall, Dice coefficient, Jaccard index, and specificity to measure segmentation accuracy. For interpretability, we assessed the model's predictions using Deletion Area Under the Curve (DAUC) and Insertion Area Under the Curve (IAUC) metrics. The model's performance was tested on the BraTS 2020 dataset, including multimodal MRI scans of brain tumors. We compared the baseline CAM method with its advanced variants: Grad-CAM, Grad-CAM++, and Score-CAM. The discussion highlights the strengths and weaknesses of each technique, emphasizing their clinical applicability and reliability. This analysis provides insights into improving the transparency and trustworthiness of AI-driven medical imaging solutions.

3.1. Segmentation Performance

This study evaluated the performance of a 3D U-Net model for brain tumor segmentation using MRI scans. The dataset included multiple MRI modalities, such as T1, T1ce, T2, and FLAIR, which provided comprehensive information for accurate segmentation. The 3D U-Net architecture, featuring an encoder-decoder structure and skip connections, is particularly well-suited for capturing spatial hierarchies and fine details in volumetric data.

Fig. 3 demonstrates the model's effectiveness, segmentation results for two sample cases. The figure illustrates the original MRI scans, the ground truth segmentations, and the model's predicted segmentations. Different colors were used to represent various tumor regions, highlighting the model's ability to accurately identify and segment these areas. This visual representation underscores the model's robustness and precision in handling complex medical imaging tasks.

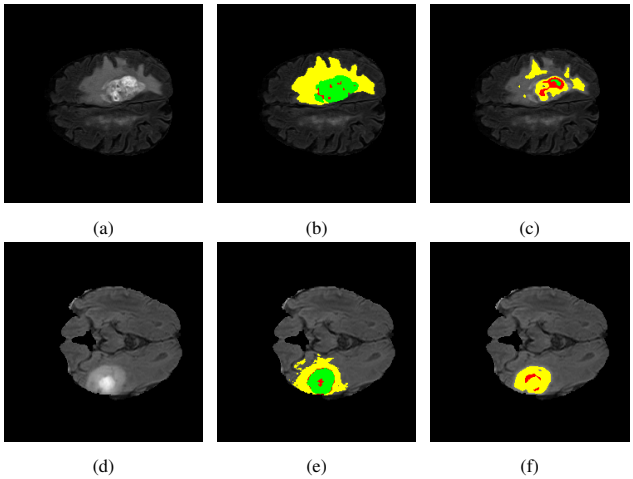

Fig. 3. Brain tumor segmentation results for two samples. (a) and (d) present the original MRI scans. (b) and (e) show the ground truth segmentations, with colors indicating different tumor regions (yellow: edema, green: enhancing tumor, red: necrotic core). (c) and (f) display the model's predicted segmentations, using the same color scheme as the ground truth.

The performance of the 3D U-Net model was evaluated using the BraTS 2020 dataset. The segmentation results were quantified using recall, Dice coefficient, Jaccard index, and specificity. Table 2 presents summary statistics for four key metrics used in the evaluation of tumor segmentation performance. These metrics were provided for three types of tumor regions: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). Each metric and tumor type combination includes several statistical measures: count (sample size), mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum.

Recall measures the ability of the model to correctly identify all relevant instances of the tumor regions. The mean recall values were found relatively high for all tumor types with Whole Tumor (WT) having the highest mean recall of 0.8939, followed by Enhancing Tumor (ET) at 0.7941, and Tumor Core (TC) at 0.7846. This indicates that the model is generally effective at capturing most of the tumor regions though there were some variabilities as indicated by the standard deviations, particularly for TC.

The dice coefficient is another metric used to gauge the similarity between the predicted and actual tumor regions. The Dice scores were found quite high with WT achieving a mean of 0.9065, indicating very good overlap between predicted and actual tumor regions. TC and ET had slightly lower mean Dice scores of 0.8180 and 0.7715, respectively, which still reflected a strong performance but indicated room for improvement, especially in the more challenging ET regions.

Jaccard index, similar to the Dice coefficient but more stringent, showed slightly lower mean values compared to Dice with WT at 0.8293, TC at 0.6932, and ET at 0.6287. These lower values were expected given the Jaccard index's stricter calculation. The standard deviations were also higher for TC and ET, suggesting more variability in the model's performance across different cases.

Specificity quantifies the model's accuracy in correctly de-

Table 2. Performance metrics for brain tumor segmentation

Statistic	Recall			Dice			Jaccard			Specificity		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
Count	34	34	34	34	34	34	34	34	34	34	34	34
Mean	0.8939	0.7846	0.7941	0.9065	0.8180	0.7715	0.8293	0.6932	0.6287	0.9992	0.9995	0.9995
Std Deviation	0.0206	0.0497	0.0232	0.0131	0.0310	0.0266	0.0218	0.0457	0.0351	0.0001	0.0001	0.0001
Minimum	0.8502	0.7155	0.7582	0.8755	0.7786	0.7246	0.7786	0.6375	0.5681	0.9989	0.9994	0.9993
25th Percentile	0.8778	0.7502	0.7700	0.8963	0.7951	0.7532	0.8121	0.6598	0.6041	0.9991	0.9995	0.9994
Median	0.9042	0.7779	0.7979	0.9096	0.8136	0.7764	0.8342	0.6858	0.6345	0.9992	0.9995	0.9994
75th Percentile	0.9106	0.8013	0.8151	0.9186	0.8311	0.7932	0.8495	0.7110	0.6573	0.9992	0.9995	0.9995
Maximum	0.9276	0.9073	0.8324	0.9223	0.9099	0.8100	0.8558	0.8347	0.6806	0.9997	0.9998	0.9996

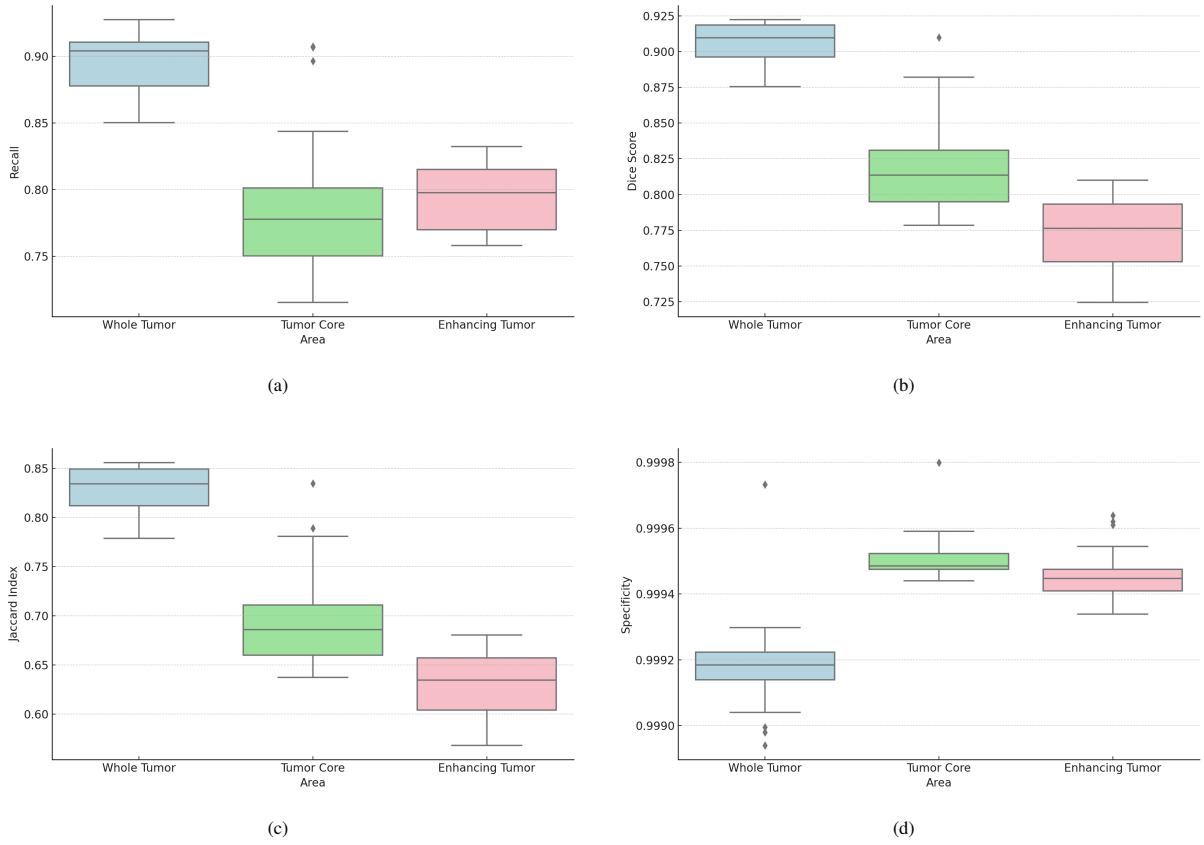

Fig. 4. Boxplots of segmentation performance metrics for different tumor regions: (a) Recall, (b) Dice Score, (c) Jaccard Index, and (d) Specificity.

tecting non-tumor regions. The specificity values were exceptionally high across all tumor types with WT, TC, and ET all having mean values above 0.999. This indicated that the model is highly effective at correctly identifying non-tumor regions, thereby minimizing false positives. The very low standard deviations further highlight the consistency of the model's performance in terms of specificity.

The best-performing metric for brain tumor segmentation, according to the summary statistics table, was Specificity. This metric showed exceptionally high values across all tumor types (WT, TC, and ET) with mean values all above 0.999. The high specificity indicated that the model is very effective at correctly identifying non-tumor regions, thereby minimizing false positives. Additionally, the low standard deviations suggest that the model's performance is consistent and reliable in terms of specificity across different cases.

While other metrics like Recall and Dice also showed strong performance, particularly for WT regions, Specificity stood out due to its near-perfect mean values and minimal variability, making it the best-performing metric in this case.

The high specificity in brain tumor segmentation was likely due to the larger proportion of non-tumor regions in the dataset, the distinct characteristics of non-tumor areas, the training strategies that penalized false positives, and the inherent properties of the segmentation models used. This combination of factors made the model very effective at correctly identifying non-tumor regions, leading to exceptionally high specificity values.

Fig. 4 illustrates the distribution of the four brain tumor segmentation performance metrics. The boxplots highlight the model's strong performance in identifying whole tumor regions and its effectiveness in minimizing false positives, while also indicating areas for improvement in segmenting more challenging tumor core and enhancing tumor regions.

The performance metrics used in this study, such as the Dice coefficient and Jaccard index, were deemed important for evaluating the accuracy of brain tumor segmentation in clinical settings. The Dice coefficient measured how much the predicted segmentation overlaps with the actual tumor, helping to assess how well the model captures the tumor region. A high Dice score is crucial in clinical applications as it ensures the precise identification of tumor boundaries, which is vital for treatment planning. The Jaccard index, which measures the intersection over union of the predicted and true segments, provides a stricter measure of similarity. A high Jaccard index indicates consistency between the model's predictions and radiologists' manual segmentations, building trust in the model for diagnostic use.

Recall and specificity provided further insight into the model's reliability. Recall measured how well the model detected tumor areas, minimizing the chances of missing malignant regions. Specificity, on the other hand, reflected the model's accuracy in identifying non-tumor areas, helping to reduce any false positives and unnecessary treatments. Together, these metrics provided a well-rounded view of the model's performance, supporting clinical decisions and enhancing patient safety.

A comparative analysis with other state-of-the-art segmentation models was conducted to benchmark the performance of the proposed 3D U-Net model. As shown in Table 3, the proposed 3D U-Net demonstrated competitive performance across all tumor regions. For WT segmentation, it achieved a Dice score of 0.9065, slightly lower than RMU-Net [30] but still higher than most other models, indicating strong capability in identifying the complete tumor region. For TC, the Dice score of 0.8180 was lower than models like RMU-Net [30] and nn-UNet [13] but still outperformed several other models, showing good reliability. In ET segmentation, the proposed model achieved 0.7715, performing moderately well compared to top performers like RMU-Net [30] but significantly better than models like MCN. The main advantage of the proposed model lied in its balanced performance across all tumor regions, with consistently high scores demonstrating effective tumor segmentation. Its competitive results, combined with efficiency and robustness, have made it a suitable option for real-world clinical applications where both accuracy and practicality are crucial.

Table 3. Performance comparison of the proposed model with state-of-the-art methods on the BraTS 2020 dataset.

Model	Dice		
	WT	TC	ET
nn-UNet [13]	0.8895	0.8506	0.8203
Res-U-Net [28]	0.8920	0.7880	0.7230
R2AU-Net [29]	0.8784	0.7993	0.7426
Residual Mobile UNet (RMU-Net) [30]	0.9135	0.8813	0.8326
IRDNU-Net [31]	0.8760	0.8400	0.8010
Deep Residual UNet (dRes-UNet) [27]	0.8660	0.8357	0.8004
Mixture of Calibrated Networks (MCN) [32]	0.6482	0.5548	0.6825
Proposed	0.9065	0.8180	0.7715

3.2. Explainability Evaluation

The explainability of the model was rigorously assessed using the DAUC and IAUC metrics for various CAM techniques, including Grad-CAM, Grad-CAM++, and Score-CAM. These explainable AI (XAI) techniques were evaluated on a test set comprising 34 samples to determine their effectiveness in highlighting key regions identified by the model. The results, as detailed in Table 4, illustrated the extent to which each method contributed to the understanding of critical features in segmentation tasks. This comprehensive analysis ensured that the model's predictions are not only accurate but also interpretable and trustworthy, which is crucial for their application in medical imaging. By providing clear visual explanations, these techniques enhanced the transparency of the model, thereby facilitating its acceptance and reliability in clinical settings.

Based on the statistical analysis in Table 4, Grad-CAM was found as the most effective and consistent XAI technique for visual explanations, showing the highest mean DAUC and low variability, indicating reliable performance in highlighting relevant segmentation areas. It also had a relatively high mean IAUC, suggesting that it can occasionally improve model performance. Grad-CAM++ also performed well with a high mean DAUC but showed greater variability, making its outcomes less predictable. Despite this, it is valuable for more complex tasks requiring advanced capabilities. In general, CAM, with moderate DAUC and low IAUC variability, was reliable but less impactful in enhancing interpretability. It serves well as a baseline technique, providing consistent visual explanations. Score-CAM had the lowest mean DAUC and highest variability, indicating less reliability. However, it showed potential in specific cases, suggesting usefulness in certain contexts but still needing refinement for consistency.

Figure 5 shows the XAI techniques with the lowest DAUC values, indicating poor performance in generating accurate visual explanations. CAM, shown in Fig. 5c, offered little improvement in interpretability and did not match the ground truth segmentations shown in Fig. 5b. Grad-CAM, illustrated in Fig. 5f, aligned moderately with the ground truth in Fig. 5e but did not clearly highlight the tumor regions. Grad-CAM++, depicted in Fig. 5i, provided more detail than CAM but showed some inconsistencies in identifying tumor areas compared to the ground truth in Fig. 5h. Score-CAM, presented in Fig. 5l, performed the worst, with visual explanations that poorly matched the ground truth shown in Fig. 5k. These results highlight the difficulties these techniques face in accurately identifying tumor regions for medical imaging tasks.

Fig. 6 presents a comparative analysis of XAI techniques with the highest IAUC values. Grad-CAM, shown in Fig. 6f, produced the most effective visual explanations, demonstrating a strong alignment with the ground truth segmentations shown in Fig. 6e. This close correspondence enhances both the in-

Table 4. The explainability metrics for XAI techniques

Statistic	CAM		Grad-CAM		Grad-CAM++		Score-CAM	
	DAUC	IAUC	DAUC	IAUC	DAUC	IAUC	DAUC	IAUC
Count	34	34	34	34	34	34	34	34
Mean	0.8246	0.0015	0.0019	**0.9607**	0.0108	0.9290	**0.0013**	0.6886
Std Deviation	0.1640	0.0009	0.0016	0.0315	0.0358	0.1726	0.0011	0.2251
Minimum	0.3729	0.0004	0.0004	0.8892	0.0004	0.0087	0.0004	0.2712
25th Percentile	0.7310	0.0006	0.0005	0.9396	0.0006	0.9442	0.0004	0.4784
Median	0.8805	0.0013	0.0019	0.9742	0.0020	0.9691	0.0012	0.6579
75th Percentile	0.9668	0.0021	0.0027	0.9846	0.0050	0.9864	0.0018	0.9400
Maximum	0.9944	0.0037	0.0066	0.9970	0.1992	0.9985	0.0051	0.9983

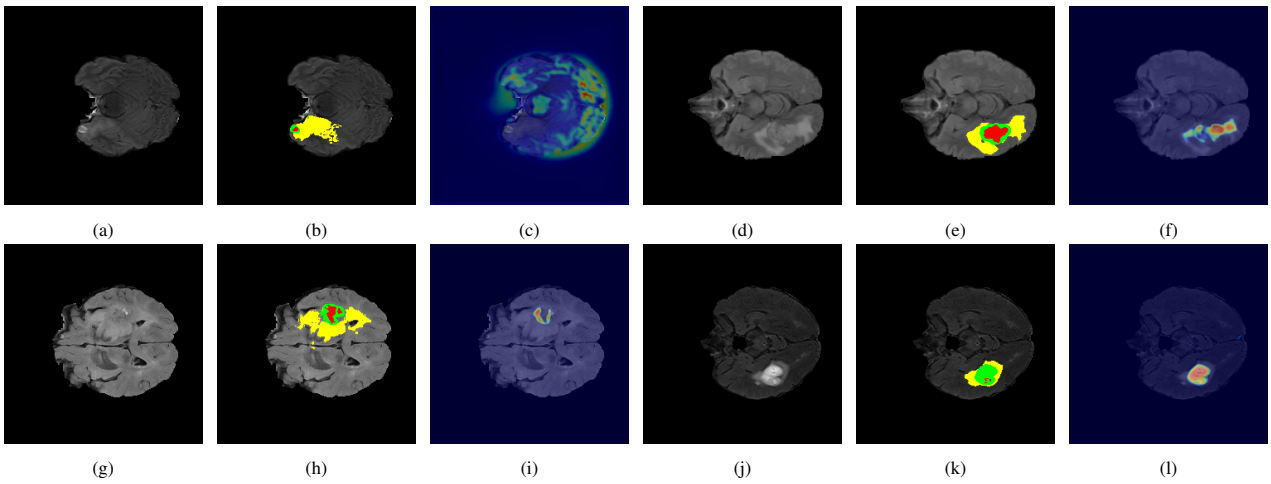

Fig. 5. Visualization results for XAI techniques with the lowest DAUC value. (a-c): CAM techniques, (d-f): Grad-CAM techniques, (g-i): Grad-CAM++ techniques, (j-l): Score-CAM techniques. Original MRI scans for different samples (a, d, g, j), ground truth segmentations showing various tumor regions (b, e, h, k), and visual explanations generated by the corresponding XAI techniques overlaying the original MRI scans (c, f, i, l).

terpretability of the model's decisions and overall performance. Grad-CAM++, depicted in Fig. 6i also exhibited high performance, generating visual explanations that closely aligned with the ground truth in Fig. 6h. However, Grad-CAM++ displayed slightly higher variability in heatmap quality compared to Grad-CAM. While offering some improvements, CAM, illustrated in Fig. 6c, and Score-CAM, presented in Fig. 6l produced visual explanations that were less consistent and lacked the detailed localization that Grad-CAM and Grad-CAM++ achieved. These findings indicate that Grad-CAM remains a robust choice for generating interpretable and reliable model explanations, especially where accurate localization of features is critical.

The figures demonstrate that Grad-CAM provided the most reliable and accurate visual explanations, particularly when considering IAUC values, enhancing model interpretability by closely aligning with ground truth segmentations. This is because Grad-CAM uses a simpler approach that often results in more stable and reliable output [33]. The straightforward nature of Grad-CAM allows for easier interpretation of results [33,34]. Grad-CAM tends to produce more coherent and focused heatmaps, especially when identifying larger features or dominant regions. In practice, Grad-CAM is generally better for applications that require broader or simpler localization.

In this study, Grad-CAM++ also performed well though with more variability. CAM and Score-CAM were less consistent. These findings emphasize the importance of selecting the appropriate XAI technique based on specific evaluation metrics such as DAUC and IAUC to optimize model explainability and reliability in clinical applications.

3.3. Comparative Insights

The comparative analysis of CAM, Grad-CAM, Grad-CAM++, and Score-CAM in brain tumor segmentation using a 3D U-Net model highlighted their varying performance and clinical applicability.

CAM provided basic visual explanations with moderate performance and modest improvements in interpretability. Its explanations often lacked precision and alignment with ground truth segmentations. Grad-CAM stood out with the highest DAUC and IAUC values, offering clear, detailed, and consistently accurate visual explanations that closely matched ground truth segmentations. This then made it highly reliable and valuable in clinical settings for accurately highlighting tumor regions. Grad-CAM++ also performed well, delivering detailed and generally accurate visual explanations though with slightly more variability than Grad-CAM. Despite this, it remained a strong option for clinical use. Score-CAM showed the lowest mean DAUC and higher variability, indicating less reliable performance. Its visual explanations were less consistent and precise, making it less effective compared to the other techniques.

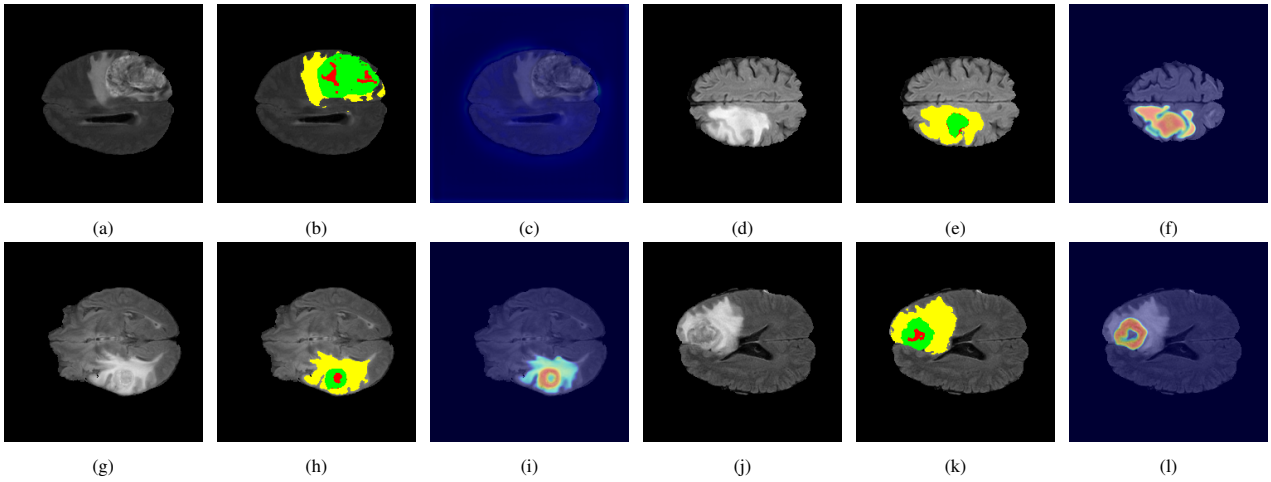

Fig. 6. Visualization results for XAI techniques with the highest IAUC value. (a-c): CAM techniques, (d-f): Grad-CAM techniques, (g-i): Grad-CAM++ techniques, (j-l): Score-CAM techniques. Original MRI scans for different samples (a, d, g, j), ground truth segmentations highlighting various tumor regions (b, e, h, k), and visual explanations generated by each XAI technique overlaid on the original MRI scans (c, f, i, l).

These XAI techniques significantly enhance clinical applicability by providing comprehensible visual explanations, thereby enabling radiologists to verify model predictions and gain insights into the decision-making process of the 3D U-Net model. By offering clear and interpretable visualizations, these techniques facilitate a deeper understanding of the model's reasoning, which is essential for informed clinical decision-making. Consequently, the integration of XAI methods improves the reliability and trustworthiness of AI-driven medical imaging solutions, fostering greater confidence in their deployment within clinical settings. This increased transparency not only aids in validating the accuracy of the AI models but also ensures that the intricate decisions made by these systems are comprehensible and justifiable, ultimately leading to enhanced patient care and outcomes.

Grad-CAM became the preferred XAI technique for brain tumor segmentation using a 3D U-Net model due to its superior accuracy, detail, and consistency. Grad-CAM++ followed closely with strong performance but slightly more variability. CAM offered moderate, stable performance, while Score-CAM showed potential but less consistent. These findings highlight the importance of selecting the right XAI technique to optimize model explainability and reliability in clinical applications.

3.4. Limitations and Future Works

The study's reliance on the BraTS2020 dataset may limit the generalizability of the findings to other datasets or medical imaging types. There is variability in the performance of XAI techniques like Grad-CAM++ and Score-CAM, which can lead to inconsistencies and affect their reliability in clinical applications. The significant computational resources required for implementing 3D U-Net and XAI techniques may restrict their use in settings with limited access to high-performance computing. Additionally, the study primarily used DAUC and IAUC metrics, which may not capture all aspects of model interpretability and clinical utility. The effectiveness of visual explanations also depends on accurate interpretation by clinicians, which might vary.

Future study should test the model and XAI techniques on a broader range of datasets to evaluate their generalizability and robustness. Developing hybrid approaches that combine the strengths of multiple XAI techniques could improve the precision and reliability of visual explanations. Efforts should be made to optimize the computational efficiency of these techniques to enable their use in real-time clinical settings. Incorporating additional evaluation metrics, including qualitative assessments from clinical experts, would provide a more comprehensive understanding of model interpretability. Creating user-friendly interfaces to integrate XAI techniques into clinical workflows could help clinicians to interpret visual explanations more effectively. Longitudinal studies are required to assess the long-term impact of using XAI-enhanced models on clinical outcomes, providing valuable insights into their practical benefits and limitations. Addressing these areas will enhance the applicability and effectiveness of 3D U-Net models and XAI techniques in medical imaging.

As part of our future work, we plan to conduct a detailed qualitative study involving clinical experts, specifically radiologists, to assess the practical applicability of the visual explanations generated by the CAM variants. This study will systematically gather feedback on how these explanations align with the clinical judgments of radiologists, providing insights into the model's utility in a real-world clinical environment. Such qualitative insights will help to establish the level of agreement between model explanations and expert assessments, enhancing our understanding of the model's reliability and usability.

4. Conclusion

This study systematically evaluated the performance and interpretability of several explainable AI techniques including CAM, Grad-CAM, Grad-CAM++, and Score-CAM in conjunction with a 3D U-Net model for brain tumor segmentation. Utilizing the BraTS 2020 dataset, the segmentation model demonstrated a robust performance across various metrics. Mean recall values were 0.8939 for Whole Tumor (WT), 0.7941 for Enhancing Tumor (ET), and 0.7846 for Tumor Core (TC). Dice coefficients were 0.9065 for WT, 0.8180 for TC, and 0.7715 for ET. The Jaccard index values were 0.8293 for WT, 0.6932 for TC, and 0.6287 for ET. Particularly high specificity with mean val-

ues above 0.999 for all tumor types indicated the reliable identification of non-tumor regions.

Of the explainable AI techniques, Grad-CAM emerged as the most effective, providing the highest mean IAUC of 0.9607 and a relatively low mean DAUC of 0.0019. Grad-CAM++ also showed promise with a mean IAUC of 0.9290 and DAUC of 0.0108, though it exhibited greater variability. CAM had a mean IAUC of 0.8246 and DAUC of 0.0015. Score-CAM had the lowest mean IAUC of 0.6886 and an DAUC of 0.0013, indicating less reliable performance.

The high specificity observed in segmentation results underscores the model's potential in minimizing false positives, crucial for clinical applications. However, the variability in performance metrics, especially for more challenging tumor regions, highlights the need for further refinement. The study's reliance on the BraTS 2020 dataset suggests a need for future study to explore the generalizability of these findings across different datasets and medical imaging modalities.

Future work should focus on improving the computational efficiency of these techniques to facilitate their integration into real-time clinical settings. Additionally, hybrid approaches that combine multiple explainable AI techniques could enhance the precision and reliability of visual explanations. Incorporating qualitative assessments from clinical experts will also be essential in evaluating the practical utility of these models in medical practice. Longitudinal studies assessing the long-term impact of using explainable AI-enhanced models on clinical outcomes will provide deeper insights into their benefits and limitations. Addressing these areas will be pivotal in advancing the applicability and effectiveness of 3D U-Net models and explainable AI techniques in medical imaging.

Acknowledgment

The study work was funded by the Directorate of Research, Universitas Gadjah Mada, through the *Final Project Recognition Grant* Universitas Gadjah Mada *Number* 5075/UN1.P.II/Dit-Lit/PT.01.01/2023.

References

1. M. A. Al Nasim, A. Al Munem, M. Islam, M. A. H. Palash, M. M. A. Haque, and F. M. Shah, *Brain tumor segmentation using enhanced u-net model with empirical analysis*, in 2022 25th International Conference on Computer and Information Technology (ICCIT), IEEE, 2022, pp. 1027–1032.
2. D. Patel, D. Patel, R. Saxena, and T. Akilan, *Multi-class brain tumor segmentation using graph attention network*, in 2023 8th International Conference on Signal and Image Processing (ICSIP), IEEE, 2023, pp. 196–201.
3. T. Magadza and S. Viriri, *Deep learning for brain tumor segmentation: A survey of state-of-the-art*, Journal of Imaging 7 (2) (2021) 19.
4. H. A. Nugroho, T. Kirana, V. Pranowo, and A. H. T. Hutami, *Optic cup segmentation using adaptive threshold and morphological image processing*, Commun. Sci. Technol. 4(2) (2019) 63-67.
5. S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, *A survey of mri-based medical image analysis for brain tumor studies*, Phys. Med. Biol. 58 (13) (2013) R97–R129.
6. B. H. Menze, et al., *The multimodal brain tumor image segmentation benchmark (BRATS)*, IEEE Trans. Med. Imaging 34(10) (2015) 1993–2024.
7. A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, *Brain tumor detection based on deep learning approaches and magnetic resonance imaging*, Cancers, 15(16) (2023) 4172.
8. U. Baid, et al., *A novel approach for fully automatic intra-tumor segmentation with 3d u-net architecture for gliomas*, Front. Comput. Neurosci. 14 (feb 2020).
9. L. Weninger, O. Rippel, S. Koppers, and D. Merhof, *Segmentation of Brain Tumors and Patient Survival Prediction: Methods for the BraTS 2018 Challenge*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11384 LNCS, 2019, pp. 3–12.
10. O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9901 LNCS, 2016, pp. 424–432.
11. N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, *U-net and its variants for medical image segmentation: A review of theory and applications*, IEEE Access 9 (2021) 82031–82057.
12. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, *Unet++: Redesigning skip connections to exploit multiscale features in image segmentation* (2019). doi:10.48550/ARXIV.1912.05074.
13. F. Isensee, P. F. Jager, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, *Automated design of deep learning methods for biomedical image segmentation* (2019). doi:10.48550/ARXIV.1904.08128.
14. J. B. Abraham, *Malaria parasite segmentation using U-Net: Comparative study of loss functions*, Commun. Sci. Technol. 4(2) (2019) 57-62.
15. W. Samek, T. Wiegand, and K.-R. Muller, *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*, 2017. doi:10.48550/ARXIV.1708.08296.
16. F. A. Zaman, X. Wu, W. Xu, M. Sonka, and R. Mudumbai, *Trust, but verify: Robust image segmentation using deep learning*, in 2023 57th Asilomar Conference on Signals, Systems, and Computers, IEEE, 2023, pp. 1070–1074.
17. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, *Learning deep features for discriminative localization*, 2015. doi:10.48550/ARXIV.1512.04150.
18. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, *Grad-cam: Visual explanations from deep networks via gradient-based localization*, in 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 618–626.
19. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, *Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks*, in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 839–847.
20. H. Wang, et al., *Score-cam: Score-weighted visual explanations for convolutional neural networks*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2020, pp. 111–119.
21. P. Natekar, A. Kori, and G. Krishnamurthi, *Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis*, Front. Comput. Neurosci. 14 (feb 2020) 6.
22. H. Saleem, A. R. Shahid, and B. Raza, *Visual interpretability in 3d brain tumor segmentation network*, Comput. Biol. Med. 133 (2021) 104410.
23. Y. Liu, et al., *Mixed-UNet: Refined class activation mapping for weakly-supervised semantic segmentation with multi-scale inference*, Front. Comput. Sci. 4 (2022) 135.
24. L. Zhu, et al., *A multi-task two-path deep learning system for predicting the invasiveness of craniopharyngioma*, Comput. Methods Programs Biomed. 216 (2022) 106651.
25. S. Bakas, et al., *Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge* (2018). doi:10.48550/ARXIV.1811.02629.
26. V. Petsiuk, A. Das, and K. Saenko, *Rise: Randomized input sampling for explanation of black-box models* (2018). doi:10.48550/ARXIV.1806.07421.
27. R. Raza, U. I. Bajwa, Y. Mehmood, M. W. Anwar, and M. H. Jamal, *dResU-net: 3d deep residual u-net based brain tumor segmentation from multimodal MRI*, Biomed. Signal Process. Control 79 (2023) 103861.
28. J. Sun, Y. Peng, D. Li, and Y. Guo, *Segmentation of the Multimodal Brain Tumor Images Used Res-U-Net*, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12658 LNCS (2021) 263–273.
29. Q. Zuo, S. Chen, and Z. Wang, *R2AU-Net: Attention Recurrent Residual Convolutional Neural Network for Multimodal Medical Image Segmentation*, Secur. Commun. Netw. 2021 (2021) 1-10.
30. M.U. Saeed, et al., *RMU-Net: A Novel Residual Mobile U-Net Model for Brain Tumor Segmentation from MR Images*, Electronics 10 (2021) 1962.
31. N. M. AboElenein, P. Songhao, and A. Afifi, *IRDNU-Net: Inception residual dense nested u-net for brain tumor segmentation*, Multimed. Tools Appl. 81 (2022) 24041–24057.
32. J. Hu, X. Gu, Z. Wang, and X. Gu, *Mixture of calibrated networks for domain generalization in brain tumor segmentation*, Knowl. Based Syst. 270 (2023) 110520.
33. M. Lerma and M. Lucas, *Grad-CAM++ is equivalent to Grad-CAM with positive gradients*, in 24th Irish Machine Vision and Image Processing Con-

ference (IMVIP), Irish Pattern Recognition and Classification Society, 2022, pp. 113-120.
34. N. O. Pinciroli Vago, F. Milani, P. Fraternali, and R. da Silva Torres, *Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis*, Journal of Imaging 7 (2021) 106.