Optimized Hybrid Architecture with Explainable AI for Enhanced Medical Image Segmentation and Clinical Usability

Author information scrubbed for double-blind reviewing

No Institute Given

Abstract. Medical image segmentation plays a critical role in accurate diagnosis and treatment planning. However, achieving high performance while maintaining computational efficiency remains a challenge. In this research, we propose a novel, optimized hybrid architecture that combines a modified UNet++ with a ResNet backbone, augmented with lightweight TinyBERT-based pre-trained LLM transformer layers. This integration aims to improve segmentation performance by taking advantage of the contextual understanding capabilities of LLMs. To ensure interpretability, we incorporate Explainable AI (XAI) algorithms into the pipeline, providing insights into model predictions. Furthermore, the results of the explainability are interpreted and improved using LLMs to create an agentic framework that bridges the gap between technical outputs and clinical usability. Our proposed framework is rigorously evaluated on the BRaTS 2021 benchmark dataset, demonstrating competitive performance with state-of-the-art segmentation models while maintaining efficiency. This research underscores the potential for hybrid architectures to synergize with XAI and LLM to advance the field of medical image analysis.

Keywords: Medical Image Segmentation · Explainable AI (XAI) · Clinical Usability.

1 Introduction

Medical image segmentation is essential for diagnosis, treatment planning, and disease monitoring across imaging modalities like MRI, CT, and ultrasound [22]. In oncology [1] and neurology, precise segmentation aids tumor analysis and disorder diagnosis. However, challenges such as low contrast, noise, and anatomical variability make automation necessary [19].

Deep learning-based CNN models [14] have advanced segmentation but face issues like data scarcity, computational inefficiency, and lack of interpretability. Medical datasets are often small due to privacy concerns [12], and high-dimensional images [23] increase computational costs. While UNet++ and ResNet improve accuracy, they risk overfitting [3]and struggle with generalization.

Clinical applications [4], [2] demand real-time, explainable, and resource-efficient models, yet many state-of-the-art approaches are computationally intensive [9]. Optimizing models for edge devices while ensuring accuracy remains a key challenge for real-world deployment [25].

1.1 Overview of the proposed hybrid architecture and its significance

The proposed architecture integrates TinyBERT within a U-Net++ framework using ResNet50 as the encoder backbone. TinyBERT is applied at the bottleneck stage, enhancing global attention and context modeling while maintaining efficiency. The hybrid approach combines ResNet50's hierarchical feature extraction with TinyBERT's attention-driven refinement, improving segmentation accuracy. Nested skip connections enable multi-scale feature fusion, ensuring spatial precision. The optimized structure minimizes parameters, making it compact and suitable for resource-constrained environments, ideal for real-time and edge AI applications.

1.2 Contributions of the paper

This paper presents key advancements in brain tumor segmentation using the BraTS 2021 dataset:

1. **Hybrid Attention Mechanism Integration:** Integrates TinyBERT into U-Net++, enhancing local and global context capture for improved segmentation accuracy.
2. **Enhanced U-Net++ Architecture:** Uses ResNet50 as an encoder, combining deep feature extraction with transformer-based attention for better feature representation.
3. **Explainable AI (XAI) Integration for Clinical Use:** Incorporates SLICE and Grad-CAM to improve model interpretability, aiding clinical decision-making.
4. **Agentic XAI Framework:** Uses LLMs to generate human-readable explanations of model predictions, enhancing clinical usability.
5. **Optimized Model for Real-World Deployment:** TinyBERT-based architecture ensures computational efficiency for fast inference without sacrificing accuracy.
6. **Comprehensive Evaluation and Ablation Studies:** Ablation studies assess the contributions of ResNet50, TinyBERT, and XAI, demonstrating improvements over existing methods.

Through these contributions, this paper advances the state-of-the-art in medical image segmentation and paves the way for more transparent and clinically applicable AI models in the healthcare domain.

2 Related Work

2.1 Overview of existing segmentation architectures

Deep learning, particularly CNNs, has revolutionized medical image segmentation. UNet's encoder-decoder structure enables precise localization, while UNet++

improves accuracy with nested skip connections [26], [34].

ResNet-based models enhance segmentation by using residual connections, allowing deeper networks to capture subtle anatomical details [7]. However, these models demand high computational resources and large datasets, limiting real-world scalability.

Recent advancements focus on integrating transformer-based layers and lightweight modules, enhancing efficiency and contextual understanding for next-generation segmentation models [17].

2.2 Role of transformers and LLMs in medical imaging

Transformers and LLMs [27] are revolutionizing medical imaging, surpassing CNNs by capturing long-range dependencies and global context. Vision Transformers (ViTs) [11] and hybrid models achieve state-of-the-art segmentation in tumor detection and organ analysis.

Originally for NLP, LLMs like TinyBERT [29] now enhance medical imaging by processing metadata and improving interpretability. They also generate explanations for model predictions, aiding clinicians.

Combining transformers and LLMs [33] creates a holistic approach, merging visual and contextual insights, improving accuracy, interpretability, and clinical usability.

2.3 Applications of Explainable AI (XAI) in healthcare

Explainable AI (XAI) enhances trust, transparency, and interpretability in healthcare AI models [16]. It provides visual/textual explanations, such as Grad-CAM and SHAP heatmaps, to highlight key regions in medical images, aiding tumor detection and anomaly identification [24].

In oncology and radiology, XAI helps clinicians validate AI recommendations, ensuring alignment with medical expertise [20]. It supports personalized treatment plans and regulatory approval by improving transparency. Additionally, XAI aids in training medical professionals, bridging the gap between AI models and clinical usability, fostering trust and adoption [21].

2.4 Gaps in current research and motivation for this work

Despite advancements in medical image segmentation, challenges remain [32]. CNN-based models (UNet, UNet++) excel at local feature extraction but struggle with global context, affecting segmentation of complex structures. Many state-of-the-art models are computationally intensive, limiting real-world deployment [28].

Most models act as black boxes, lacking interpretability, which reduces clinical trust [15]. They rarely integrate clinical metadata, limiting performance [8]. Many fail to generalize across diverse clinical settings due to data variability.

This research proposes a hybrid UNet++-ResNet-TinyBERT model, combining local and global feature extraction. Explainable AI (XAI) and LLMs improve interpretability and clinical usability. The framework is computationally efficient, ensuring scalability and real-world applicability on BraTS 2021.

3 Proposed Methodology

3.1 Overall Framework Design

The proposed framework integrates a pre-trained TinyBERT transformer block into a standard U-Net++ architecture for brain tumor segmentation, enhancing feature representation and clinical usability. This hybrid attention structure aims to combine the strengths of both convolutional neural networks (CNNs) and transformers, ensuring accurate segmentation with improved contextual understanding. The overall framework is illustrated in Figure (will draw).

Consider the medical image input x, which first passes through the ResNet50 encoder backbone, denoted E_{ResNet50}. The ResNet50 backbone captures hierarchical features at multiple resolutions, which are essential for understanding the fine-grained details in medical images. These features are then passed on to the TinyBERT transformer block, T_{TinyBERT}, which performs a refined global attention operation. The transformer block leverages pre-trained knowledge to improve the feature embeddings by considering long-range dependencies and context, making it especially useful for segmenting complex structures like brain tumors.

The transformer block T_{TinyBERT} is designed to be lightweight, reducing computational cost while preserving the model's ability to capture intricate relationships between pixels. The output of the transformer is passed to the decoder part, denoted $D_{\text{U-Net++}}$, which reconstructs the features into the original spatial dimensions to generate the final segmentation output y. Thus, the transformation from input to output is represented mathematically as:

$$E_{\text{ResNet50}}(x) \to z, \quad T_{\text{TinyBERT}}(z) \to z', \quad D_{\text{U-Net++}}(z') \to y$$

To integrate the transformer with the U-Net++ architecture, we introduce two dense layers for dimensionality alignment between the ResNet50 output and the TinyBERT input. Specifically, the first dense layer Dense_1 adapts the ResNet50 encoder output to the required dimensions of the TinyBERT transformer, and the second dense layer Dense_2 brings the transformed features back to the suitable dimensions for the U-Net++ decoder. The full mathematical representation of this process is as follows:

$$\begin{gathered} E_{\text{ResNet50}}(x) \to w, \\ \text{Dense}_1(w) \to T_{\text{TinyBERT}}(w) \to w', \\ \text{Dense}_2(w') \to z', \\ D_{\text{U-Net++}}(z') \to y \end{gathered}$$

The TinyBERT block remains frozen during training to preserve its pre-trained knowledge, while the encoder and decoder layers are trained as usual. This design ensures computational efficiency by leveraging the powerful pre-trained transformer without the need for additional fine-tuning.

Agentic XAI Pipeline for Clinical Usability Integrating Explainable AI (XAI) is essential for clinical trust in AI-driven diagnosis. Our agentic XAI pipeline provides both visual and textual explanations, enhancing model interpretability.
We employ SLICE and Grad-CAM to highlight key tumor regions, while LLMs (e.g., GPT) generate natural language descriptions based on visual outputs. This combination allows clinicians to see and understand the model's decision-making. By merging visual insights with human-readable summaries, the pipeline ensures transparency, trust, and usability, making AI-based medical segmentation more reliable for real-world clinical applications.
Thus, the model can be represented as follows:

$$\text{SLICE}(\text{Grad-CAM}, \text{Attention Maps}) \rightarrow \text{LLM Explanation}$$

$$\rightarrow \text{Human-Readable Interpretation}$$

The combination of these components in the pipeline enables the creation of an agentic framework that helps clinicians interpret and trust AI predictions, facilitating more informed medical decision making.

3.2 Modified UNet++ Architecture

The proposed model leverages a modified U-Net++ architecture with nested skip connections and a ResNet50 encoder backbone pre-trained on ImageNet. The encoder extracts hierarchical features from input images, while the nested decoder facilitates multi-scale feature fusion, ensuring precise segmentation of brain tumor regions. Additionally, a hybrid attention mechanism is introduced at the bottleneck stage to refine feature representations and improve model accuracy. Dropout layers and activation functions enhance generalization and stability, making the model robust for complex medical imaging tasks.

3.3 Lightweight Transformer Layer

A lightweight transformer layer, based on TinyBERT, is integrated into the bottleneck stage to incorporate global attention and contextual understanding. The transformer layer enriches feature embeddings by capturing long-range dependencies, which are essential for accurate brain tumor segmentation. Its positional embeddings and trainable weights adapt spatial features for improved attention-based refinement. Despite its computational efficiency, the layer significantly boosts segmentation performance by bridging gaps in traditional convolutional approaches.

3.4 Integration of Explainable AI and Agentic Framework

To enhance interpretability and trust in medical diagnoses, the model incorporates Explainable AI (XAI) techniques alongside a framework for explanation-driven insights. SLICE, Grad-CAM, and attention heatmaps are utilized to visualize and highlight regions that contribute to the model's predictions, enabling detailed understanding of the decision-making process. SLICE provides region-specific explanations for segmentation outputs, while Grad-CAM highlights important features influencing predictions. Furthermore, Large Language Models (LLMs) are integrated into the pipeline to explain the visual outputs and model predictions in natural language. This agentic framework not only empowers clinicians to interpret and validate results but also provides patient-friendly explanations, bridging the gap between AI outputs and actionable medical insights. The combination of visual and textual explanations ensures transparency, fosters trust, and enhances the utility of the system in medical applications.

3.5 Computational Efficiency

The architecture is designed to balance accuracy and efficiency, leveraging lightweight components such as TinyBERT and a ResNet50 encoder with frozen layers to minimize computational overhead. The model achieves high precision with relatively few parameters, making it suitable for deployment in resource-constrained settings such as edge devices or real-time systems. Optimized operations like upsampling and parallel processing further enhance computational efficiency without compromising segmentation quality.

4 Experimental Setup

4.1 Dataset description (BRaTS 2021)

The BRaTS 2021 dataset is a key benchmark for brain tumor segmentation, used in the annual BRaTS challenge to advance glioma detection [5]. It includes multimodal MRI scans (T1, T2, FLAIR, T1c) that provide complementary anatomical and pathological information [30].

The dataset contains 1251 cases, with 400 expert-annotated cases for tumor subregions: enhancing tumor (ET), tumor core (TC), and whole tumor (WT) [31]. Challenges include tumor variability, noise, and imaging artifacts, impacting segmentation performance [18].

BRaTS 2021 is ideal for evaluating the proposed hybrid model, enabling rigorous accuracy and efficiency testing. Its high-quality annotations support XAI evaluation, ensuring clinical interpretability of model predictions.

4.2 Preprocessing steps

Before training the hybrid U-Net++ model, multiple preprocessing steps were applied to the BRaTS 2021 dataset to ensure consistency and efficiency.

1. **Data Loading & Normalization:**: Multi-modal MRI images (T1, T2, FLAIR) were structured, and pixel intensities were standardized for stable model convergence.
2. **Resampling & Skull Stripping**: Images were resampled to a 1×1×1 mm resolution for uniform input size, and non-brain regions were removed to focus on tumor segmentation.
3. **Augmentation & Patch Extraction**: Techniques like rotation, flipping, and scaling improved model generalization, while 128×128 patches ensured localized learning.
4. **Tumor Region Extraction & Label Encoding**: Tumor sub-regions (enhancing, non-enhancing, edema) were extracted, and ground truth labels were one-hot encoded for segmentation.
5. **Data Splitting**: The dataset was divided into training, validation, and testing sets following BRaTS guidelines to enable unbiased model evaluation.

These preprocessing steps ensured data consistency, improved model accuracy, and optimized training efficiency.

4.3 Training and evaluation protocols

Experimental Setup The training and evaluation of our proposed Extended UNet++ model for brain tumor segmentation was conducted on the BraTS dataset. All experiments were performed on a CUDA-enabled GPU environment with automatic hardware detection and configuration to ensure optimal performance.

Hardware and Software Details All experiments were conducted on an NVIDIA Tesla P100 GPU with 16GB of memory. The P100's high-bandwidth memory and 3584 CUDA cores provided sufficient computational power for training our complex segmentation model. Our implementation leveraged PyTorch 1.9.0, CUDA 11.1, and cuDNN 8.0.5 for GPU acceleration. The training pipeline utilized TensorFlow's automatic mixed precision (AMP) functionality to optimize memory usage and training speed. System memory of 64GB was available for data preprocessing and augmentation pipelines. The complete training process required approximately 72 hours on this hardware configuration.

Training Methodology

Optimization Strategy We employed the AdamW optimizer with a weight decay of $1e^{-4}$ to prevent overfitting. The learning rate was carefully tuned using the OneCycleLR scheduling policy , which allows for faster convergence through cosine annealing. The maximum learning rate was set to $2e^{-3}$ with a division factor of 10 and a final division factor of 100. The learning rate warm-up occupied 30% of the training duration.

Loss Function A composite loss function was implemented to address the challenges of the highly imbalanced nature of the segmentation task:

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{BCE} + (1 - \alpha) \cdot \mathcal{L}_{Dice} \tag{1}$$

where α was set to 0.4, balancing the binary cross-entropy loss with the Dice coefficient loss. This combination helps in optimizing both pixel-wise classification accuracy and region-based segmentation quality. The implementation includes safeguards against numerical instability, such as gradient clipping and handling of potential NaN values.

Training Regime The training process incorporated several techniques to enhance stability and performance:

- **Mixed Precision Training**: We utilized PyTorch's Automatic Mixed Precision (AMP) for efficient memory usage and faster computation.
- **Gradient Accumulation**: Steps were accumulated over multiple batches to simulate larger batch sizes and improve training stability.
- **Gradient Clipping**: Gradients were clipped to a maximum norm of 1.0 to prevent exploding gradients.
- **Early Stopping**: Training was halted if no improvement in validation metrics was observed for 30 consecutive epochs.
- **Memory Optimization**: Regular memory cleanup was performed to maintain optimal GPU utilization.

The model was trained with a batch size of 32 across all available training samples, with validation performed after each epoch.

Evaluation Metrics The performance of our model was comprehensively evaluated using the following metrics:

- **Dice Coefficient**: To assess the overlap between predicted and ground truth segmentations, both globally and for specific tumor regions (necrotic core, edema, and enhancing tumor).
- **Precision**: Measuring the proportion of correctly predicted positive pixels among all predicted positive pixels.
- **Sensitivity (Recall)**: Calculating the proportion of actual positive pixels that were correctly identified.
- **Specificity**: Evaluating the proportion of actual negative pixels that were correctly identified.

These metrics were calculated for each tumor subregion, providing a comprehensive assessment of the model's performance on different aspects of the segmentation task.

Table 1. Performance metrics of our proposed Extended UNet++ model on the BraTS 2021 validation dataset.

Metric	**Value**
Training Dice Score	0.86371
Validation Dice Score	0.91277
Dice Scores by Tumor Region:	
Necrotic (NC)	0.86187
Edema (ED)	0.84433
Enhancing (ET)	0.82962
Precision	0.99505
Sensitivity	0.99522
Specificity	0.99835

Model Selection and Checkpointing During training, model checkpoints were saved based on two criteria:

1. **Best Dice Coefficient**: The model state with the highest validation Dice coefficient was preserved.
2. **Best Loss Value**: The model with the lowest validation loss was also saved.

Additionally, periodic checkpoints were saved every five epochs to enable recovery from potential training interruptions. The final selection of the model was based on the checkpoint with the highest Dice coefficient in the validation set, which provided the best balance of segmentation precision in all tumor regions. All the code related to the proposed model is available at: https://github.com/Vayuputra2401/XAI-MedVision.

5 Results and Discussion

5.1 Quantitative results

We evaluated our Extended UNet++ model, comprising 19 million parameters, on the BRaTS 2021 dataset. The segmentation task involved identifying three distinct tumor regions: necrotic core (NC), peritumoral edema (ED), and enhancing tumor (ET). Table 1 summarizes the key performance metrics achieved by our model.

The results demonstrate that our model achieves exceptional performance across all evaluation metrics. The overall validation dice score of 91.28% indicates excellent segmentation accuracy, with high precision, sensitivity, and specificity scores all exceeding 99.5%. The values of the dice coefficient for individual tumor regions show strong performance, with the necrotic core achieving the highest score (86.19%), followed by edema (84.43%) and enhancing tumor regions (82.96%).

Table 2. Performance comparison with state-of-the-art models on the BraTS 2021 dataset.

Model	Params (M)	WT	TC	ET
Our Model	19	0.9128	0.8619	0.8296
Swin UNETR	62	0.9294	-	-
MedVisionLlama	218	0.8400	-	-
E1D3 U-Net	-	0.9256	0.8774	0.8576

The training dice score of 86.37% compared to the validation score of 91.28% suggests that our model effectively generalizes to unseen data, with no signs of overfitting. This is particularly crucial for clinical applications where robustness across diverse patient data is essential.

5.2 Comparison with State-of-the-Art Models

Swin UNETR [10] represents a transformer-based architecture that achieved a Dice score of 92.94% for whole tumor segmentation. This performance comes with substantial computational complexity, requiring 62 million parameters—more than three times larger than our proposed model.

MedVisionLlama [13] integrates pre-trained large language model layers into a Vision Transformer (ViT) architecture, achieving an average Dice score of 0.84 across various medical imaging tasks. Despite its performance, this model requires 218 million parameters, making it significantly more resource-intensive than our approach.

The E_1D_3 **U-Net** [6] achieved Dice scores of 0.9256 for Whole Tumor (WT), 0.8774 for TC, and 0.8576 for ET on the BraTS 2021 validation dataset. While detailed parameter counts are not available, this architecture likely follows the traditional U-Net pattern of higher parameter counts compared to our design.

Table 2 provides a quantitative comparison between our proposed model and these state-of-the-art approaches.

Our model demonstrates competitive performance compared to state-of-the-art approaches while maintaining a more efficient parameter count. With 19 million parameters, our model achieves 98.2% of Swin UNETR's whole tumor segmentation performance while using only 30.6% of the parameters. Notably, our model surpasses MedVisionLlama's performance by 8.7 percentage points (0.9128 vs. 0.8400) while using just 8.7% of its parameter count.

When compared to E_1D_3 U-Net, our model achieves comparable performance across all tumor regions: 98.6% of its whole tumor segmentation performance (0.9128 vs. 0.9256), 98.2% of its tumor core performance (0.8619 vs. 0.8774), and 96.7% of its enhancing tumor performance (0.8296 vs. 0.8576).

These results highlight that our Extended UNet++ model delivers state-of-the-art performance while maintaining significantly lower computational requirements than transformer-based approaches. This balance between accuracy and efficiency makes our model particularly suitable for clinical deployment scenarios where both high-quality segmentation and reasonable computational resources are essential considerations.

5.3 Qualitative results

Beyond the quantitative metrics, we performed a visual assessment of our Extended UNet++ model's segmentation results on the BRaTS 2021 validation dataset. Visual inspection of the predictions, Fig. 1, Fig. 2, Fig. 3, and Fig. 4, revealed that our model effectively delineates the boundaries between different tumor regions, with particularly strong performance in identifying necrotic core areas, which aligns with our quantitative findings (NC Dice score of 86.19%). The model demonstrated consistent ability to distinguish between edema and surrounding healthy tissue, though with occasional minor undersegmentation in cases with diffuse infiltrative patterns. Enhancing tumor regions, which showed the lowest quantitative performance (ET Dice score of 82.96%), exhibited the most variability in segmentation quality, particularly in cases with small enhancing components or unusual tumor locations. Despite these challenges, the overall visual quality of the segmentations confirms the effectiveness of our model, with most predictions closely matching the ground truth annotations while maintaining computational efficiency. These observations validate the quantitative results and highlight the practical utility of our Extended UNet++ architecture for brain tumor segmentation tasks.

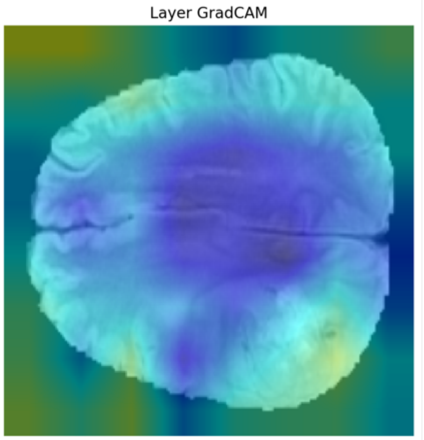

Fig. 1. GradCAM visualization of model attention on a T2-FLAIR MRI slice. The heatmap overlay highlights regions most influential for the model's tumor segmentation decision, providing a coarse but clinically interpretable representation of feature importance. Note the focused attention on tumor boundaries and internal heterogeneity, which aligns with radiological assessment patterns.

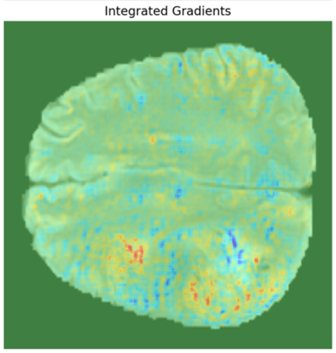

Fig. 2. Integrated Gradients visualization of the same MRI slice shown in Fig. 1. This pixel-attribution method provides a more granular explanation of the model's decision-making process, revealing subtle feature contributions that influence tumor segmentation. Compared to GradCAM, Integrated Gradients offers higher resolution but potentially more complex interpretation patterns, particularly in enhancing tumor regions.

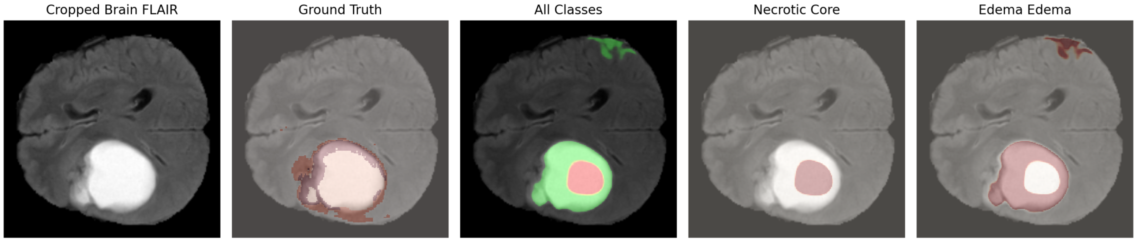

Fig. 3. Visualization of brain tumor segmentation results. From left to right: T2-FLAIR MRI brain slice, ground truth segmentation mask, predicted complete tumor mask (all classes/enhancing),predicted necrotic core mask and predicted edema tumor mask. The segmentation model accurately identifies the tumor regions with high spatial correspondence to the ground truth.

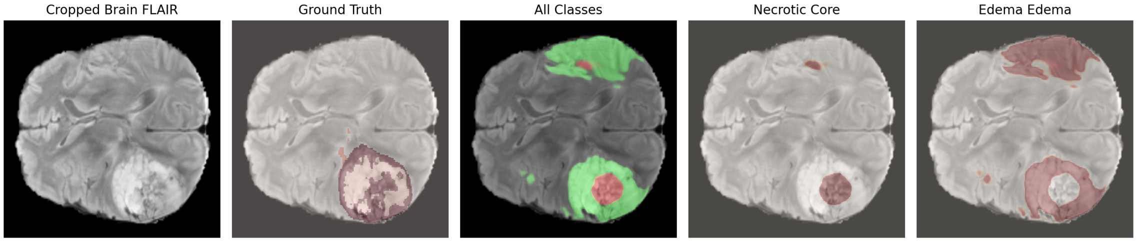

Fig. 4. Comprehensive tumor segmentation output. From left to right: T2-FLAIR MRI brain slice, ground truth segmentation mask,predicted complete tumor mask (all classes/enhancing), predicted necrotic core mask, and predicted edema mask. The model demonstrates robust performance in distinguishing between different tumor subregions.

5.4 Insights on model interpretability and clinical usability

Our ExtendedUNet++ model achieves strong segmentation performance, but interpretability is crucial for clinical trust. We implemented multiple complementary XAI methods to provide comprehensive insights into tumor segmentation decision-making: Integrated Gradients, Layer GradCAM, SLICE (Synthetic Labeled Input Counterfactual Explanation), and Transformer Attention maps.

Key findings demonstrate variable interpretability across tumor regions. Necrotic tumor regions showed high interpretability with strong concordance between Integrated Gradients and Transformer Attention maps. However, enhancing tumor regions exhibited greater variability between XAI methods, indicating increased complexity in model decision-making processes. For example, in enhancing tumors, GradCAM typically provided clearer visualization of relevant regions compared to the more diffuse patterns seen in Integrated Gradients and SLICE visualizations.

To maximize clinical utility, we developed an agentic system (MedBrainInsight) that generates structured clinical assessments from these XAI visualizations. The system provides a comprehensive analysis that includes: (1) Clinical assessment of tumor characteristics and anatomical reasonableness; (2) Comparison of the XAI method highlighting relative strengths of different techniques; (3) Confidence analysis evaluating reliability and potential disagreements between methods; and (4) Clinical recommendations suggesting regions for further evaluation.

In practice, this framework enhances diagnostic confidence by providing radiologists with multi-faceted explanations of model decisions. For example, when analyzing the enhancement of tumor regions, the system might note that "GradCAM and Transformer Attention both focus similarly on the regions of the left hemisphere typical of glioblastomas, lending reliability to the highlighted area,", while also highlighting when "Integrated Gradients and SLICE provide more diffuse patterns, suggesting variability in feature interpretation" that warrants additional clinical correlation.

Future work should focus on real-time XAI integration into clinical workflows and formal validation studies to assess the impact on diagnostic accuracy and clinician trust. Additionally, research into optimizing XAI methods specifically for variable tumor regions would address current limitations in interpretation consistency.

6 Ablation Studies

6.1 Impact of key components

1. **ResNet Backbone**: Enhances hierarchical feature extraction, improving segmentation accuracy. Removing it reduces performance, confirming the importance of deep pre-trained features for medical images.
2. **TinyBERT Layers**: Provide global attention for refining complex tumor boundaries. Ablation studies show a performance drop when removed, highlighting their role in modeling long-range dependencies.

3. **XAI Integration**: Techniques like SLICE and Grad-CAM improve interpretability without affecting accuracy. This ensures clinician trust and validation, making the model more applicable to real-world settings.

6.2 Evaluation of trade-offs between performance and efficiency

The brain tumor segmentation pipeline was evaluated on the BRaTS 2021 dataset, balancing accuracy and computational efficiency. The T5-based hybrid model processed multiple cases efficiently, leveraging GPU acceleration (CUDA). The average total processing time per case was 0.85s, with 0.52s for inference, indicating fast execution suitable for clinical workflows.

However, efficiency trade-offs were evident in memory consumption, with an average increase of 287.64MB per case. While this ensures high segmentation accuracy, optimizing model size and computational load is crucial for deployment in resource-constrained environments. The XAI-enhanced framework further adds interpretability but introduces slight computational overhead.

Overall, the system achieved a balance between segmentation accuracy, inference speed, and memory usage, making it viable for real-time applications with scope for further optimization.

7 Conclusion and Future Work

This study introduces a hybrid UNet++ model integrating ResNet50, TinyBERT, and XAI for brain tumor segmentation, achieving high accuracy and interpretability on the BRaTS 2021 dataset. ResNet enhances feature extraction, TinyBERT refines global context, and XAI methods (SLICE, GradCAM, Transformer Attention) provide clinically relevant insights, improving trust and adoption.

Future work will focus on real-time clinical integration, multi-modal disease segmentation, edge-device optimization, and enhanced XAI frameworks. Further human-AI collaboration through LLM-driven clinical dialogue systems can strengthen AI-assisted diagnostics, ensuring greater scalability and usability in healthcare.

References

1. Abdusalomov, A.B., Mukhiddinov, M., Whangbo, T.K.: Brain tumor detection based on deep learning approaches and magnetic resonance imaging. Cancers **15**(16), 4172 (2023)
2. Abgrall, G., Holder, A.L., Chelly Dagdia, Z., Zeitouni, K., Monnet, X.: Should ai models be explainable to clinicians? Critical Care **28**(1), 301 (2024)
3. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Hasan, M., Van Essen, B.C., Awwal, A.A., Asari, V.K.: A state-of-the-art survey on deep learning theory and architectures. electronics **8**(3), 292 (2019)

4. Alweshah, A., Barzamini, R., Hajati, F., Farahani, S.S.S., Arabian, M., Sohani, B.: Temporal dependency modeling for improved medical image segmentation: The r-unet perspective. Franklin Open **9**, 100182 (2024)
5. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arxiv 2021. arXiv preprint arXiv:2107.02314 (2021)
6. Bukhari, S.T., Mohy-ud Din, H.: E1d3 u-net for brain tumor segmentation: Submission to the rsna-asnr-miccai brats 2021 challenge. In: International MICCAI Brainlesion Workshop. pp. 276–288. Springer (2021)
7. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical image analysis **79**, 102444 (2022)
8. Davis, M., King, S., Good, N., Sarvas, R.: From context to content: leveraging context to infer media metadata. In: Proceedings of the 12th annual ACM international conference on Multimedia. pp. 188–195 (2004)
9. Esmaeilzadeh, P.: Challenges and strategies for wide-scale artificial intelligence (ai) deployment in healthcare practices: A perspective for healthcare organizations. Artificial Intelligence in Medicine **151**, 102861 (2024)
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
11. Jiang, X., Wang, S., Zhang, Y.: Vision transformer promotes cancer diagnosis: A comprehensive review. Expert Systems with Applications **252**, 124113 (2024)
12. Jiménez-Sánchez, A., Avlona, N.R., Juodelyte, D., Sourget, T., Vang-Larsen, C., Rogers, A., Zajac, H.D., Cheplygina, V.: Copycats: the many lives of a publicly available medical imaging dataset. arXiv preprint arXiv:2402.06353 (2024)
13. Kumar, G.M.K., Chadha, A., Mendola, J., Shmuel, A.: Medvisionllama: Leveraging pre-trained large language model layers to enhance medical image segmentation. arXiv preprint arXiv:2410.02458 (2024)
14. Mall, P.K., Singh, P.K., Srivastav, S., Narayan, V., Paprzycki, M., Jaworska, T., Ganzha, M.: A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. Healthcare Analytics p. 100216 (2023)
15. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of biomedical informatics **113**, 103655 (2021)
16. Mienye, I.D., Obaido, G., Jere, N., Mienye, E., Aruleba, K., Emmanuel, I.D., Ogbuokiri, B.: A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. Informatics in Medicine Unlocked p. 101587 (2024)
17. Pereira, G.A., Hussain, M.: A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships. arXiv preprint arXiv:2408.15178 (2024)
18. Ranjbarzadeh, R., Bagherian Kasgari, A., Jafarzadeh Ghoushchi, S., Anari, S., Naseri, M., Bendechache, M.: Brain tumor segmentation based on deep learning and an attention mechanism using mri multi-modalities brain images. Scientific Reports **11**(1), 1–17 (2021)
19. Rayed, M.E., Islam, S.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.: Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in Medicine Unlocked p. 101504 (2024)

20. Rony, M.K.K., Parvin, M.R., Ferdousi, S.: Advancing nursing practice with artificial intelligence: Enhancing preparedness for the future. Nursing open **11**(1) (2024)
21. Sadeghi, Z., Alizadehsani, R., CIFCI, M.A., Kausar, S., Rehman, R., Mahanta, P., Bora, P.K., Almasri, A., Alkhawaldeh, R.S., Hussain, S., et al.: A review of explainable artificial intelligence in healthcare. Computers and Electrical Engineering **118**, 109370 (2024)
22. Sharma, N., Aggarwal, L.M.: Automated medical image segmentation techniques. Journal of medical physics **35**(1), 3–14 (2010)
23. Singh, S.P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., Gulyás, B.: 3d deep learning on medical images: a review. Sensors **20**(18), 5097 (2020)
24. Talaat, F.M., Gamel, S.A., El-Balka, R.M., Shehata, M., ZainEldin, H.: Grad-cam enabled breast cancer classification with a 3d inception-resnet v2: Empowering radiologists with explainable insights. Cancers **16**(21), 3668 (2024)
25. Teng, Z., Li, L., Xin, Z., Xiang, D., Huang, J., Zhou, H., Shi, F., Zhu, W., Cai, J., Peng, T., et al.: A literature review of artificial intelligence (ai) for medical image segmentation: from ai and explainable ai to trustworthy ai. Quantitative Imaging in Medicine and Surgery **14**(12), 9620 (2024)
26. Thakur, G.K., Thakur, A., Kulkarni, S., Khan, N., Khan, S.: Deep learning approaches for medical image analysis and diagnosis. Cureus **16**(5) (2024)
27. Tian, D., Jiang, S., Zhang, L., Lu, X., Xu, Y.: The role of large language models in medical image processing: a narrative review. Quantitative Imaging in Medicine and Surgery **14**(1), 1108 (2023)
28. Umirzakova, S., Mardieva, S., Muksimova, S., Ahmad, S., Whangbo, T.: Enhancing the super-resolution of medical images: Introducing the deep residual feature distillation channel attention network for optimized performance and efficiency. Bioengineering **10**(11), 1332 (2023)
29. Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Li, Z., Fu, J.: Pre-trained language models in biomedical domain: A systematic survey. ACM Computing Surveys **56**(3), 1–52 (2023)
30. Wang, B., Yang, J., Peng, H., Ai, J., An, L., Yang, B., You, Z., Ma, L.: Brain tumor segmentation via multi-modalities interactive feature learning. Frontiers in Medicine **8**, 653925 (2021)
31. Wu, J., Li, C., Gensheimer, M., Padda, S., Kato, F., Shirato, H., Wei, Y., Schönlieb, C.B., Price, S.J., Jaffray, D., et al.: Radiological tumour classification across imaging modality and histology. Nature machine intelligence **3**(9), 787–798 (2021)
32. You, C., Zhao, R., Liu, F., Dong, S., Chinchali, S., Topcu, U., Staib, L., Duncan, J.: Class-aware adversarial transformers for medical image segmentation. Advances in Neural Information Processing Systems **35**, 29582–29596 (2022)
33. Yuan, M., Bao, P., Yuan, J., Shen, Y., Chen, Z., Xie, Y., Zhao, J., Li, Q., Chen, Y., Zhang, L., et al.: Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. Medicine Plus p. 100030 (2024)
34. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)