IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 July 2025, accepted 17 July 2025, date of publication 24 July 2025, date of current version 1 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3592239

RESEARCH ARTICLE

Visualizing UNet Decisions: An Explainable AI Perspective for Brain MRI Segmentation

D. JEYA MALA[1], (Member, IEEE), MAINAK CHATTOPADHYAY[1], PARTHIBA MUKHOPADHYAY[1], AND ROOPAK SINHA[2], (Senior Member, IEEE)

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India
[2]School of Information Technology, Deakin University, Burwood, VIC 3125, Australia

Corresponding author: D. Jeya Mala (jeyamala.d@vit.ac.in)

This work was supported in part by the Department of Science and Technology (DST), India, under the Fund for Improvement of Science and Technology Infrastructure in Universities and Higher Educational Institutions (FIST) Program under Grant SR/FST/ET-I/2022/1079; and in part by VIT University.

ABSTRACT In recent years, medical image analysis, particularly neuroimaging, has experienced remarkable advancements, with Magnetic Resonance Imaging (MRI) greatly helping in diagnosing complex neurological disorders, including brain tumors. However, accurately segmenting brain tumors from MRI scans remains a significant challenge, necessitating sophisticated computational techniques. This article presents research findings from brain MRI segmentation utilizing the UNet architecture and enhancing model interpretability using explainable AI methods to harness UNet's effectiveness in semantic segmentation tasks. We study intricacies of UNet's adaptation to brain MRI segmentation, the dataset employed, and the methodology for model development, training, and validation. In addition to discussing segmentation outcomes, we incorporate several explainable AI (XAI) techniques like Grad-CAM, Saliency Maps, Vanilla Gradient and Layer-wise Relevance Propagation (LRP) to generate necessary visualizations showing the internal workings of the opaque system nature of UNet. A comprehensive analysis of the results highlights the clinical implications of these findings, addressing the relative utility of different XAI methods in visualizing UNet's outputs using metrics like fidelity, unambiguity and stability. The Vanilla Grad method stands out with its high unambiguity and consistent fidelity scores in complex scenarios. We also find that while LRP also offers high stability, the combination of high fidelity and clarity from the Vanilla Grad model makes it the preferred method for enhancing the interpretability of AI systems in brain tumor segmentation. Overall, this research work represents a significant advancement in leveraging the trustworthiness of UNet in accurate and efficient brain tumor segmentation via XAI methods, ultimately aiming to support clinicians in diagnosis and treatment planning while fostering a deeper understanding of the model's decision-making processes.

INDEX TERMS Neuro image analysis, brain MRI segmentation, UNet, explainable AI, Grad-CAM, Vanilla Grad, saliency maps.

I. INTRODUCTION

MRI or Magnetic Resonance Imaging has profoundly transformed the landscape of neurological diagnosis and treatment, allowing clinicians to dive deep inside the structure of the human brain and its complexities [1]. Among its many applications, the detection and characterization of brain tumors stand out as crucial, necessitating precise image segmentation techniques to accurately delineate tumor regions [2]. However, as advanced deep learning techniques [3], [4], [5], particularly the UNet architecture [10], become integral to these processes, the need for explainable AI (XAI) methods [11] grows increasingly important.

While traditional segmentation methods can struggle to differentiate nuanced tumor boundaries in MRI scans, deep learning solutions like UNet show more promise [6]. UNet's

The associate editor coordinating the review of this manuscript and approving it for publication was Gina Tourassi.

TABLE 1. Summary of key literature on Explainable AI (XAI) and its applications in healthcare and information systems.

Source	Focus Area	Key Contributions	Limitations
Ullah et al. [1]	Brain tumor segmentation	Combine handcrafted features with CNNs for improved segmentation	High computational complexity; challenges with interpretability & scalability
Ranjbarzadeh et al. [2]	Deep learning for brain tumor segmentation	Introduce Cascade CNN with Distance-Wise Attention (DWA) for improved accuracy	Challenges with larger tumor volumes and precision in specific regions
Ejaz et al. [3]	Hybrid method for brain tumor segmentation	Develop FKM-SOM with DWT and PCA for accurate detection	High complexity and computational demands; needs optimization for clinical use
Işin et al. [4]	Segmentation	Evaluation of CNN-based segmentation.	High computation needs; poor interpretability.
Sallam and Seddik [5]	Segmentation	Review of traditional segmentation methods.	Sensitive to noise and manual effort needed.
Liu and Xu [6]	Segmentation	Enhanced Unet with MA-FPN and edge modules.	Implementation complexity.
Xavier et al. [7]	Segmentation	HART-UNet using ResNet50 and attention.	Requires validation on broader datasets.
Huang et al. [9]	Segmentation	3D-UNet for volumetric segmentation.	Dependence on CNN generalization.
Kumar et al. [10]	Segmentation	Attention-Unet with 2D decomposition.	Needs improvement in boundary delineation.
Saranya et al. [11]	XAI	Tackled the black-box issue in ML systems.	Contextual adaptation needed.
Nagahisarchogha et al. [12]	Systematic review of XAI	Categorize XAI methods into intrinsic and post-hoc; discuss global/local interpretability	Highlights inconsistency in effectiveness, limited user comprehension, and lack of structured literature
Brasse et al. [13]	XAI in information systems	Structured review of 180 papers, identifying key XAI concepts & research directions in IS	Limited source selection, potential search term gaps, exclusion of transparent systems
Xu et al. [14]	Evolution of XAI methods	Assess techniques like LRP and saliency maps for transparency	Challenges with DNN opacity, bias in training data, and misaligned predictions
Ying et al. [15]	Explainability in Graph Neural Networks	Proposed GNNEXPLAINER, a model-agnostic method to interpret GNN predictions; validated on synthetic/real-world datasets	Computationally intensive; limited effectiveness on complex graphs
Ghorbani et al. [16]	Robustness of interpretability methods	Show saliency maps are highly sensitive to small input changes	Lacks broader method analysis
Selvaraju et al. [17]	Visual interpretability	Introduce Grad-CAM for class-discriminative visual explanations in CNNs	Coarse outputs; limited precision for fine-grained tasks
Holzinger et al. [18]	Causability in medical AI	Extend explainability to causal reasoning for medical AI	Technical focus; Lacks practical application
Kaissis et al. [19]	Federated learning	Explore federated learning and privacy-preserving techniques for medical diagnosis	High computational complexity; communication overhead limits scalability
Singh et al. [20]	XAI in medical domains	Report XAI applications in radiology and pathology, improving diagnostic accuracy	Challenging to balance complexity and interpretability
Ronneberger et al. [21]	UNet architecture with data augmentation	Elaborated the use of contracting path and a symmetric expanding path for precise localization in the UNet architecture.	Showcased that this network can be trained end-to-end and outperforms the existing sliding-window convolutional network.
Simonyan et al. [22]	Visualisation of CNNs with saliency maps	Introduce Vanilla Gradient to highlight class-specific regions	Simple but noisy; limited precision compared to later techniques
Selvaraju et al. [23]	Visual explanations for CNNs	Propose Grad-CAM to generate class-discriminative localization maps using gradients	Coarse resolution; limited pixel-level precision
Bach et al. [24]	Pixel-level interpretability in DNNs	Introduce LRP to backpropagate relevance scores through network layers	May produce noisy results; sensitive to model architecture
Zeiler & Fergus [25]	Visual insights into CNNs	Use deconvolutional networks to generate saliency maps and interpret feature activations	Interpretations can be coarse and lack class specificity
Chiaburu et al. [26]	Uncertainty in XAI	Emphasise modelling uncertainty to improve fidelity & human trust in AI explanations	Challenges in quantifying and communicating uncertainty remain
Amparore et al. [27]	LEAF framework for XAI	Proposes LEAF to assess fidelity of local linear explanations and their impact on user trust	Highlights that high fidelity alone may not guarantee trust
Burger et al. [28]	Similarity in surrogate models	Analyse impact of similarity metrics on stability of local surrogate models in text-based XAI	Instability across metrics affects explanation reliability
Leemann et al. [29]	Effect of exemplar choice	Investigates how selected examples influence trust and perceived clarity in XAI explanations	Poor exemplar choice may reduce user confidence
Pawlicki et al. [30]	Multi-metric evaluation in XAI	Using multiple metrics gives a more comprehensive assessment of explainability quality	Single-metric approaches may be misleading
Adebayo et al. [31]	Evaluating saliency maps	Emphasize contextual evaluations to improve fidelity	Lacks broader method applicability
Montavon et al. [32]	Layer-wise Relevance Propagation (LRP)	Introduce LRP to explain DNN decisions	Lacks broader method applicability
Hoffman et al. [33]	Evaluation standards for XAI	Highlight need for subjective and objective evaluation metrics for XAI	Lack of standardized metrics; challenges in capturing stakeholder perspectives
Singh et al. [34]	Evaluation of counterfactual explanations	Propose metrics for quality, interpretability & effectiveness of counterfactual explanations	Limited to counterfactuals

specialized encoder-decoder architecture, enhanced by skip connections, allows for precise localization of anatomical structures while preserving critical spatial information. This capability enables UNet to capture both global context and fine-grained details essential for accurate brain tumor segmentation [7], [8], [9].

TABLE 2. Advantages of UNet architecture and XAI techniques.

Limitation in Literature	Advantage of UNet Architecture	Advantage of XAI Techniques
High computational complexity and poor scalability (e.g., [1], [3]–[5])	UNet's encoder-decoder structure is computationally efficient with fewer parameters and skip connections for better memory use.	Enables quick generation of visual explanations without excessive overhead.
Challenges in interpretability (e.g., [1], [4], [20])	UNet's structured design with clear layer connections aids in understanding feature localization.	Grad-CAM, LRP, and Saliency Maps enhance transparency and help visualize model focus areas.
Difficulty with precise boundary segmentation (e.g., [2], [10])	UNet handles spatial information well using skip connections, leading to improved boundary delineation.	XAI helps validate how well edges and regions are being focused on in segmentation outputs.
Implementation and architectural complexity (e.g., [6], [7])	UNet is a standard, lightweight model requiring minimal tuning for medical images.	XAI frameworks like Grad-CAM can be integrated post hoc without model changes.
Poor generalizability and dataset dependence (e.g., [7], [9])	UNet generalizes well on medical image tasks and is easily adapted to 2D and 3D data.	XAI allows understanding of generalization gaps by visualizing attention drift.
Lack of structured evaluation and explainability standards (e.g., [31]–[34])	UNet provides interpretable intermediate outputs (e.g., feature maps).	LRP, Saliency, and Grad-CAM offer subjective and objective evaluation metrics for interpretability.

Deep learning models and their opaque system nature raises challenges in their trustworthiness and their interpretability in clinical settings [12]. To address these concerns, our research explores several XAI techniques for visualizing and interpreting model decisions, providing insights MRI scan features that significantly impact tumor predictions [13]. By enhancing transparency, these XAI models allow greater acceptance and integration of deep learning technologies into clinical practice [14]. This article aims to leverage UNet's segmentation capabilities while foregrounding the significance of XAI in neuroimaging [19]. We demonstrate UNet's effectiveness in identifying tumor regions and developing a classification model for discerning tumor presence, with XAI assisting with improved interpretability. Assisting advanced segmentation techniques with XAI underscores the importance of fostering trust and accountability in the deployment of deep learning solutions in neuroimaging.

II. RELATED WORKS

Table 1 shows a summary of key literature covering the application of XAI in healthcare-related information systems. A large number of works provide approaches to explain segmentation done by AI algorithms. Several works on XAI segmentation cover high-stakes areas such as the detection of brain tumors. Current approaches face challenges in efficiency (time and space complexity), scalability, sensitivity to noise, the need for manual intervention, and generalizability. These works adopt a diverse array of approaches aimed at enhancing the accuracy, interpretability, and clinical applicability of AI models in medical imaging. Ullah et al.'s hybrid model [1] illustrated the benefits of combining handcrafted features with CNNs, despite challenges in scalability. Similarly, Ranjbarzadeh et al. [2] proposed a Cascade CNN with attention mechanisms that improve segmentation accuracy but require further refinement for complex cases. The work of Ejaz et al. [3], which integrated Fuzzy K-Mean and SOM, highlights the promise of hybrid approaches in tumor detection, although optimization remains essential for clinical feasibility. The exploration of advanced architectures like UNet, RESNET50, and Hybrid Attention-Residual UNET by authors such as Xavier et al. [7] underlined the importance of model modifications to address segmentation intricacies in MRI images. Tong et al. [8] have applied UNet++ model for improved segmentation of Brain MRI images. Meanwhile, contributions from Huang et al. [9] and Kumar et al. [10] demonstrated improvements in volumetric segmentation and boundary delineation through attention mechanisms and model adaptations, pointing to promising directions for overcoming traditional limitations in brain tumor segmentation. In the realm of XAI, foundational research by Selvaraju et al. [17] on Grad-CAM and Holzinger et al. [18] on causability expanded the discourse on interpretability, emphasizing the need for intuitive explanations in medical AI. Works by Kaissis et al. [19] and Singh et al. [20] delved into privacy-preserving techniques, particularly federated learning, which protects patient data while enabling collaboration across institutions–a critical advancement for large-scale healthcare AI deployment. Collectively, these studies underscore the imperative for continued innovation and interdisciplinary collaboration in AI for healthcare.

Table 2 lists the limitations presented in the literature we surveyed and then lists the relative advantages that the UNet architecture may deliver. Compared to other techniques, UNet can deliver improved efficiency, transparency, resolution, and higher control such as through enhanced visualization support.

III. PROPOSED METHODOLOGY

The methodology proposed for the entire workflow is given in Figure 1. Each component is explained below:

1) **Dataset Loading**: The BraTS 2020 Training and Validation dataset [35] was loaded into the working environment. This dataset contains pre-operative MRI images and manual segmentation annotations for gliomas, including multiple tumor sub-regions.

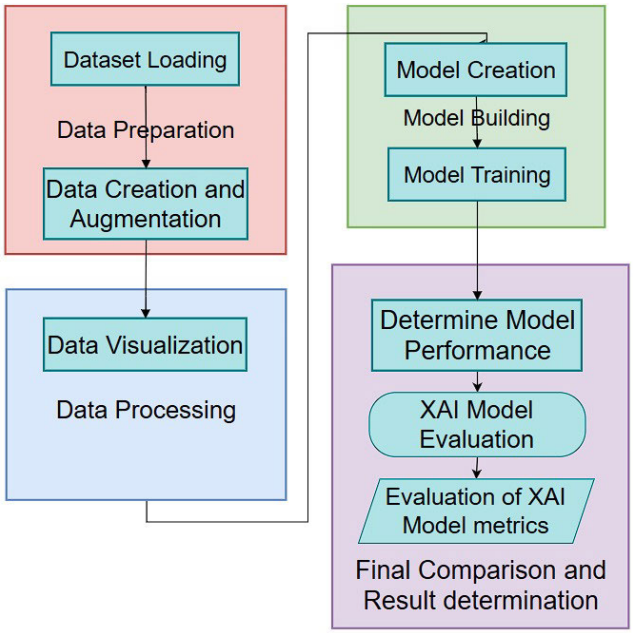

FIGURE 1. Workflow for proposed methodology.

2) **Data Creation and Augmentation**: Data generators were employed to create batches of training data [36]. Augmentation techniques such as rotation, flipping, and scaling were applied to enhance the diversity of the training dataset and prevent overfitting.
3) **Data Visualization**: Initial visualizations of the MRI images and the corresponding segmentations were performed to understand the characteristics of the dataset [38]. This included plotting sample images alongside their annotated segmentations to visually assess the quality of the data.
4) **Model Creation**: A UNet architecture was designed for the segmentation task. The model was configured with appropriate layers, activation functions, and loss functions suitable for multi-class segmentation tasks [9], [10].
5) **Model Training**: The UNet model was trained on the augmented dataset, monitoring the learning process through metrics such as loss, accuracy, and mean Intersection over Union (IoU). Learning curves were plotted to visualize the model's performance over epochs [21].
6) **Determining Training and Validation Metrics**: Training and validation metrics, including accuracy, precision, recall, F1 score, Dice coefficient, and specificity, were calculated after each epoch to assess the model's performance [21].
7) **Application of Explainable AI Techniques**: Explainable AI techniques were applied to the model's predictions to generate visual explanations of the segmentation results [20].
8) **Evaluation of Explainability Metrics**: For each explainable AI model, different metrics were computed to quantify the reliability and effectiveness of the explanations. These metrics provided insights into how well the visualizations aligned with the model's predictions [28], [30], [33].
9) **Visualizations and Explanations**: The generated visual explanations from each technique were compared. Visualizations included overlaying explanation maps on the original MRI images to illustrate the focus areas of the model.
10) **Comparison of Scores**: Comparison curves were plotted to visualize the performance of each explainable AI technique. This facilitated an assessment of which method provided the most reliable explanations [34].
11) **Final Comparison and Results Determination**: The results of the explainable AI techniques were compared and discussed in relation to the training and validation metrics, highlighting their strengths and weaknesses [33], [34]. The findings were synthesized to conclude the interpretability of the UNet model in the context of glioma segmentation.

In this study, we applied four Explainable AI (XAI) techniques to interpret the results of our UNet model for brain tumor segmentation. Each technique generates visual explanations to highlight regions within the MRI images that influenced the model's decisions. Grad-CAM [17], for instance, uses the gradient information from the model's final convolutional layer to create a heatmap that shows areas of high relevance. Saliency Maps [31] highlight pixels most influential to the model's predictions by computing the gradient of the model's output concerning the input. Vanilla Gradient and LRP [32] similarly provide relevance-based visual explanations, though they differ in how they assign weights to each pixel based on their contribution to the model's outcome.

To evaluate the effectiveness of each technique, we applied three metrics [33]: stability, fidelity, and unambiguity. Stability measures the consistency of explanations across similar input variations, assessing how stable the interpretation is in response to minor changes. Fidelity indicates how accurately the explanation aligns with the model's predictions, showing how well the highlighted regions represent the model's decision-making process. Unambiguity assesses the clarity of the explanation, ensuring that important regions are distinctly and reliably represented in the visualizations. Together, these metrics provide a comprehensive assessment of each XAI technique's reliability and interpretability for the model's segmentation task.

IV. UNet ARCHITECTURE AND EXPLAINABLE AI MODELS

A. UNet OVERVIEW AND USE

UNet has emerged as a powerful architecture for image segmentation, especially in biomedical imaging applications. Its specialized design enables accurate segmentation, making it highly applicable in fields such as medical image analysis, specifically for tasks like tumor detection. The reason for UNet's widespread use is its capacity to precisely segment data by capturing both local attributes and global context [21].

B. UNet MODEL ARCHITECTURE

The UNet architecture, as seen in Figure 2, features an encoder-decoder design with interconnected paths, tailored to handle unlabeled masks and generate precise segmentation results [21].

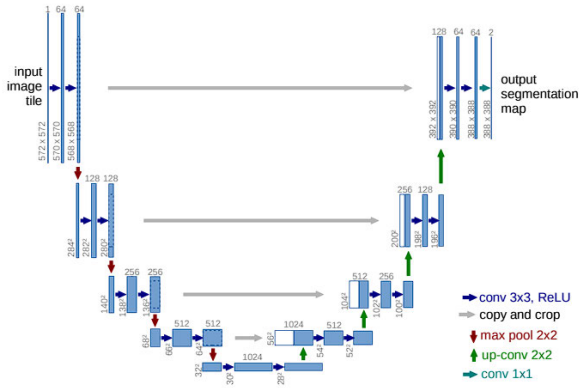

FIGURE 2. UNet architecture diagram [21].

1) **Encoding path:** The encoding path of UNet focuses on extracting high-level features and reducing spatial dimensions through a series of 3×3 convolutional layers with ReLU activation, along with max-pooling operations. This process diminishes spatial features while doubling the channels to compensate for spatial resolution reduction.

2) **Decoding path:** The decoding path involves 3×3 convolutional layers with ReLU activation and upsampling using 2×2 convolutional layers. This step enriches spatial features while halving the channels. Additionally, symmetrical features from the encoding path are concatenated to ensure accurate segmentation.

3) **Connecting paths:** The integration of low-level and high-level characteristics is made easier by connecting paths, which allow information flow across the encoding and decoding paths. Fusing spatial and semantic information ensures pixel-perfect segmentation results.

4) **Bottleneck:** The bottleneck, where the encoder transitions to the decoder, involves downsampling, convolution, and upsampling. During upsampling, segments are concatenated with encoding path channels to enhance clarity and distinctiveness for subsequent layers.

5) **Key functions in UNet:**

- 3×3 **Convolution:** Employed for effective feature extraction while maintaining computational efficiency.
- 2×2 **Maxpooling:** Used for downsampling, preserving essential spatial information crucial for segmentation tasks.
- **ReLU and Sigmoid Activation:** Introduce non-linearity and accelerate convergence during training.
- **Adam Optimizer:** Facilitates fast convergence and robustness to noisy gradients during training.
- **Cross Entropy Loss:** Penalizes prediction errors by measuring dissimilarity between predicted probabilities and ground truth labels.
- **Dice Coefficient Loss:** Emphasises spatial agreement for precise segmentation by measuring the overlap between the ground truth and anticipated masks.
- **Concatenation:** Establishes skip connections between encoder and decoder paths to preserve spatial information and propagate low-level details.
- **Transpose Convolution:** Upsamples feature maps to restore spatial resolution lost during downsampling, aiding in accurate segmentation.

C. BRATS 2020 (TRAINING AND VALIDATION) DATASET

The BraTS 2020 dataset [35], as seen in Figure 3, provides a comprehensive breakdown of data across the training, test, and validation sets, supporting effective model training and evaluation for glioma segmentation in pre-operative MRI images. It includes detailed annotations for various tumor sub-regions, such as necrotic and non-enhancing core (NCR/NET), peritumoral edema (ED), and enhancing tumor (ET), each uniquely labeled for precise segmentation. The dataset offers multiple imaging modalities, including T1-weighted, T2-weighted, and FLAIR sequences, which enhance the reliability of segmentation. Figure 4 illustrates a sample FLAIR image from the dataset, showcasing the four types anat, epi, img, and mask with a region of interest (ROI) highlighted to demonstrate the segmentation of critical areas [35], [36], [37], [38], [39].

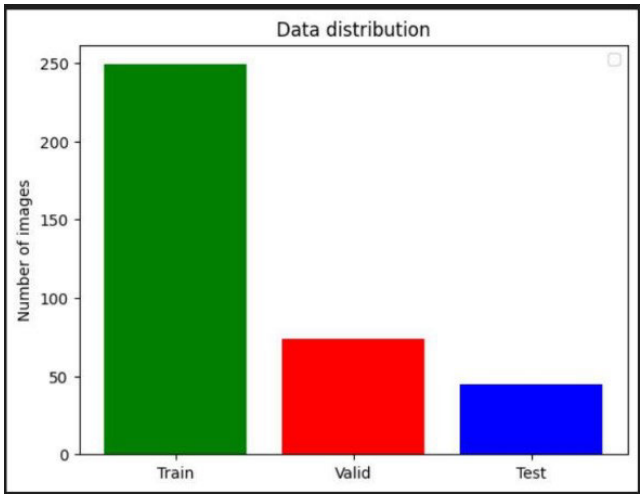

FIGURE 3. Data distribution of images across the datasets.

D. EXPLAINABLE AI MODELS USED

1) VANILLA GRADIENT

One of the most straightforward techniques in the explainable AI family is Vanilla Gradient [20]. The goal is to calculate the model's output gradient in relation to the input. This gradient shows the impact of input modifications on the prediction. It provides information about the pixels (or areas) in the input image that are most crucial for prediction. This gradient-based technique aids in producing a saliency map that identifies key areas for an image classification challenge.

a: MATHEMATICAL FORMULA

Let us assume a model $f(x)$ and an given input image x, the gradient at each pixel i is computed as:

$$G_i = \frac{\partial f(x)}{\partial x_i} \tag{1}$$

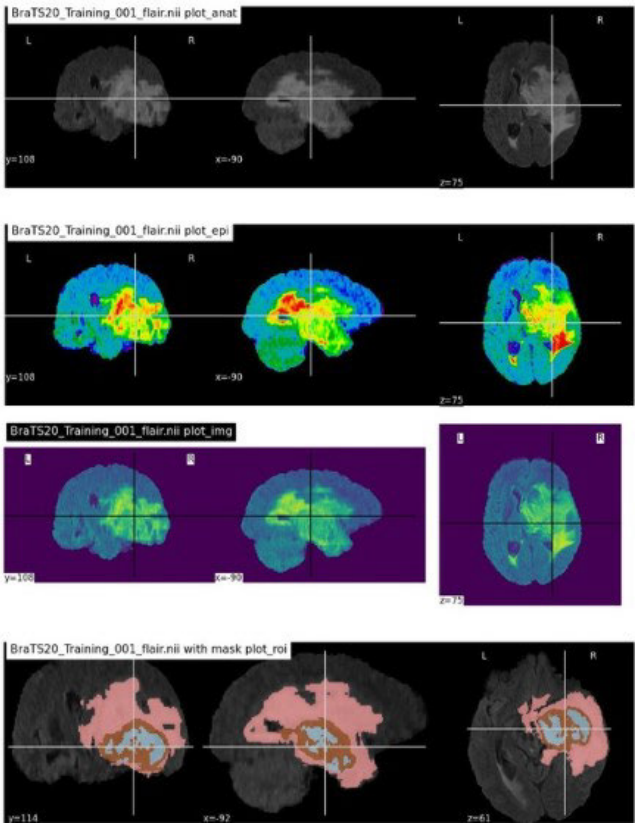

FIGURE 4. Representation of images with different highlighted regions.

G_i represents the gradient of the model's output with respect to the i-th pixel of the input image x. This gradient is often visualized as a heatmap, explaining the model's decision process.

In the research work [22], the authors introduced Vanilla Gradient as a simple method to generate class-specific saliency maps. Using this method, regions of the picture that are most important for classification are highlighted by taking the gradient of the class score with respect to the input image pixels. By selecting pixel areas with the most effect, the study illustrated the usefulness of this method in visualising the Convolutional Neural Networks' (CNNs') decision-making process and providing insight into how they arrive at particular classifications. This method laid the foundation for further improvements in understanding deep networks by visually interpreting their learned representations. The paper's primary contribution was making CNNs more interpretable and facilitating debugging in classification tasks [22].

Vanilla Gradient is a foundation for many gradient-based interpretability techniques. While it gives a rough estimate of how sensitive the model is to changes in each pixel, it can be noisy and lack clarity in regions of interest, especially in medical image segmentation tasks like brain MRI analysis. However, it is still useful as a quick, baseline visualization of importance in the input image.

2) GRAD-CAM

A more sophisticated method called Grad-CAM creates a heatmap of significant areas by using the gradients of a model's output in relation to its convolutional feature maps [23]. It focuses on the final convolutional layer, which captures spatial information and can thus localize important image regions related to a specific class.

Grad-CAM (Gradient-weighted Class Activation Mapping) was proposed by Selvaraju et al. [17], exploring the application of Grad-CAM as a way to produce visual explanations for decisions made by CNN-based models. This technique highlights the significant areas of an image that helped with the prediction by creating a coarse localisation map using the gradients that flow into the final convolutional layer. Without needing any architectural changes, Grad-CAM is adaptable and may be used for a variety of applications, including image categorisation, visual question answering (VQA), and picture captioning. The paper demonstrated the method's ability to generate faithful, class-discriminative visualizations that help users understand and trust the model's predictions [23].

a: MATHEMATICAL FORMULA

Let A^k represent the feature maps of the last convolutional layer, and y_c be the score for class c. The gradients of y_c with respect to the feature maps A^k are:

$$\frac{\partial y_c}{\partial A^k} \tag{2}$$

The importance weights α_k^c for each feature map are calculated by:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k} \tag{3}$$

Z is a normalization factor (the number of pixels). The Grad-CAM heatmap $L_{Grad-CAM}^c$ is then:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{4}$$

ReLU is used to focus on the regions that positively influence the class score.

Grad-CAM's capacity to highlight regions of interest makes it incredibly valuable for applications like brain MRI segmentation, which is why it is commonly employed in medical image analysis. It provides a more interpretable visualization by focusing on coarse regions and offering a balance between sensitivity and specificity.

3) LAYER-WISE RELEVANCE PROPAGATION (LRP)

Layer-wise Relevance Propagation (LRP) [24] is a technique that, depending on how much each neuron contributes to the final choice, backpropagates the relevance score from the output layer to the input layer. This score indicates which input elements were most important to the model's conclusion. Unlike gradients, LRP is based on the decomposition

of the prediction, and it respects the structure of the deep network layers [33].

a: MATHEMATICAL FORMULA

The relevance ratings are spread backwards through the network, beginning with the output. For a neuron j in one layer and i in the previous layer, the relevance R_j is calculated as:

$$R_j = \sum_i \frac{z_{ij}}{\sum_{j'} z_{ij'}} R_i \tag{5}$$

z_{ij} represents the activation from neuron j to i.

Layer-wise Relevance Propagation (LRP) was first introduced by Bach et al. [24]. LRP aims to decompose a classifier's prediction into the contributions of each input pixel, distributing the relevance backwards through the network layers. The key idea is to assign relevance scores to pixels that contributed most to the classification outcome, allowing for a deeper understanding of how neural networks process inputs. LRP has been instrumental in analyzing non-linear models and improving model interpretability across a variety of domains, including image classification and medical diagnostics.

LRP provides fine-grained relevance maps that help interpret a model's decisions at the level of individual neurons, making it highly useful in sensitive areas like medical imaging. For brain MRI segmentation, it can provide detailed relevance maps that explain how different regions of the brain contribute to a model's classification or segmentation output.

4) SALIENCY MAPS

Saliency maps are another gradient-based method [32] where the magnitude of the gradient of the output with respect to each pixel is used to indicate the "saliency" of that pixel. Unlike Vanilla Gradient, saliency maps focus on the absolute value of the gradient to better highlight important regions without considering the direction of the change.

The concept of Saliency Maps was expanded in the work by Zeiler and Fergus [25]. They introduced deconvolutional networks, which provide insight into the activations of intermediate layers in CNNs. By visualizing saliency maps, the authors were able to identify which parts of the input image were most influential in triggering certain activations. This work contributed significantly to understanding how CNNs process and filter image data at various stages, offering a more granular view of network decision processes.

a: MATHEMATICAL FORMULA

Given a model $f(x)$ and input x, the saliency map S_i is computed as:

$$S_i = \left| \frac{\partial f(x)}{\partial x_i} \right| \tag{6}$$

This method considers both positive and negative gradients.

Saliency maps are useful in providing insights into which areas of the input image influence the model's prediction the most. They can highlight key regions in brain MRI segmentation tasks, identifying areas that significantly contribute to identifying tumors or other abnormalities. However, saliency maps can sometimes be too diffuse, making them less interpretable compared to methods like Grad-CAM.

V. RESULTS, VISUALISATIONS AND ANALYSIS

A. LEARNING CURVE

Graphical representation of the model's learning curve over training epochs is shown in Figure 5, and it illustrates its convergence behavior and performance trends over training and validation datasets. It is to be noted that no patient contributed data used for testing is used in both the training and validation datasets.

The curve shows a steady increase in training and validation accuracy, along with a decrease in loss, indicating effective convergence over epochs. The dice coefficient and mean IoU metrics also demonstrate consistent improvements, with values closely tracking each other, showing good generalization on the validation set.

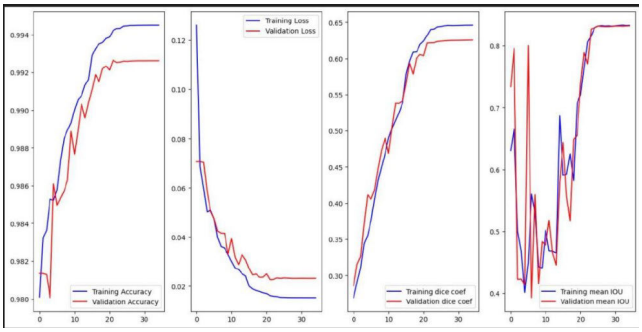

FIGURE 5. Learning curve for model training across four metrics.

B. FINAL PREDICTIONS

The final predictions from the UNet model on test dataset is obtained as seen in Figure 6.

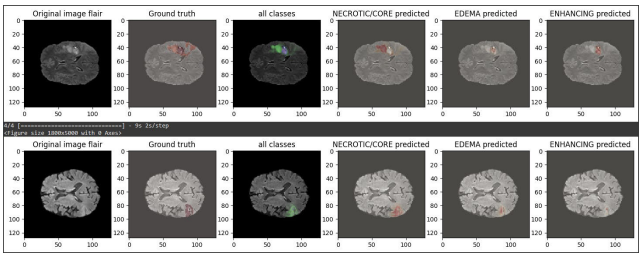

FIGURE 6. Classwise predictions from UNet.

C. METRICS

Table 3 shows the quantitative assessment of model performance using predefined metrics, providing a comprehensive evaluation of segmentation accuracy and reliability.

TABLE 3. Test metrics for the UNet brain tumor segmentation model.

Metric	Value
Test Loss	0.017342811450362206
Test Accuracy	0.9939939379692078
Mean IoU	0.828941822052002
Dice Coefficient	0.657872200012207
Precision	0.9942461848258972
Sensitivity (Recall)	0.9926532506942749
Specificity	0.9980550408363342
Dice-Coefficient of Necrotic	0.674229642593384
Dice-Coefficient of Edema	0.7835088968276978
Dice-Coefficient of Enhancing	0.708930492401123
F1 Score	0.9934490792176661

The UNet model's performance analysis indicated higher test accuracy which makes this model highly efficient in brain tumor segmentation.

The model demonstrated strong segmentation capability, achieving high specificity, indicating its reliability in distinguishing tumor regions from healthy tissue.

The high Dice scores underscore the model's effectiveness in segmenting distinct tumor subregions, while precision and recall values highlight its balanced performance in identifying true positives without overestimating segmentation regions.

D. EXPLAINABLE AI MODEL VISUALIZATIONS

1) GRAD-CAM

Grad-CAM visualisation offers a means of deciphering the areas of a picture that have a major impact on a model's prediction, as seen in Figure 7.

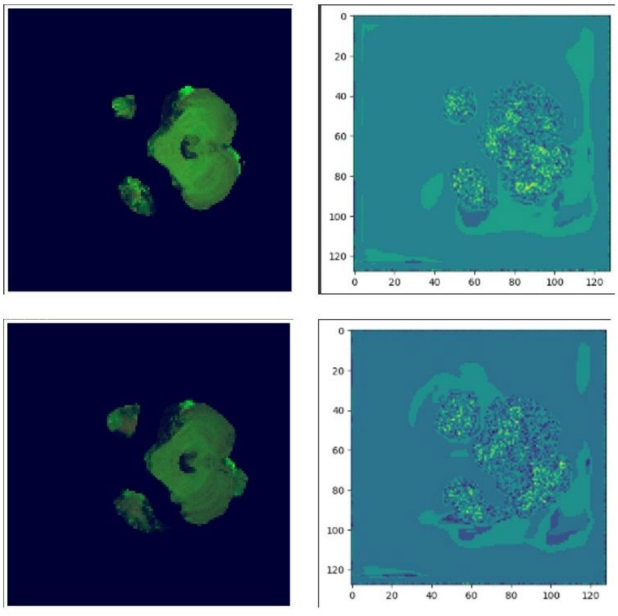

FIGURE 7. Grad-CAM visualization.

2) SALIENCY MAPS

Saliency maps illustrate the most influential pixels in an image for a given prediction as seen in Figure 8. By enhancing interpretability, saliency maps play a vital role in validating model outputs in sensitive domains such as healthcare.

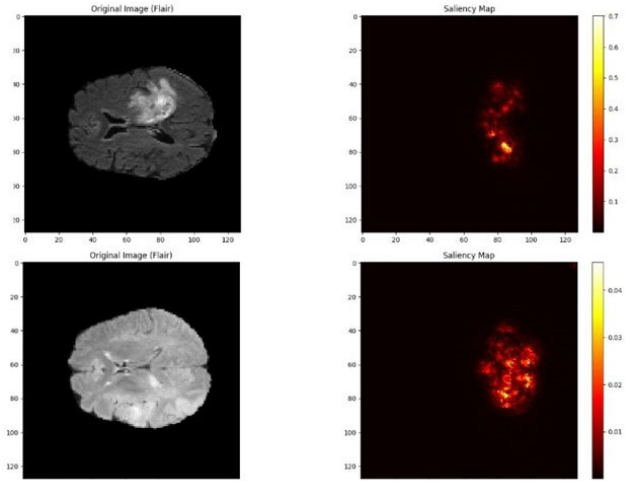

FIGURE 8. Saliency map visualization.

3) VANILLA GRAD

By visualising the output's gradient in relation to the input image, the Vanilla Grad technique gives a clear picture of how input changes impact the model's predictions. These kinds of insights are very helpful for evaluating the robustness of the model and can be seen in Figure 9.

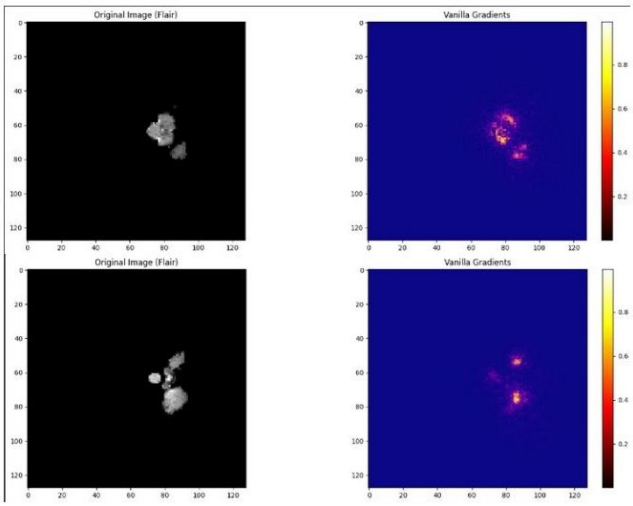

FIGURE 9. Vanilla grad visualization.

4) LRP (LAYER-WISE RELEVANCE PROPAGATION)

Layer-wise Relevance Propagation (LRP) is a powerful technique that traces the contributions of each input feature back through the layers of the network to determine their relevance to the final prediction. This method enhances transparency as seen in Figure 10 and is particularly useful

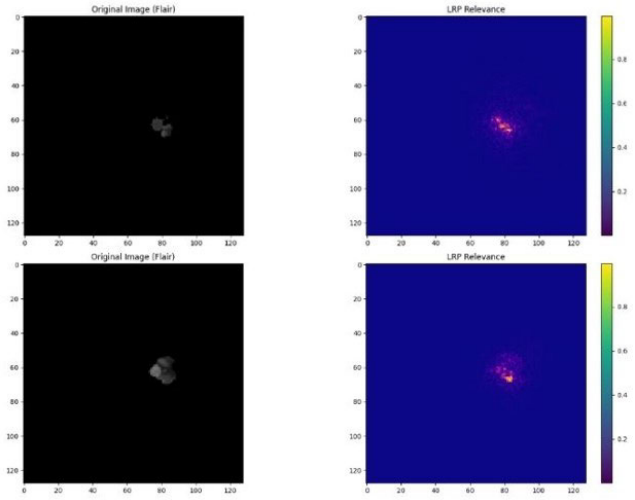

FIGURE 10. LRP visualization.

in domains where understanding model behavior is critical for trust and safety.

E. METRICULAR ANALYSIS OF EXPLAINABLE AI MODELS

The explainability of the four Explainable-AI models used for brain tumour segmentation is assessed in this section. Three important metrics–fidelity, unambiguity, and stability–are used to evaluate the models. The examination of these measures over five example scenarios reveals the advantages and disadvantages of each approach, and each one offers distinct insights into the interpretability of the models. Each statistic is explained in detail in the next subsections, along with a comparison of the models' performance and an explanation of their applicability.

The complexities of uncertainty in XAI and its effects on human perception of explanations are highlighted across several studies, which collectively emphasize the importance of fidelity, stability, and trust in AI systems. For instance, in a research work [26], it is discussed how modeling uncertainty is crucial for improving interpretability, as fidelity significantly impacts user trust and understanding. Another study [27], introduced the LEAF framework, underscoring how fidelity metrics influence user trust in local linear XAI methods. Furthermore, a research [28], revealed how different similarity measures lead to variations in the stability of local surrogate models, thus affecting the trustworthiness of explanations. The choice of exemplars is also critical, as explored in [29], where findings indicate that exemplar selection can impact explanation ambiguity and user confidence. Finally, the research work [30] argued for a multi-metric framework, asserting that relying on a single metric may lead to an incomplete evaluation of explainability. Together, these studies call for better modeling of uncertainty, careful selection of metrics, and more standardized evaluation practices in XAI.

1) FIDELITY

The degree to which an Explainable AI model's explanations faithfully match the underlying model's predictions is known as fidelity. By ensuring that the explanations accurately reflect the AI system's core decision-making process, it boosts user confidence and promotes better openness.

The fidelity scores as obtained in Table 4 indicate that Grad-CAM generally outperformed Saliency Maps, Vanilla Grad, and LRP across cases, showing higher alignment with the UNet model's predictions. However, the performance of all methods was relatively close, suggesting consistency in interpretability but with Grad-CAM showing a slight advantage overall, with proper visualization appearing in Figure 11.

TABLE 4. Fidelity scores in 10^{-2} format.

Case	Grad-CAM	Saliency Maps	Vanilla Grad	LRP
Case 1	0.11	0.04	0.05	0.05
Case 2	0.31	0.25	0.26	0.26
Case 3	0.14	0.10	0.12	0.12
Case 4	0.37	0.37	0.28	0.28
Case 5	0.05	0.05	0.04	0.04

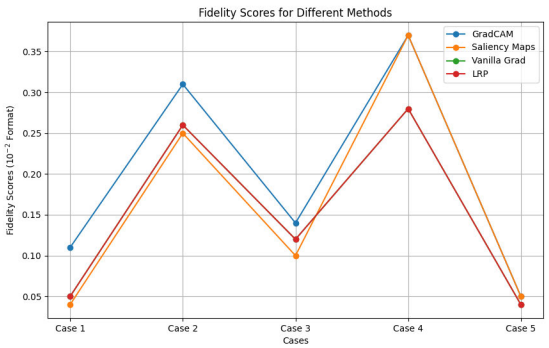

FIGURE 11. Fidelity scores comparison.

2) UNAMBIGUITY

Unambiguity refers to the clarity and specificity of the explanations generated by the Explainable AI model. It measures whether the explanations are non-redundant and effectively highlight only the essential features that significantly contribute to the model's predictions, ensuring that the information provided is clear and focused.

The unambiguity values tabulated in Table 5 show a notable distinction between Grad-CAM and Saliency Maps, which yield consistently low values, and Vanilla Grad and LRP, which produce much higher values, especially in Cases 3, 4, and 5. This suggests that Vanilla Grad and LRP provide more concentrated and specific explanations, while Grad-CAM and Saliency Maps tend toward broader, less specific interpretations and that can be seen in Figure 12.

TABLE 5. Unambiguity values.

Case	Grad-CAM	Saliency Maps	Vanilla Grad	LRP
Case 1	0.297	0.029	0.004	0.004
Case 2	0.419	0.029	0.854	0.854
Case 3	0.358	0.029	14.743	14.743
Case 4	0.400	0.029	28.082	28.082
Case 5	0.485	0.029	21.642	21.642

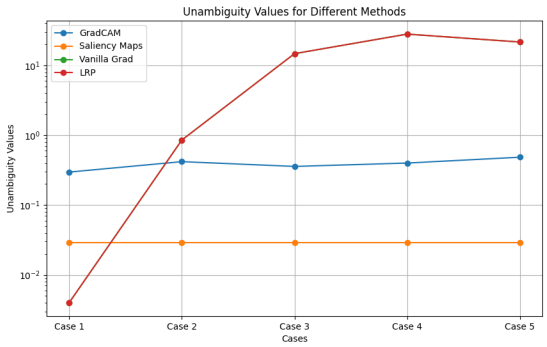

FIGURE 12. Unambiguity scores comparison.

3) STABILITY

Stability assesses the consistency of the explanations generated by the model under minor input perturbations or changes in the model architecture. A stable Explainable AI model will produce similar explanations when small alterations are made to the input, indicating that the explanations are not overly sensitive to random noise or variations in the input data.

The stability values are consolidated in Table 6, and they are consistently high across all methods, with Grad-CAM, Vanilla Grad, and LRP showing nearly identical stability, indicating reliable interpretability across cases. Saliency Maps exhibit slightly lower stability, suggesting minor variability, but still maintain comparable stability to the other methods. As seen in Figure 13, this consistency in stability across Grad-CAM, Vanilla Grad, and LRP highlights their robustness in producing dependable visual explanations, whereas the slight drop in Saliency Maps may imply a trade-off between detail and stability in their representations.

TABLE 6. Stability values in 10^{-4} format.

Case	Grad-CAM	Saliency Maps	Vanilla Grad	LRP
Case 1	0.61035	0.60788	0.61035	0.61035
Case 2	0.61035	0.60104	0.61035	0.61035
Case 3	0.61035	0.60446	0.61035	0.61035
Case 4	0.61035	0.60446	0.61035	0.61035
Case 5	0.61035	0.60446	0.61035	0.61035

Based on the analysis of unambiguity, stability, and fidelity scores across the four explainable AI models Grad-CAM, Saliency Maps, Vanilla Grad, and LRP, we can draw insightful conclusions regarding their performance.

Overall, the Vanilla Grad model stands out with its high unambiguity and consistent fidelity scores, making it a reliable choice for providing clear and interpretable

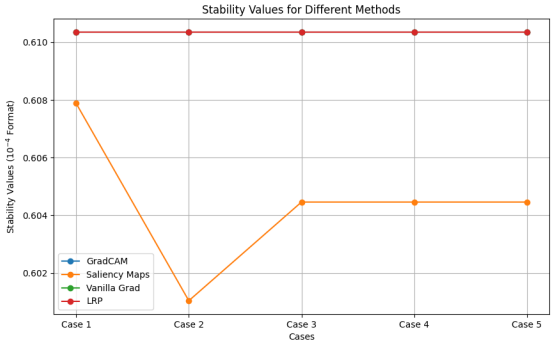

FIGURE 13. Stability scores comparison.

explanations in complex scenarios. The findings suggest that while LRP offers solid stability, the combination of high fidelity and clarity from the Vanilla Grad model makes it the preferred method for enhancing the interpretability of AI systems in brain tumor segmentation. These insights underscore the importance of selecting models that not only perform well on technical metrics but also enhance user trust and understanding in critical applications.

VI. PERFORMANCE ANALYSIS

In this study, the UNet architecture was employed for brain MRI segmentation. A total of 369 subjects were included, out of which approximately 295 subjects were utilized for training, while around 37 subjects were allocated for both validation and testing, respectively. The training set comprised approximately 45,725 image slices (295 × 155). Both the validation and testing sets consisted of 5,735 slices each. The UNet model implemented in this work consists of a five-layer deep architecture with 3 × 3 convolutional kernels at each layer. The experimentation is done with the experimental set-up of Colab GPU 15GB with a memory size of 112GB (100GB used), with a system configuration of RAM 16GB and hard disk space of 1TB. The computational time required for each phase of the model pipeline, such as training, validation, and testing, is recorded; these details are presented in Table 7.

TABLE 7. Estimated running time with standard UNet.

Phase	# Images	Time per Epoch	Total Time (36 Epochs)
Training	∼45,725	6–8 minutes	3.5–4.8 hours
Validation	∼5,735	30–60 seconds	∼30 minutes
Testing	∼5,735	5–10 minutes	One-time inference

A rigorous analysis is conducted to get the time and space complexity of the implemented UNet architecture by utilizing the underlying working mechanism of it [21] and the inferences given in [40]. For computational purposes, the following parameter naming conventions are used: N = (Height X Width) be the input spatial size; D = Number

of levels (depth); C = Number of base channels; and K = Number of Kernels. This is depicted in Table 8.

TABLE 8. Time and space complexity (UNet).

Metric	Value
Time Complexity	$\mathcal{O}(N \cdot D \cdot K^2 \cdot C^2)$
Where: N = image size, C = channels, K = Kernels and D = depth	
Space Complexity	
Training	$\mathcal{O}(D \cdot K^2 \cdot C^2 + D \cdot C \cdot N)$;
Inference	$\mathcal{O}(D \cdot K^2 \cdot C^2 + C \cdot N)$
Model Params	∼31M (UNet, 5 levels)
RAM Usage	8–10 GB (Batch size = 32, 512×512 slices)
Memory Bottleneck	Intermediate feature maps for backpropagation

Based on the above analysis, it is inferred that UNet is superior to other deep neural network models such as VGG, ResNet etc., as it provides higher accuracy with few parameters. In addition, its space complexity increases only linearly and not quadratically with depth, and does not require a massive width and height of the channels. During inference, the UNet architecture requires less activation memory as it reuses the buffer. Hence, it is highly recommended for low-resource applications.

As UNet has skip connections between the encoders and decoders at the same level, the decoders are allowed to access high-resolution feature maps, which reduces the time taken. This also increases boundary accuracy and improves localization, which is highly essential for image segmentation use cases. Hence, the UNet architecture is superior to other similar counterparts, especially for medical image segmentation.

Further, the time taken by different XAI methods to generate visual explanations was computed. The average running time per image for each method is presented in Table 9.

TABLE 9. Average computation time per image for different XAI methods.

Method	Avg. Time/Image	Notes
Vanilla Grad	∼20–30 ms	Single backpropagation pass
Saliency Map	∼30–40 ms	Same as Vanilla Grad but visualized differently
Grad-CAM	∼60–90 ms	Backprop to specific layer and Grad × Feature Map
LRP	∼120–200 ms	Rule-based backward pass through the full network

The theoretical time complexity of the XAI methods used in this work is summarized in Table 10. The complexity is measured with respect to the number of forward (F), backwards (B), additional processing overhead (A) and rule-based (R) operations.

Apart from that, the space complexity of each XAI method was also analyzed, as shown in Table 11. The space requirements are expressed in terms of - input size (n), space for intermediate activations (A), number of channels (f) and dimensions of the layer selected (h,w).

TABLE 10. Theoretical time complexity of different XAI methods.

Method	Time Complexity	Explanation
Vanilla Grad	$O(B)$	Gradient will be calculated during the single backward pass
Saliency Map	$O(B)$	Same as Vanilla Grad with standard backward pass time complexity of B
Grad-CAM	$O(F + B + A)$	It computes gradient with respect to feature maps in each convolutional layer and uses weighted sum of all these values hence, additional cost A is added
LRP	$O(F + B_R)$	It does the forward pass once and then it performs a custom backward pass to redistribute the score using layer-wise rules R.

TABLE 11. Space complexity of different XAI methods.

Method	Space Complexity	Notes
Vanilla Grad	$O(n + A)$	Need to store each input size and activations in intermediate layers
Saliency Map	$O(n + A)$	Same as above , only one backward pass will be carried out and the gradient value will be stored
Grad-CAM	$O(n + A + f \cdot h \cdot w)$	Stores input and input gradient, activations and feature maps of target convolutional layer
LRP	$O(n + 2A)$	Backward relevance scores and intermediate activation are stored

VII. CONCLUSION

This research has demonstrated the vital role of explainable AI models in enhancing the interpretability of deep learning systems for brain tumor segmentation. Through a comprehensive evaluation of unambiguity, stability, and fidelity metrics across Grad-CAM, Saliency Maps, Vanilla Grad, and LRP models, we find that the Vanilla Grad model provides the most reliable explanations with superior clarity and consistent performance. These findings emphasize that model selection requires not just accuracy but also trust and understanding among medical practitioners. Implementing explainable AI techniques will be crucial for advancing diagnostic processes through AI and ensuring safe and effective patient care. As future work, hybrid UNet models with XAI can be experimented to find the best model for brain tumor identification using image segmentation.

ACKNOWLEDGMENT

The authors are grateful to DST-FIST and VIT management for the resources provided for this work. In addition, they thank Deakin University, Australia, for their support in the successful completion of this research work.

REFERENCES

[1] F. Ullah, H. A. Khan, S. T. Hussain, and M. J. M. I. Manzoor, "Hybrid deep learning model for brain tumor classification," *IEEE Access*, vol. 8, pp. 225628–225640, 2020, doi: 10.1109/ACCESS.2023.3330919.

[2] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghoushchi, S. Anari, M. Naseri, and M. Bendechache, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," *Sci. Rep.*, vol. 11, no. 1, p. 10930, May 2021, doi: 10.1038/s41598-021-90428-8.

[3] K. Ejaz, R. Muhammad, and M. A. Khan, "Hybrid fuzzy K mean-self organization mapping approach for brain tumor segmentation," *Appl. Soft Comput.*, vol. 93, Apr. 2020, Art. no. 106332, doi: 10.1016/j.asoc.2020.106332.

[4] A. Işın and M. A. Acar, "MRI-based brain tumor segmentation using deep learning techniques," *J. Biomed. Informat.*, vol. 96, Feb. 2019, Art. no. 103225, doi: 10.1016/j.jbi.2019.103225.

[5] W. Sallam and A. F. Seddik, "A review on brain MRI image segmentation," in *Proc. 2nd Int. Conf. New Paradigms Electronics Information Technologies (PEIT)*, Luxor, Egypt, Nov./Dec. 2013.

[6] S. C. Liu and T. Y. Xu, "Research on brain tumor segmentation and identification technology based on improved unet," in *Proc. 36th Chin. Control Decis. Conf. (CCDC)*, May 2024, pp. 5161–5168.

[7] S. Xavier, P. S. K. Sathish, and G. Raju, "Advancing brain tumor segmentation in MRI scans: Hybrid attention-residual UNET with transformer blocks," *Int. J. Online Biomed. Eng.*, vol. 20, no. 6, pp. 103–115, Apr. 2024, doi: 10.3991/ijoe.v20i06.46979.

[8] K. Tong, J. Ding, and X. Li, "Improved MRI brain tumor segmentation method based on UNet++," *Academic J. Sci. Technol.*, vol. 10, no. 1, pp. 250–254, Mar. 2024, doi: 10.54097/pm55y427.

[9] L. Huang, E. Zhu, L. Chen, Z. Wang, S. Chai, and B. Zhang, "A transformer-based generative adversarial network for brain tumor segmentation," *Frontiers Neuroscience*, vol. 16, 2022, Art. no. 1054948, doi: 10.3389/fnins.2022.1054948.

[10] E. K. Kumar, A. Ajay, K. H. Vardhini, R. Vemu, and A. A. Padmanabham, "Residual edge attention in U-net for brain tumour segmentation," *IJRITCC*, vol. 11, no. 4, pp. 324–340, 2023, doi: 10.17762/ijritcc.v11i4.6457.

[11] S. Saranya and R. Subhashini, "A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends," *Decis. Anal. J.*, vol. 7, Jun. 2023, Art. no. 100230, doi: 10.1016/j.dajour.2023.100230.

[12] M. Nagahisarchoghaei, N. Nur, L. Cummins, N. Nur, M. M. Karimi, S. Nandanwar, S. Bhattacharyya, and S. Rahimi, "An empirical survey on explainable AI technologies: Recent trends, use-cases, and categories from technical and application perspectives," *Electronics*, vol. 12, no. 5, p. 1092, Feb. 2023, doi: 10.3390/electronics12051092.

[13] J. Brasse, H. R. Broder, M. Förster, M. Klier, and I. Sigler, "Explainable artificial intelligence in information systems: A review of the status quo and future research directions," *Electron. Markets*, vol. 33, no. 1, pp. 26–43, Dec. 2023, doi: 10.1007/s12525-023-00644-5.

[14] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham, Switzerland: Springer, 2019, ch. 51.

[15] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," 2019, *arXiv:1903.03894*.

[16] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 3681–3688.

[17] 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," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[18] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Er, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 2, p. e1372, 2021, doi: 10.1002/widm.1372.

[19] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.*, vol. 2, no. 6, pp. 305–311, Jun. 2020, doi: 10.1038/s42256-020-0186-1.

[20] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imag.*, vol. 6, no. 6, p. 52, Jun. 2020, doi: 10.3390/jimaging6060052.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[22] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[23] 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 *Proc. ICCV*, 2017, pp. 618–626.

[24] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140, doi: 10.1371/journal.pone.0130140.

[25] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.

[26] T. Chiaburu, F. Haußer, and F. Bießmann, "Uncertainty in XAI: Human perception and modeling approaches," *Mach. Learn. Knowl. Extraction*, vol. 6, no. 2, pp. 1170–1192, May 2024, doi: 10.3390/make6020055.

[27] E. Amparore, A. Perotti, and P. Bajardi, "To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods," *PeerJ Comput. Sci.*, vol. 7, p. e479, Apr. 2021, doi: 10.7717/peerj-cs.479.

[28] C. Burger, C. Walter, and T. Le, "The effect of similarity measures on accurate stability estimates for local surrogate models in text-based explainable AI," 2024, *arXiv:2406.15839*.

[29] T. Leemann, Y. Rong, T. T. Nguyen, E. Kasneci, and G. Kasneci, "Caution to the exemplars: On the intriguing effects of example choice on human trust in xAI," in *Proc. NeurIPS*, 2023, pp. 1–8, doi: 10.1007/s41539-025-00301-w.

[30] M. Pawlicki, A. Pawlicka, F. Uccello, S. Szelest, S. D'Antonio, R. Kozik, and M. Choraś, "Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination," *Neurocomputing*, vol. 602, Oct. 2024, Art. no. 128282, doi: 10.1016/j.neucom.2024.128282.

[31] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," 2018, *arXiv:1810.03292*.

[32] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Lecture Notes in Computer Science), vol. 11700, W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and K. R. Müller, Eds., Cham, Switzerland: Springer, 2019, pp. 193–209, doi: 10.1007/978-3-030-28954-6_10.

[33] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," 2018, *arXiv:1812.04608*.

[34] V. Singh, K. Cyras, and R. Inam, "Explainability metrics and properties for counterfactual explanation methods," in *Explainable and Transparent AI and Multi-Agent Systems* (Lecture Notes in Computer Science), vol. 13283, D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling, Eds., Cham, Switzerland: Springer, 2022, doi: 10.1007/978-3-031-15565-9_10.

[35] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[36] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, no. 1, Sep. 2017, Art. no. 170117, doi: 10.1038/sdata.2017.117.

[37] 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, *arXiv:1811.02629*.

[38] S. Bakas, A. Hamed, S. Aristeidis, B. Michel, R. Martin, K. Justin, and D. Christos, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection [data set]," National Cancer Institute, Maryland, Cancer Imaging Archive, 2017, doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q.

[39] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, and C. Davatzikos, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection [data set]," National Cancer Institute, Maryland, Cancer Imaging Archive, 2017, doi: 10.7937/K9/TCIA.2017.GJQ7R0EF.

[40] C. Yao, W. Liu, W. Tang, J. Guo, S. Hu, Y. Lu, and W. Jiang, "Evaluating and analyzing the energy efficiency of CNN inference on high-performance GPU," *Concurrency Comput., Pract. Exp.*, vol. 33, no. 6, p. e6064, Mar. 2021, doi: 10.1002/cpe.6064.

D. JEYA MALA (Member, IEEE) is currently a Professor with the School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, Tamil Nadu, India. She is a part of the National Work Group on Quantum Mission, Government of India, and an Expert Evaluation Committee Member for AICTE-NEAT cell. She has more than 20 years of teaching and research and four years of industrial experience. She has been granted with a design patent and published two utility patents from IP, Government of India. She has published more than 60 papers in reputed, refereed, SCI and Scopus-indexed journals, conferences, and book chapters. She has completed several funded projects and published more than four books and MOOC courses for Udemy. She has formed the reviewer board of several international journals and conferences. Her research interests include artificial intelligence, explainable AI, deep learning, machine learning, software engineering, healthcare analytics, cyber security, and quantum computing. She is a member of editorial boards and technical program committees of several reputed journals and conferences. She is a member of ACM, Indian Science Congress Association (DST, Government of India), the Computer Society of India, and i-Soft, and an Invited Member of Machine Intelligence Research Laboratories.

MAINAK CHATTOPADHYAY is currently pursuing the B.Tech. degree in computer science engineering, specializing in artificial intelligence and machine learning with Vellore Institute of Technology, Chennai. He has cracked offers from Infosys, Cognizant, and Tata Consultancy Services. He has interned with Teach For India, managing mobile devices and working on Salesforce Cloud, and with HMT Architects, where he developed the company's website. His projects include PAW-sitive (an animal healthcare portal), Chaters (a full-stack chat app), and Breast Tumor Segmentation using deep learning models. He has held leadership roles, including the Strategy and Operations Lead of the Tech Researchers Club, VIT Chennai, and a Technical Content Writer of the IEEE Computer Society. Additionally, he led the winning team at Hack4Bengal 3.0. His research focuses on artificial intelligence, machine learning, and full-stack web development.

PARTHIBA MUKHOPADHYAY is currently pursuing the B.Tech. degree in computer science engineering with a specialization in artificial intelligence and machine learning with Vellore Institute of Technology, Chennai. He is an Upcoming Intern with JPMorgan Chase and Company. He has previously interned as a Full-Stack Developer at Alpha Code Laboratories, leading the development of a finance microservice, and as a Web Developer at HMT Architects, where he designed and optimized the company's website. His notable projects include PAW-sitive (an animal healthcare portal), Crypto Nexus (a cryptocurrency tracking systems), and Brain MRI Segmentation using deep learning models. He has held leadership roles, including the Project Manager of the Pehchaan The Street School and the CP Representative at Google Developers Student Club. His research interests include AI, machine learning, web development, and healthcare technology. He was part of the winning team at Hack4Bengal 3.0 and a finalist at JPMC Code For Good 2024.

ROOPAK SINHA (Senior Member, IEEE) is currently an Internationally Recognized Expert in systematically designing safe and secure industrial software, with interests in requirements engineering, design and architectures, code generation, formal methods, research commercialization, and industrial standards. He has previously held research positions at Auckland University of Technology, The University of Auckland, New Zealand, and INRIA.