2025 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA) | 979-8-3315-3308-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/COGSIMA64436.2025.11079472

AXONS-3: An XAI-Augmented Approach for Advancing Trust and Transparency in 3D Brain Tumor Segmentation

Jacqueline Abyasa
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
jacqueline.abyasa@binus.ac.id

Rissa Rahmania
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
rissa.rahmania@binus.ac.id

***Abstract*—Early brain tumor detection remains a critical challenge in medicine due to its impact on patient outcomes. While magnetic resonance imaging (MRI) is a key tool, the challenges accumulated from the grayscale nature of MRI and high volume of data, coupled with human cognitive limitations and time pressures in radiology, create a potentially large margin of diagnostic uncertainty and error. Despite their performance, deep learning solutions face resistance in clinical adoption due to the lack of trust that roots from the opacity of such models. This research introduces the *AXONS-3* workflow with the aim of bridging model outputs with the practical needs of clinicians by integrating interpretability and transparency into artificial intelligence (AI) systems for clinical decision-making. First, a 3D U-Net model is trained on the BraTS2020 dataset using T1-Gd, T2, and FLAIR MRI sequences to segment brain tumors into sub-regions of NCR/NET, ED, and ET. Then, post-hoc visual Explainable AI (XAI) techniques, including gradient-based methods and uncertainty quantification, are augmented to the workflow to interpret the process of reaching the predicted segmentation. The proposed *AXONS-3* workflow provides visually intuitive feedback and justifications to foster greater stakeholder comprehension and trust, contributing to the transparency of AI-driven systems needed for reliable adoption in clinical settings.**

***Keywords*—3D U-Net, brain tumor segmentation, deep learning, Explainable AI (XAI), healthcare AI, interpretability, neuroimaging, post-hoc, transparency**

I. Introduction

Brain tumors in general are categorized as benign (non-cancerous) or malignant (cancerous). Regardless of their type, brain tumors are potentially life-threatening due to the disruptions they create toward surrounding brain tissue [1]. Early detection of tumors enables early diagnosis which is critical to the patient's outcome. Magnetic resonance imaging (MRI) is a non-invasive medical imaging technique that is commonly used for brain imaging (neuroimaging) and is effective for visualizing soft tissues, blood vessels, and abnormalities. However, the resulting high volume of images/slices presents challenges for manual evaluation by clinicians. The cognitive and visual system has limitations in focus and sensitivity, creating a margin for human error [2], [3]. The repetitive nature of this task is mentally demanding because each slice is examined and requires high focus to accurately interpret the results [4]. At a certain point, fatigue coupled with neurocognitive limitations could accumulate and start to negatively affect diagnostic accuracy, resulting in errors, delayed diagnoses, and inconsistent decision-making [5], [6].

Deep Learning (DL) models, particularly Convolutional Neural Networks (CNNs), are highly prospective in the context of neuroimaging because of their ability to catch intricate characteristics in the image, which are often not as easily distinguished by the human eye, mitigating missed detections [7], [8]. Among these, the 3D U-Net architecture [9] stands out for its ability to process volumetric data, making it ideal for brain tumor segmentation in MRI [10], [11]. The integration of AI in neuroimaging analysis acts as a tool that enhances, aids, and supports the work of clinicians, providing cues/hints to areas of interest in a medical image, allowing clinicians, specifically radiologists, to spend less cognitive energy in manually dissecting each image, and instead focus on higher-level tasks such as diagnostic synthesizing. However, the implementation of these systems requires not only high technical accuracy but also features like interpretability and transparency to gain the trust of clinicians and patients.

Friction in implementing AI in healthcare prevails due to the system's opacity in its decision-making. For AI systems to be integrated effectively, clinicians must trust the models and understand how predictions are generated [12]. According to a paper by [13], the rate of model interpretability decreases as model accuracy increases. Higher-accuracy, "black-box", DL models are considered as low interpretability methods because of their architectural complexity which involves a significantly higher number of trainable parameters, complicating the process of tracing their decision boundaries. This results in a lack of trust and transparency from users involved such as clinicians and patients, especially where clinical decisions can have profound consequences [14].

Explainable AI (XAI) is an emerging field that addresses the challenge of interpretability of AI systems, enabling end-

users to better understand and comprehend their decisions [15]. In the context of CNNs, visual XAI methods provide intuitive feedback and justifications by revealing the regions that influence model predictions, enabling clinicians to assess their validity and identify potential areas of uncertainty [16], [17]. Providing a rationale for a model's predictions alongside the development of an AI-driven brain tumor segmentation system enhances interpretability, bridging the gap between technical performance and clinical usability, with a study [18] proving that explainable systems receive more positive feedback on the acceptability of the system compared to those without. This alignment with the needs of clinicians and patients fosters a higher level of trust and confidence among users, advancing seamless adoption of these AI systems in clinical settings.

The contributions of this research are as follows.

- Presents an enhanced brain tumor segmentation process that redefines conventional radiology workflows by addressing existing challenges, optimizing diagnostic accuracy, and ensuring reliability.
- Analyzes the utility of post-hoc visual XAI techniques for increasing interpretability and transparency by bridging the gap between AI predictions and confidence from clinicians and non-technical stakeholders.
- Provides insights and reference points for future research, offering actionable recommendations for integrating XAI, uncertainty quantification, risk considerations, along with strategies for real-world clinical implementation.

II. Related Works

This research draws on recent studies in deep learning for brain tumor segmentation, resistance to AI in healthcare, segmentation interpretability, XAI stakeholders, and XAI guidelines with a user-oriented approach. Throughout the years, researchers have continuously aimed to improve diagnostic accuracy and mitigate human errors using probabilistic methods. Specifically for brain MRI tumor segmentation, the 3D U-Net [9], an extension of the U-Net [19], has garnered significant interest among researchers, consistently obtaining exceptional results across different variations [10], [11]. However, deep learning approaches often lack transparency in their underlying decision-making processes [14].

Research on segmentation interpretability have explored methods to better understand model predictions [17], dissecting XAI taxonomies, discussing methods, and evaluation approaches. Components of XAI in a real-world context for defined groups of diverse user needs are discussed in [13], [16]. A common theme observed highlights the need for end-user comprehension and trust through making the transparency of these systems more accessible and easier to understand. A study [20] has proposed a systematic directive for developing transparent AI in medical image analysis with an emphasis on user-oriented design. Despite these advancements, there is a notable gap in connecting these fields to develop a comprehensive, tailored solution for brain tumor segmentation.

This study bridges this gap by accommodating the need to address challenges in conventional radiological processes and the resistance to AI-driven systems with notions of segmentation interpretability using a more user-oriented XAI design. It seeks to bridge these areas to create a workflow that not only undertakes the limitations of manual brain tumor segmentation, but also ensures interpretability and usability for relevant stakeholders.

III. Proposed Method

The *AXONS-3* workflow is outlined in Fig. 1. The name *AXONS-3* is inspired by the axon of a neuron and is an abbreviation for *Augmented eXplainable Outputs for Neuroimaging Segmentation in 3D*. Data collection and preparation is conducted to develop a dataset of 3D brain MRI volumes. This data is then used to train the 3D U-Net model, which segments the brain tumor regions. Once the model is trained, post-hoc visual XAI methods are applied and qualitatively evaluated to provide explanations to stakeholder groups.

A. Data Collection and Experiment Setup

The Multimodal Brain Tumor Segmentation (BraTS2020) from the Center for Biomedical Image Computing and Analytics (CBICA) Image Processing Portal of Perelman School of Medicine at the University of Pennsylvania is used as the data source [21]–[23]. The BraTS2020 dataset includes 3 main classes: necrotic & non-enhancing tumor core (NCR/NET), peritumoral edema (ED), GD-enhancing tumor (ET); non-tumor regions and background are labeled 0. The data is originally given as a 3D volume and in the Neuroimaging Informatics Technology Initiative (NIfTI) format.

A combined stack of T1-Gd, T2, and FLAIR MRI channels is used. To optimize storage space and processing time, the data is cropped to focus on the area of the brain on the axial, sagittal, and coronal axes of the MRI. The resulting data file has a dimension of (3, 128, 128, 128). This dimension represents the 3D brain volume, in which each voxel has 3 associated values corresponding to the 3 MRI channels that are used simultaneously. The preprocessed dataset consists of 300 samples and is split into three parts: 75% for training, 20% for validation, and 5% for testing. Due to the relatively large dataset volume and limitations in computational resources, data augmentation is not implemented in this research.

B. Segmentation with 3D U-Net Model

This study employs a segmentation model based on the 3D U-Net architecture, illustrated in Fig. 2, which is adequately trained to serve as a viable medium for implementing and analyzing XAI methods. The 3D U-Net starts with a contracting path where in each block, a sequence of a $3 \times 3 \times 3$ 3D Convolution, a batch normalization, followed by a ReLU activation, is applied twice. Each sequence produces an output of the same size as the input. Max pooling is done to decrease the spatial dimensions of the 3D feature map while increasing the number of feature filters. In the expanding path, a key element of the U-Net, skip connections, are performed

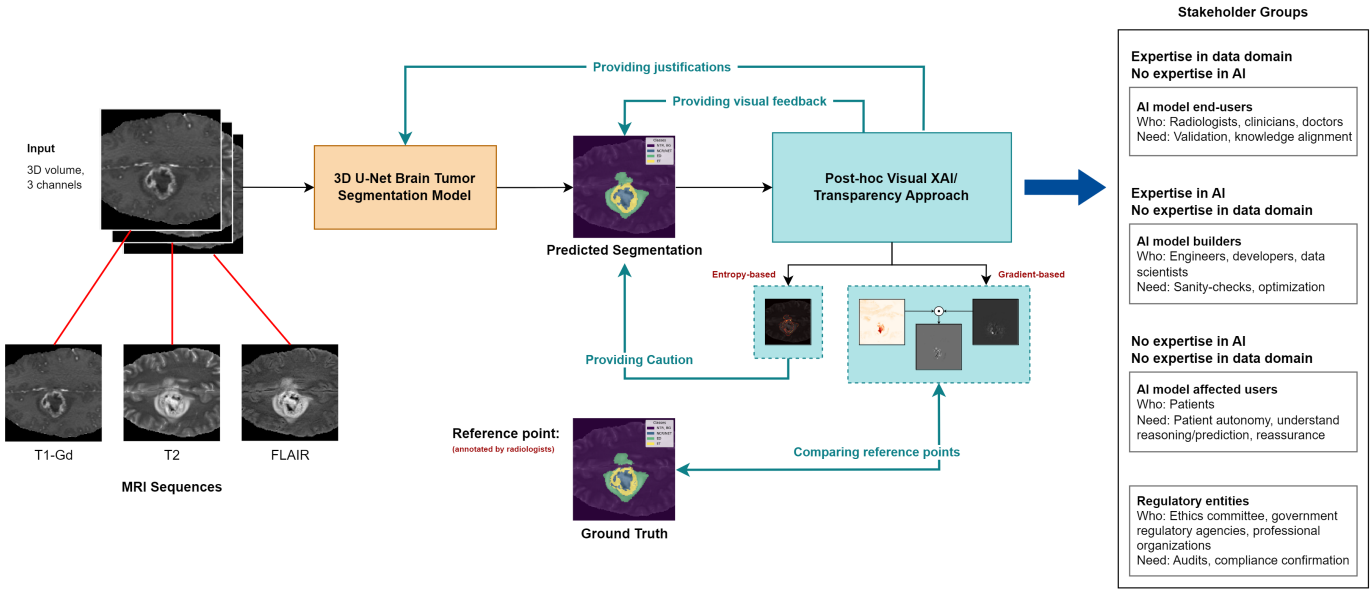

Fig. 1. *AXONS-3* workflow

where feature concatenation from a corresponding layer in the contracting path is performed to the upsampled output. This is done with the objective of reusing information and providing contextual guidance from fine-grained details learned in the contracting path at the same dimension level, allowing the network to leverage both global context and local details to reconstruct the spatial details, leading to more accurate segmentation results. Training is carried out with a batch size of 1, through 100 epochs, using the Adam and AdamW optimizers with a learning rate of 0.0001.

Dice coefficient is used as an evaluation metric to measure model performance. Dice coefficient, also called the Dice Similarity Coefficient (DSC), is a commonly used metric for 3D medical image segmentation tasks [24] to compare the overlap between the predicted segmentation and the ground truth segmentation. This measures the similarity between the two sets, as defined in (1):

$$Dice = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{1}$$

where A is the predicted segmentation, B is the ground truth, $|A \cap B|$ represents the intersection (overlap) of the two sets, and $|A| + |B|$ is the sum of their sizes. The Dice coefficient ranges from 0 to 1, where 1 indicates a perfect match, and 0 indicates no overlap at all. The type of Dice coefficient used in this research is the soft Dice approach which leverages predicted probabilities directly, instead of converting them into one-hot encoded binary masks, reflecting a more refined representation of the model's confidence in each class. This approach aims to take the existing layer of uncertainty into account and mitigate creating misleadingly high confidence levels.

C. XAI Implementation

From the broad taxonomy of XAI, this research focuses on post-hoc visual XAI techniques, including Grad-CAM, Guided Backpropagation, Guided Grad-CAM, and entropy-based uncertainty quantification, to provide insights to explain the model's decision-making process for the resulting segmentation predictions. As human cognitive abilities favor the understanding of visual data [13], Visual XAI is used since it is more intuitive and straightforward in presenting complex AI concepts to non-technical end-users. Post-hoc explanations are generated after the model is trained and are done completely without altering the model.

1) Grad-CAM: Gradient-weighted Class Activation Mapping (Grad-CAM) [25], is a gradient-based approach that visualizes class-discriminative regions of an image that contribute most to the model's prediction by analyzing the gradient information that goes into the last convolutional layers of the CNN. Convolutional layers are used on the basis that they retain spatial information, notably in the last convolutional layers, which have an ideal balance of both detailed spatial information and high-level semantics. Each neuron in these layers is assigned importance values for a particular decision of interest. In a 3D context, a modified version of the original Grad-CAM equation is used. The 3D class-specific localization map is defined as: $\mathcal{L}^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v \times w}$ of width u, height v, and depth w for any class c. Equation (2) acquires the neuron importance weights α^c_k through backpropagation by taking the gradient of the given score of class c, y^c, with respect to feature map activations A^k of a convolutional layer, resulting in $\frac{\partial y^c}{\partial A^k}$. This is averaged spatially, using global-average-pooled (GAP) over the width, height, and depth dimensions (indexed by p, q, and r respectively), with Z being the number of voxels in the feature map.

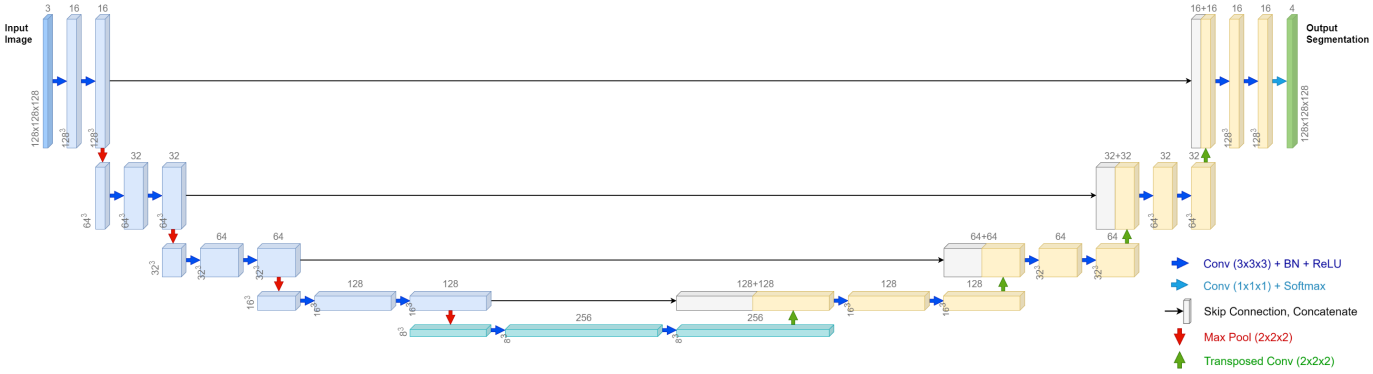

Fig. 2. 3D U-Net architecture

$$\alpha_k^c = \underbrace{\frac{1}{Z}\sum_p\sum_q\sum_r}_{\text{GAP}} \overbrace{\frac{\partial y^c}{\partial A_{pqr}^k}}^{\text{gradients}} \tag{2}$$

A linear combination of feature map activations weighted by their importance is then computed. This is done from the conclusion that α_k^c represents a partial linearization of the deep network downstream from A, and captures the importance of feature map k for a target class c [25].

$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \tag{3}$$

ReLU is applied to ensure only positive contributions are visualized, as seen in (3), emphasizing features with a positive impact on the class score y^c. Negative contributions, likely associated to other classes, are excluded. This prevents highlighting unrelated voxels to the desired class. This process produces a saliency map of the same size as the convolutional feature maps, visualized as heat maps, enabling further interpretation to give context for a model's predictions.

2) Guided Backpropagation: Guided backpropagation (GBP) adds an additional guidance signal from the higher layers to the usual backpropagation, where activated neurons are guided by setting negative gradients to zero using ReLU, ultimately revealing areas where significant features are present in the image [26]. GBP modifies the standard backpropagation algorithm by applying a ReLU activation function to the gradients during the backward pass. This means that only positive gradients are preserved and propagated, while negative gradients are set to zero. This selective propagation of gradients highlights the positive contributions of neurons to the final output, effectively focusing on the most influential features, producing a more focused and informative visualization.

$$R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1} \tag{4}$$

Equation (4) represents a formal definition of GBP for propagating an output activation out back through a ReLU unit in layer l. There are 2 binary conditions: $(f_i^l > 0)$ and $(R_i^{l+1} > 0)$, which each equals 1 if the condition is true and 0 otherwise. These act as indicator functions that indicate whether f_i^l and R_i^{l+1} are positive. Both conditions need to be true (resulting in 1) in order for R_i^{l+1} to be considered as a relevant and contributing signal to be propagated back, preventing the backward flow of negative gradients. It should be noted, however, that GBP does not compute a true gradient but an approximated version instead.

3) Guided Grad-CAM: Guided Grad-CAM is a variation of Grad-CAM that fuses the class-discriminative nature of Grad-CAM with the fine-grained detail of GBP [25]. It produces visualizations that are both high-resolution, identifying fine-grained details, and class-specific, highlighting only the regions relevant to the predicted class. As Grad-CAM, in itself, produces relatively coarse saliency maps, Guided Grad-CAM pursues a more refined, voxel-level visualization that gives insight to which exact voxels contribute to the prediction that was made. To combine the two methods, element-wise multiplication is performed, as in (5), obtaining a 3D Guided Grad-CAM, class-specific localization map $L^c_{\text{Guided Grad-CAM}}$:

$$L^c_{\text{Guided Grad-CAM}} = L^c_{\text{Grad-CAM}} \odot L_{\text{Guided Backpropagation}} \tag{5}$$

4) Entropy-Based Uncertainty Quantification: The uncertainty map is calculated based on Shannon Entropy. The Shannon entropy H quantifies the expected uncertainty inherent in the possible outcomes of a discrete random variable [27]. The probability distribution in this case is the output of probabilities of a voxel belonging to a certain class generated by the softmax activation function in the final layer. The entropy values are then normalized for visualization. As formulated in (6), H is the entropy of the set of probabilities $p_1, ..., p_n$:

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{6}$$

In this implementation, Shannon entropy is employed to quantify the degree of uncertainty in the voxel-wise predictions

of the 3D U-Net segmentation model. Higher entropy values indicate greater uncertainty, as the probabilities are more evenly distributed across classes, whereas lower entropy values suggest higher confidence, with the probabilities skewed toward a particular class. This entropy calculation results in an uncertainty map, providing a spatial representation of the model's confidence in its predictions, which is particularly valuable for understanding model reliability in critical areas of segmentation. The map is normalized for enhanced visualization.

IV. Results and Analysis

This section discusses the results of the experiment performed in this research along with its corresponding evaluation and analyses.

A. 3D U-Net Brain Tumor Segmentation Model

The best validation score is achieved at epoch 87 with a Dice coefficient (soft Dice) of 72.66%, which reflects the model's ability to generalize at that stage of training. The slightly lower test set Dice coefficient of 70.98%, as presented in Table I, suggests that the model generalizes reasonably well to completely unseen data, with a minor drop in performance. The validation and test Dice coefficients in the low-70% range indicate the model is reasonably effective, having a sufficient overlap between the predicted and ground truth segmentations, but may still struggle with certain class-specific challenges. The drop in the test Dice coefficient might be due to differences in data distribution between the validation and test sets or the inherent challenge of the task.

The model's performance is evaluated by analyzing Dice coefficients across four classes: necrotic/non-enhancing tumor core (NCR/NET), peritumoral edema (ED), GD-enhancing tumor (ET), and non-tumor regions/background (NTR/BG). This evaluation of Dice coefficients per class provides a more nuanced understanding of the model's performance across different segmentation categories. On the validation set, the model shows varying performance across different tumor classes, with ED, Class 2, having the highest validation performance among tumor-related classes, exhibiting a gradual increase and fluctuations around 60-71% in later epochs. ED is the outermost tumor sub-region, therefore having to also establish the boundary between the tumor within the complexity of the brain structure. Non-tumor regions, Class 0, achieve consistently high Dice scores with minimal training-validation gaps, indicating accurate background segmentation. NCR/NET, Class 1, is learned at a slower rate than other classes, gradually improving to a range of 45-54% toward the end of training. Consistent validation Dice improvement indicates continued benefit of the model from training and progressive adaptation to the complexity of brain tumor segmentation. It is worth noting that misclassification in NCR/NET voxels can propagate errors in ET and even ED, causing the overall Dice coefficient to decrease. ET, Class 3, fluctuates between 55-67%, however, the presence of a smaller gap between training and validation suggests the model achieves a better balance between training and validation performance, generalizing well on unseen data. The distinct, bright appearance of ET regions in T1-Gd potentially gives an advantage to the model to learn features for ET that transfer well between training and validation sets.

TABLE I
Test set Dice coefficient for each class and averaged score

Dice Coefficient				
NTR, BG	**NCR/NET**	**ED**	**ET**	**Average**
96.93%	53.34%	69.54%	64.10%	70.98%

B. XAI Approaches

This section discusses the XAI approaches, including gradient-based methods and an entropy-based method.

1) Gradient-Based: The three gradient-based, visual XAI techniques used in this research, Grad-CAM, Guided Grad-CAM, and Guided Backpropagation, provided insights into the 3D U-Net model's decision-making process through saliency maps, highlighting the regions of the input images that were most important and have high influence in the model's decision process.

Grad-CAM highlights broader regions but tends to produce noisier and vaguer heatmaps, offering a general sense of the model's focus. Guided Grad-CAM, on the other hand, provides a cleaner and more focused visualization on a voxel-level scale because of its GBP foundations, which may be more helpful for clinicians. GBP focuses on the whole image rather than by class, visualizing edges and textures, and highlighting areas of high saliency at a more detailed, voxel-level scale. For example, it may align with sharp anatomical boundaries visible in the MRI, but since GBP takes an imputed form of the gradients used, it might not fully correspond to clinical patterns. The finer detail captured by GBP is retained by Guided Grad-CAM, while also incorporating the class-discriminative nature of Grad-CAM, allowing a more precise view of which parts of the tumor or its surroundings contribute most to the model's predictions.

A previous study [16] has suggested that comparison between the regions of interest, considered as reference points, from the decision made by the model is to be compared to one of a radiologist's. By comparing the XAI visualizations to a radiologist's reference points, we can assess how well these techniques align with human intuition and decision-making processes. However, in this research, since the ground truth segmentations were directly annotated by professional radiologists, it serves as the acting reference point and known clinical prior, in place of a live validation process.

Fig. 3 demonstrates the performance of Grad-CAM and Guided Grad-CAM in highlighting salient regions of the tumor in a class-discriminative manner. Grad-CAM captures the region of the GD-enhancing tumor (ET) quite well with strong gradients, though with some overgeneralization. Guided Grad-CAM refines this by reducing noise, indicating the model's reliance on high-intensity signals to delineate ET regions

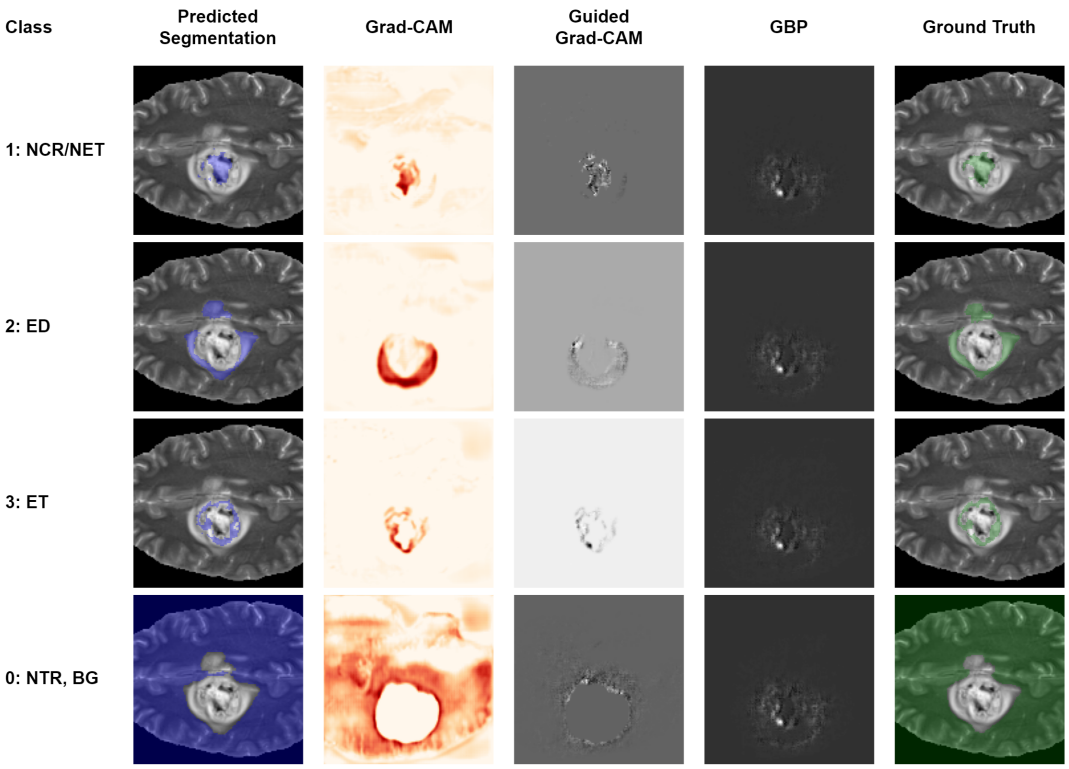

Fig. 3. Predicted segmentations and gradient-based XAI maps (Grad-CAM, Guided Grad-CAM, GBP) for each class (NCR/NET, ED, ET, NTR/BG), compared with clinically annotated ground truth

confidently. Non-tumor regions (NTR, BG) were identified with strong activations at the periphery of the tumor, reflecting the model's capability to distinguish non-tumor regions with high confidence. However, the necrotic/non-enhancing tumor core (NCR/NET) and peritumoral edema (ED) regions were less accurately segmented, especially where activations are more diffuse and sometimes misaligned with the ground truth. This reflects the inherent challenges posed by these regions, in which the lack of distinct visual cues in certain MRI sequences (e.g., T1-Gd for NCR/NET or FLAIR for ED) makes segmentation more ambiguous. Additionally, this can also be visualized in 3D to further demonstrate how the activations extend in a 3D space, providing a more intuitive understanding for clinicians, as it mimics the real-world volumetric analysis of MRI scans.

It should be noted, however, that minor discrepancies exist between the XAI visualizations and the ground truth. If the saliency intensities were high, this potentially indicates over-reliance on non-specific features, highlighting areas where the model may benefit from additional supervision to address these cases. However, since the saliency intensities are proportionately minor to the ones concerning the main tumor, it could be assumed that these areas are used as checkpoints to assess and compare surrounding tissue relative to the main tumor regions. Based on image characteristics, input sequences, especially T1-Gd, seem to be a significant guide to the model's understanding of NCR/NET and ET, while all three sequences (T1-Gd, T2, and FLAIR) contribute to ED segmentation.

Comparison with radiologists' interpretations reveals that, despite some overgeneralization, the model generally highlights clinically relevant regions. However, overgeneralization in medical segmentation directly relates to the balance between false positives (FPs) and false negatives (FNs), both of which carry distinct risks in medical contexts. FPs refer to regions incorrectly predicted as pathological (e.g., tumor, lesion) when they are actually healthy tissue. In contrast, FNs refer to regions predicted as healthy tissue, failing to detect true pathological regions.

Determining whether it is better to overestimate (resulting in more FPs) or underestimate (resulting in more FNs) depends on the clinical scenario and the consequences of these errors. In conditions where delaying or completely missing a diagnosis could be fatal or lead to rapid disease progression, especially in the context of malignant brain tumors, overestimation is preferred. While increased FPs could provide false alarms, it is more preferable to minimize FNs, as it favors patient safety, prioritizing sensitivity to reduce the risk of overlooking potential abnormalities. It provides an opportunity to rule out diseases through further diagnostic tests, ensuring critical cases are not missed. For benign tumors, slight underestimations of the region of interest might be more tolerable, with the preliminary objective of avoiding excessive intervention or overtreatment while prioritizing the most critical areas.

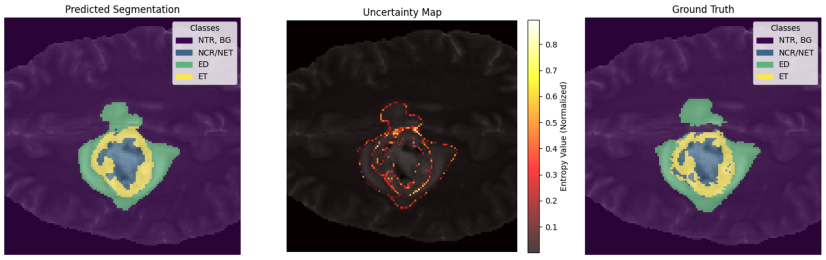

Fig. 4. Predicted segmentation and corresponding uncertainty map compared to clinically annotated ground truth

2) Entropy-Based: Fig. 4 presents the resulting uncertainty map, calculated based on the normalized Shannon entropy for each voxel, which provides insight into the model's confidence in its predictions for brain tumor segmentation across different classes. By quantifying uncertainty at the voxel level, this visualization highlights regions where the model exhibits lower confidence, as the probabilities are more evenly distributed across classes (higher entropy), which is particularly valuable in a clinical setting where errors can have significant implications. In the uncertainty map, areas of high uncertainty, indicated by higher entropy values, are predominantly located at the boundaries between tumor subregions and the surrounding non-tumor tissue, as well as in regions where class overlap or ambiguity is common. This aligns with the expected behavior of segmentation models, as uncertainty often increases in transition zones due to the subtle intensity and mixed tissue types.

Uncertainty is particularly elevated at the boundaries between different tumor classes and between tumor regions and non-tumor regions. This is expected since these areas represent transitions where the model struggles to accurately delineate the exact borders of the tumor, as they often have irregular shapes and can diffuse/blend with surrounding tissues. Increased entropy is seen in boundaries between ET and ED, NCR/NET and ET, and between ED and non-tumor regions. The sharpness of these boundary uncertainties may also indicate how well the model has been trained to resolve transitions between classes.

The distribution of uncertainty highlights the complexity of each class in terms of its visual features and separability from other classes. ED regions, being diffuse and less distinct in T2 and FLAIR sequences, exhibit higher uncertainty, while ET regions, with clearer boundaries in T1-Gd sequences, show lower uncertainty. Within tumor subregions, uncertainty is generally low, suggesting the model's confidence in clearer regions but greater uncertainty near boundaries, often linked to FPs or FNs. Through these observations, potential sources of uncertainty include class imbalance (e.g., NCR/NET's smaller voxel volume), the model's difficulty generalizing to diffuse classes like ED, and variability in patient anatomy, tumor presentation, and imaging quality, especially in unseen data.

From a clinical perspective, these uncertainty maps are to be incorporated for risk assessment and decision-making. These regions could be flagged for closer inspection by clinicians, as they may represent areas where the voxel's probability distribution is more evenly spread across the classes, indicating high-potential regions of FPs and FNs. High uncertainty regions can guide clinicians to exercise caution and consider further review or additional imaging sequences to verify results. Moreover, explicitly visualizing areas of low confidence allows for increased transparency by identifying regions where the model's predictions are less reliable. This transparency enables clinicians to trust the model's predictions while recognizing its limitations.

V. Conclusion

The 3D U-Net model closely segmented brain tumor regions to ones segmented by radiologists, with the exception of existing challenges in accurately defining necrotic and non-enhancing tumor core (NCR/NET) and peritumoral edema (ED) regions due to boundary ambiguities within the brain's highly intricate anatomical structure. The *AXONS-3* workflow addresses resistance in adopting these systems through the integration of saliency and uncertainty maps with segmentation results and clinically annotated segmentation, offering visually intuitive feedback and justifications that are designed to foster greater stakeholder comprehension and trust. An increased understanding of model behavior and clinical relevance for brain tumor segmentation tasks improves reliability and interpretability in clinical diagnostics. Gradient-based methods, such as Grad-CAM, Guided Grad-CAM, Guided Backpropagation, enhanced the interpretability of the 3D U-Net's predicted brain tumor segmentation through visualizing salient regions derived from intermediate-level features and spatial activations. Entropy-based uncertainty maps provided a voxel-wise representation of uncertainty based on the resulting probability distribution for each voxel across four classes. Uncertainty arises from the difficulty in generalizing diffuse boundaries, class imbalance, and inherent variability in patient anatomy, tumor morphology, and imaging quality that challenge the model's adaptability to unseen cases. High-entropy regions indicate potential regions of FPs or FNs, providing prediction transparency and aiding clinicians in identifying less reliable predictions. Managing FPs and FNs is crucial, in which the medical context dictates whether overestimation or underestimation is more tolerable should a choice arise.

Future research would be oriented toward evaluating XAI techniques in real-world workflows, such as having radiolo-

gists annotate MRI data with descriptive comments for labeling decisions along with establishing standardized evaluation metrics that are directly tailored to clinical priorities. This process creates a richer dataset for training AI systems and enhances XAI evaluation, especially for assessing the impact of reference points and segmentation errors on treatment planning and outcomes. Developing clinically oriented benchmarks for evaluating both segmentation accuracy and XAI outputs will further ensure these AI systems are reliable, actionable, and aligned with healthcare needs. Additionally, risk-aware modeling could also be done to address the existence of FPs and FNs through assigning higher penalties for FNs in critical areas during training while minimizing FPs in less critical areas.

ACKNOWLEDGMENTS

This research is supported by the Research and Technology Transfer Office of Bina Nusantara University through the university's STARS grant.

AUTHOR CONTRIBUTIONS

Jacqueline Abyasa is the primary contributor to this research, responsible for dataset preparation, model development, XAI implementation, result analysis, workflow integration, and manuscript organization. Rissa Rahmania provided essential supervision and guidance throughout the research process. All authors have approved the final manuscript.

REFERENCES

[1] L. M. DeAngelis, "Brain tumors," *The New England Journal of Medicine*, vol. 344, no. 2, pp. 114–123, Jan 2001. [Online]. Available: https://doi.org/10.1056/nejm200101113440207

[2] M. A. Bruno, "256 Shades of gray: uncertainty and diagnostic error in radiology," *Diagnosis*, vol. 4, no. 3, pp. 149–157, 7 2017. [Online]. Available: https://doi.org/10.1515/dx-2017-0006

[3] L. Barger, N. T. Ayas, B. E. Cade *et al.*, "Impact of Extended-Duration shifts on medical errors, adverse events, and attentional failures," *PLoS medicine*, vol. 3, no. 12, p. e487, 12 2006. [Online]. Available: https://doi.org/10.1371/journal.pmed.0030487

[4] S. Taylor-Phillips and C. Stinton, "Fatigue in radiology: a fertile area for future research," *The British Journal of Radiology*, vol. 92, no. 1099, p. 20190043, 7 2019. [Online]. Available: https://doi.org/10.1259/bjr.20190043

[5] H. H. Abujudeh, G. W. Boland, R. Kaewlai *et al.*, "Abdominal and pelvic computed tomography (CT) interpretation: discrepancy rates among experienced radiologists," *European radiology*, vol. 20, no. 8, pp. 1952–1957, 3 2010. [Online]. Available: https://doi.org/10.1007/s00330-010-1763-1

[6] G. A. Pignatiello, R. J. Martin, and R. L. Hickman, "Decision fatigue: A conceptual analysis," *Journal of health psychology*, vol. 25, no. 1, pp. 123–135, 3 2018. [Online]. Available: https://doi.org/10.1177/1359105318763510

[7] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, pp. 94–98, 6 2019. [Online]. Available: https://doi.org/10.7861/futurehosp.6-2-94

[8] T. Kalaiselvi, T. Padmapriya, P. Sriramakrishnan, and K. Somasundaram, "Advancements of MRI-based Brain Tumor Segmentation from Traditional to Recent Trends: A Review," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 18, no. 12, pp. 1261–1275, 12 2021. [Online]. Available: https://doi.org/10.2174/1573405617666211215111937

[9] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," *arXiv (Cornell University)*, 1 2016. [Online]. Available: https://arxiv.org/abs/1606.06650

[10] A. Verma, S. N. Shivhare, S. P. Singh, N. Kumar, and A. Nayyar, "Comprehensive Review on MRI-Based Brain Tumor Segmentation: A Comparative Study from 2017 Onwards," *Archives of Computational Methods in Engineering*, 5 2024. [Online]. Available: https://doi.org/10.1007/s11831-024-10128-0

[11] A. Avesta, S. Hossain, M. Lin *et al.*, "Comparing 3D, 2.5D, and 2D approaches to brain Image Auto-Segmentation," *Bioengineering*, vol. 10, no. 2, p. 181, 2 2023. [Online]. Available: https://doi.org/10.3390/bioengineering10020181

[12] C. Longoni, A. Bonezzi, and C. K. Morewedge, "Resistance to medical artificial intelligence," *The journal of consumer research/Journal of consumer research*, vol. 46, no. 4, pp. 629–650, 5 2019. [Online]. Available: https://doi.org/10.1093/jcr/ucz013

[13] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 12 2019. [Online]. Available: https://doi.org/10.1016/j.inffus.2019.12.012

[14] R. Yokoi, Y. Eguchi, T. Fujita, and K. Nakayachi, "Artificial Intelligence Is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity," *International journal of human-computer interaction*, vol. 37, no. 10, pp. 981–990, 12 2020. [Online]. Available: https://doi.org/10.1080/10447318.2020.1861763

[15] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 10 2018. [Online]. Available: https://doi.org/10.1016/j.artint.2018.07.007

[16] K. Borys, Y. A. Schmitt, M. Nauta *et al.*, "Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches," *European Journal of Radiology*, vol. 162, p. 110787, 3 2023. [Online]. Available: https://doi.org/10.1016/j.ejrad.2023.110787

[17] R. Gipiškis, C.-W. Tsai, and O. Kurasova, "Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey," *ICT Express*, 9 2024. [Online]. Available: https://doi.org/10.1016/j.icte.2024.09.008

[18] P. Sabol, P. Sinčák, P. Hartono *et al.*, "Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images," *Journal of Biomedical Informatics*, vol. 109, p. 103523, 8 2020. [Online]. Available: https://doi.org/10.1016/j.jbi.2020.103523

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-NET: Convolutional Networks for Biomedical Image Segmentation," *Cornell University*, 1 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[20] H. Chen, C. Gomez, C.-M. Huang, and M. Unberath, "Explainable Medical Imaging AI Needs Human-Centered Design: Guidelines and Evidence from a Systematic Review," *arXiv (Cornell University)*, 1 2021. [Online]. Available: https://arxiv.org/abs/2112.12596

[21] B. H. Menze, A. Jakab, S. Bauer *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *PMC*, 10 2015. [Online]. Available: https://dspace.mit.edu/handle/1721.1/110992

[22] S. Bakas, H. Akbari, A. Sotiras *et al.*, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, no. 1, 9 2017. [Online]. Available: https://doi.org/10.1038/sdata.2017.117

[23] S. Bakas, M. Reyes, A. Jakab *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," 2018. [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/291597

[24] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, 8 2015. [Online]. Available: https://doi.org/10.1186/s12880-015-0068-x

[25] R. R. Selvaraju, M. Cogswell, A. Das *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *2017 IEEE International Conference on Computer Vision (ICCV)*, 10 2017. [Online]. Available: https://doi.org/10.1109/iccv.2017.74

[26] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: the all convolutional net," *arXiv (Cornell University)*, 1 2014. [Online]. Available: https://arxiv.org/abs/1412.6806

[27] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 7 1948. [Online]. Available: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x