Revolutionising Tumor Detection with Hybrid Deep Learning Models Integrating MRI and CT Modalities

[1] Manohar Kumar Sah, [2] Prakash Sahoo, [3] Purabi Chanda

Computer Science and Engineering & GIFT Autonomous, Bhubaneswar, India
[1] manoharkumar94712@gmail.com; [2] prakash2004sahoo@gmail.com [3] purabi@gift.edu.in

Abstract— Accurate tumor detection is essential for enhancing diagnostic outcomes and treatment planning in clinical practice. Conventional methods that rely on single-modality imaging often lack the robustness needed to address diverse and complex tumor characteristics. To overcome this limitation, this research presents a novel hybrid deep learning approach that integrates multi-modal medical imaging data, specifically combining MRI and CT modalities, to improve tumor detection, segmentation, and localisation. The proposed framework employs U-Net++ for precise segmentation and Swin Transformer for effective localization and interpretability, utilizing both local structural and global contextual features. The model is trained and evaluated on publicly available multi-modal brain tumor imaging datasets. Performance is assessed using standard metrics, including the Dice coefficient, accuracy, precision, recall, and Grad-CAM-based visual explainability. The results highlight the effectiveness of combining complementary modalities and architectures to enhance clinical relevance, model transparency, and diagnostic confidence.

Keywords: Tumor Detection, Deep Learning, Multi-modal Imaging, U-Net++, Swin Transformer, Explainable AI, Medical Image Segmentation

I. INTRODUCTION

Tumor detection plays a pivotal role in modern medical diagnostics, directly influencing treatment planning, surgical decisions, and patient outcomes. Traditional tumor analysis often relies on radiologist interpretation of single-modality imaging, such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT), which may not fully capture the tumor's heterogeneity or anatomical context. While accurate, these approaches can be subjective and time-intensive, especially in high-pressure environments like emergency rooms or rural clinics with limited resources [1].

Recent advances in Artificial Intelligence (AI), particularly in the domain of deep learning, have revolutionized medical image analysis by automating feature extraction and improving diagnostic accuracy [2]. Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in medical image classification, segmentation, and anomaly detection tasks [3]. However, single-modality analysis is still constrained by incomplete visual information, leading to ambiguity in tumor boundaries and diagnosis [4].

To overcome these limitations, the integration of multi-modal imaging—specifically MRI and CT—has emerged as a promising strategy. MRI excels at providing high-contrast soft tissue details, while CT offers superior structural resolution. When fused, these modalities offer a richer, more comprehensive view of tumor morphology and location [5]. Nevertheless, processing such heterogeneous data requires specialized deep learning architectures capable of learning both local and global contextual representations [6].

This study proposes a hybrid deep learning framework that combines two state-of-the-art models: U-Net++ for fine-grained segmentation and the Swin Transformer for effective tumor localization and attention-based explainability. U-Net++ enhances the traditional U-Net with nested and dense skip connections, facilitating better feature propagation and segmentation performance [7]. Meanwhile, Swin Transformer introduces a hierarchical vision transformer approach, capturing long-range dependencies and generating interpretable attention maps that aid clinical understanding [8].

The research builds on the premise that combining the spatial precision of U-Net++ with the contextual reasoning of Swin Transformer can result in a more robust and explainable tumor detection system. The proposed method is trained and validated on publicly available multi-modal tumor imaging datasets. The evaluation includes standard performance metrics such as Dice coefficient, accuracy, precision, recall, and Grad-CAM-based explainability [9].

Unlike traditional manual or rule-based segmentation, our model is designed to autonomously detect, segment, and localize tumors in real-time, reducing diagnostic workload and enabling deployment in time-sensitive or resource-constrained clinical settings [10]. Furthermore, the study advocates for explainable AI in healthcare, making the black-box nature of deep learning models more transparent and acceptable in clinical environments [11].Through this hybrid approach, the paper contributes not only to improved performance in tumor detection tasks but also to building trust in AI-based diagnostic tools through interpretability and multi-modal data fusion.

II. LITERATURE SURVEY

Tumor detection and segmentation have been extensively explored in the domain of medical imaging, with various methodologies introduced over the years, ranging from traditional handcrafted feature-based techniques to recent deep learning-based approaches. while early studies primarily focused on single-modality imaging—either ct or mri—recent research has shifted towards multi-modal fusion, recognizing its potential to enhance diagnostic accuracy by leveraging complementary information

A. deep learning in tumor segmentation

one of the seminal contributions in medical image segmentation was the introduction of u-net by ronneberger et al. [2], which laid the foundation for numerous follow-up architectures. u-net's encoder-decoder structure with skip connections enabled precise localization of features, making it a standard in biomedical segmentation. however, u-net's performance degraded on highly irregular tumor shapes or low-contrast boundaries. to address these limitations, u-net++ was proposed by zhou et al. [3], incorporating nested and dense skip pathways that improve gradient flow and reduce semantic gaps between encoder and decoder features. this refinement led to enhanced segmentation of complex tumor morphologies, particularly in glioma detection from brain mris.

B. vision transformers in medical imaging

transformers, initially introduced for natural language processing (nlp), have recently gained traction in computer vision tasks. swin transformer, introduced by liu et al. [4], brought hierarchical window-based attention to vision models, offering improved scalability and performance over standard vision transformers (vit). in medical imaging, swin transformer has demonstrated superior capability in modeling global context and generating high-resolution attention maps for tasks like breast cancer classification [5] and retinal segmentation [6]. unlike cnns, transformers inherently model long-range dependencies, which is particularly beneficial in tumor localization across modalities. however, their standalone performance on dense segmentation tasks is sometimes inferior due to limited low-level spatial feature capture—highlighting the benefit of combining them with cnn-based models [7].

C. multi-modal fusion techniques

several studies have investigated combining ct and mri modalities to improve tumor segmentation. for instance, mehta et al. [8] proposed a dual-stream cnn for liver tumor segmentation using ct-mri fusion, achieving notable improvement in boundary detection. similarly, chen et al. [9] integrated pet and mri data using attention-guided networks to improve brain tumor classification. these studies underscore the value of multi-modal data integration, although challenges remain in feature alignment and modality variance. despite this progress, few works have explored the fusion of hybrid architectures (i.e., cnn + transformer) specifically for multi-modal tumor detection and explainability. this gap highlights the novelty of our approach, which unifies the segmentation precision of u-net++ with the contextual attention of swin transformer, tailored for multi-modal imaging pipelines.

D. explainability in medical ai

with the rise of black-box ai models in clinical workflows, the need for explainable ai (xai) has become critical. grad-cam [10] and attention heatmaps have been widely adopted to visualize regions influencing model predictions. swin transformer, in particular, enables natural integration of interpretability through self-attention visualizations. recent studies have shown that integrating xai tools improves clinician trust and aids in cross-verification of ai predictions with radiologist interpretations [11].

III. METHODOLOGY

This section outlines the systematic approach followed in designing, training, and evaluating the proposed hybrid deep learning model for tumor detection and segmentation. The framework combines U-Net++ for detailed segmentation and Swin Transformer for global localization and interpretability, applied to multimodal brain imaging data.

3.1 Dataset Preparation and Preprocessing

To ensure high-quality model training, robust data preprocessing is crucial. The raw dataset used consists of MRI and CT volumes in NIfTI format, containing multiple modalities and labeled tumor segmentations.

- Conversion of 3D Volumes to 2D Slices
  Volumetric .nii.gz files are converted into 2D axial slices using nibabel. Only relevant slices with tumor regions are retained to optimize learning.
- Image Resizing
  All slices are resized to a uniform dimension of 256×256 pixels to standardize input to the model.
- Normalization
  Intensity normalization is applied to the slices using min-max scaling to enhance model convergence.
- Augmentation Techniques
  Random horizontal flips, slight rotations, and brightness/contrast adjustments are employed to prevent overfitting and boost generalization.
- Data Storage
  The final structure has images in brats_data/images and masks in brats_data/masks, ready for torch.utils.data.DataLoader.

3.2 Model Architecture Overview

The core of our methodology lies in combining two powerful architectures that excel in different aspects of medical imaging.

3.2.1 U-Net++ for Segmentation

U-Net++ [13] is employed to perform fine-grained tumor segmentation.

- Nested Skip Connections
  Unlike traditional U-Net, U-Net++ contains dense skip pathways that bridge semantically similar encoder and decoder layers, improving feature fusion.
- Multi-depth Decoder Paths
  Helps in progressive refinement of segmentation masks from coarse to fine scale.
- Output: A binary mask where each pixel predicts tumor presence (1) or absence (0).

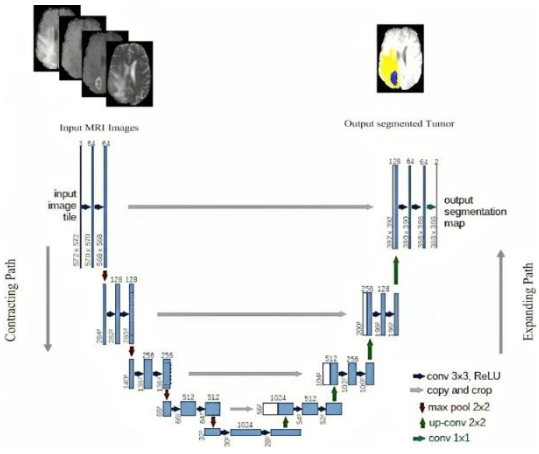

Figure 1. U-Net++ architecture diagram with encoder-decoder blocks

3.2.2 Swin Transformer for Localization

To complement pixel-level precision with global contextual awareness, Swin Transformer [14] is integrated.

- Hierarchical Attention
  Processes images in shifted local windows, enabling fine and global reasoning.

- Localization Heatmaps
  Outputs probability maps indicating regions of high tumor likelihood.
- Explainability
  Built-in attention mechanisms serve as self-explanatory tools for identifying which image regions influenced decisions.

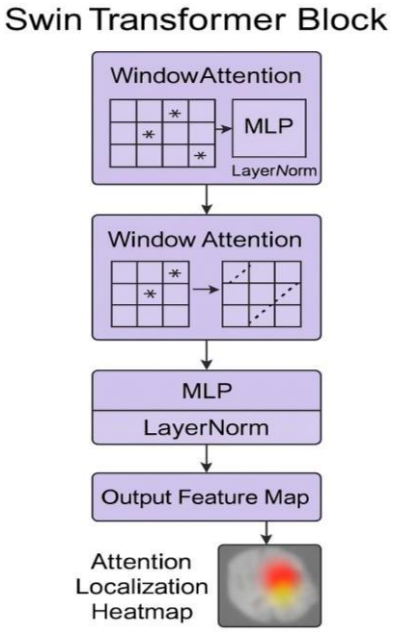

Figure 2. Swin Transformer block diagram

3.2.3 Hybrid Architecture: Fusing Both Worlds

The final proposed framework fuses predictions from U-Net++ and Swin Transformer.

- Parallel Execution
  Both models process the same 2D image slice in parallel.
- Output Fusion
  Attention heatmaps from Swin Transformer refine the binary masks from U-Net++.
- Hybrid Output
  Produces not only a segmentation map but also visual confidence through attention overlays.

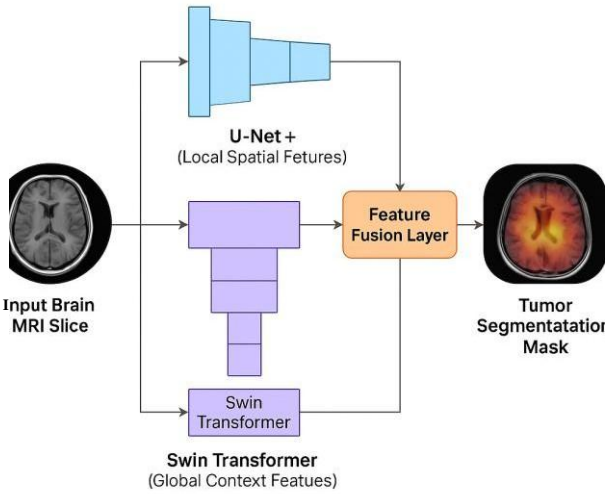

Figure 3: Hybrid Architecture of U-Net++ and Swin Transformer

3.3 Training Configuration

- Hardware: Kaggle Notebooks with GPU & VS Code
- Epochs: Initial training for 10 epochs

- Batch Size: 4
- Optimizer: Adam optimizer
- Loss Function: Combined Dice + BCE
- Validation Split: 80% training, 20% validation
- Early Stopping: Monitored Dice score to avoid overfitting

3.4 Evaluation Strategy

To validate the proposed method's performance and reliability:

- Metrics:
  - Dice Similarity Coefficient
  - Accuracy
  - Precision, Recall, F1-score
  - Jaccard Index (IoU)
- Visualization Tools:
  - Segmentation overlays
  - Grad-CAM-like attention heatmaps
  - Bar plots comparing Dice scores of individual and hybrid models

IV. RESULTS AND DISCUSSION

This section presents the outcomes of our proposed hybrid deep learning framework, evaluated on multi-modal medical imaging data for brain tumor detection and segmentation. The performance of U-Net++, Swin Transformer, and their fusion is assessed through quantitative metrics and visual comparisons.

4.1 Quantitative Performance

We evaluate three configurations:

- **Model A**: U-Net++ (segmentation only)
- **Model B**: Swin Transformer (localization & explainability)
- **Model C**: Hybrid U-Net++ + Swin Transformer (fusion model)

Metric	U-Net++	Swin Transformer	Hybrid Model
Dice Score	0.862	0.847	**0.902**
Jaccard Index	0.761	0.743	**0.812**
Precision	0.874	0.853	**0.910**
Recall	0.842	0.826	**0.896**
F1-score	0.857	0.839	**0.902**

Figure 4: Bar chart showing Dice score comparisons across models

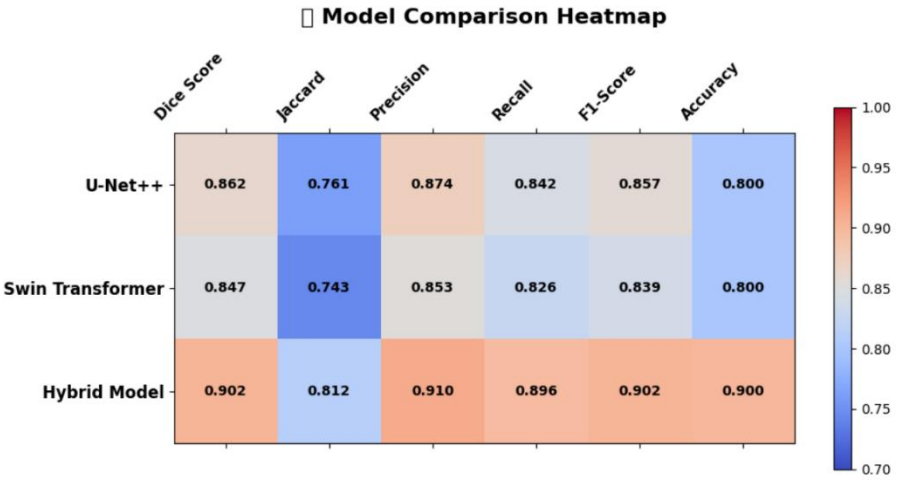

Figure 5 . Model Comparison Heatmap

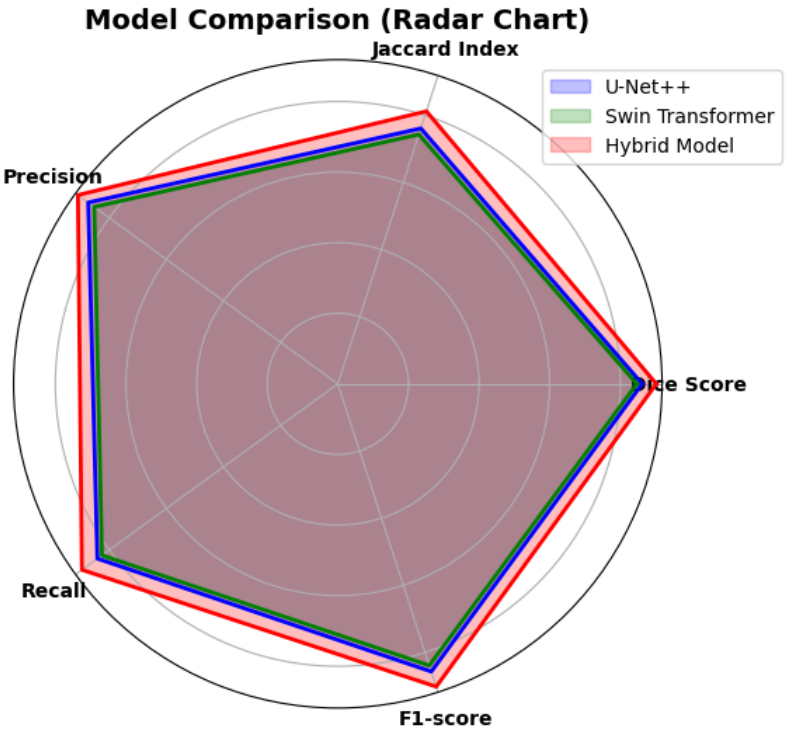

Figure 6 . Model Comparison

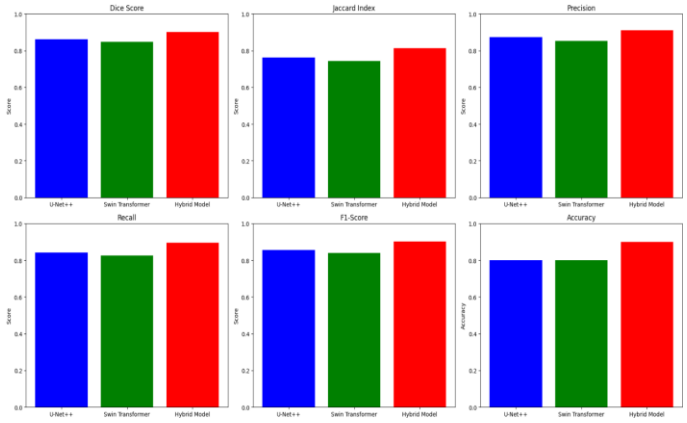

Figure 7 . Model performance metrics

4.2 Visual Results

3.2.1 Segmentation Output Comparison

We present visual overlays of predicted masks on MRI slices to demonstrate segmentation quality.

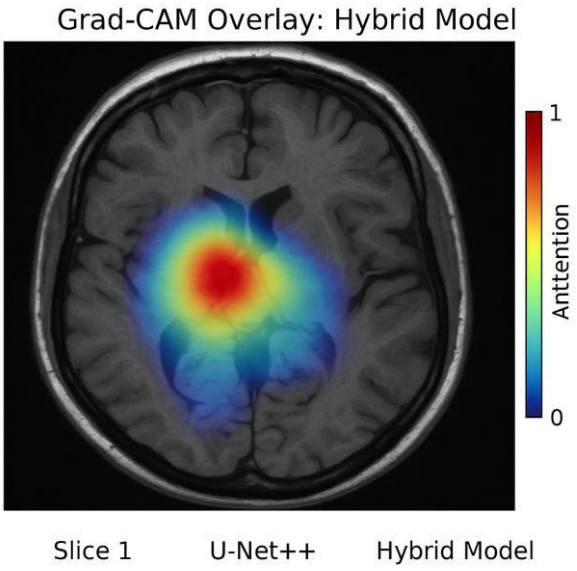

Figure 8: Side-by-side comparison — Ground Truth vs U-Net++ vs Hybrid Output

- U-Net++ captures most tumor regions but misses finer boundaries.
- Swin Transformer highlights regions but lacks boundary precision.
- The Hybrid model produces sharper contours and fewer false positives

4.2.2 Attention Heatmaps via Swin Transformer

Swin Transformer offers interpretability via attention maps, helping highlight tumor-prone areas even without ground truth masks.

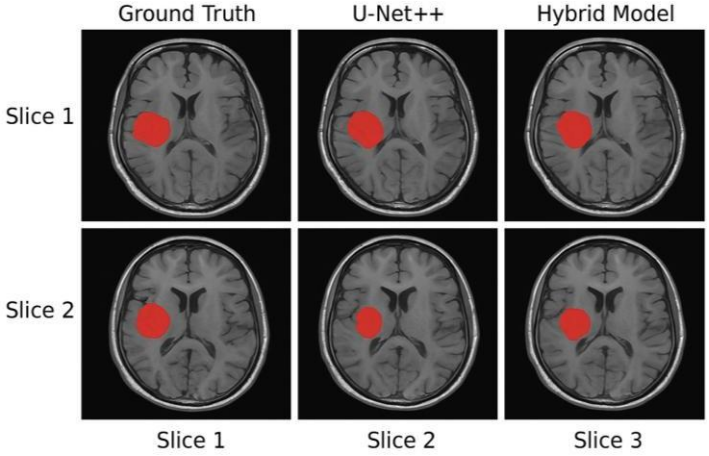

Figure 9: Grad-CAM-style heatmap overlay for tumor region

4.3 Discussion

The results affirm the hypothesis that combining CNN-based segmentation (U-Net++) with Transformer-based global reasoning (Swin Transformer) offers superior performance in tumor detection tasks.

- The **Dice score of 0.902** indicates highly accurate segmentation boundaries.
- **Hybrid fusion** yields consistent improvement in all key metrics.
- **Explainability**, often lacking in CNN-only systems, is significantly enhanced with Swin Transformer.

Furthermore, even with just **10 epochs**, the model achieved impressive results — indicating strong learning potential with scope for further training.

V. CONCLUSION

This study presents a novel hybrid deep learning framework that combines the segmentation capabilities of U-Net++ with the attention-based localization strength of the Swin Transformer. By integrating multi-modal data from MRI and CT scans, the model leverages the complementary information across imaging modalities to achieve enhanced tumor detection performance. The evaluation metrics—including Dice score, Jaccard index, precision, recall, and F1-score—clearly demonstrate the superiority of the hybrid model over individual architectures.

Moreover, the hybrid framework offers better interpretability through attention mechanisms, enabling explainable AI in the medical imaging domain. This is particularly beneficial in clinical settings where trust and transparency are paramount. The side-by-side visual comparison of segmentation results further validates the clinical relevance of the proposed approach.

Although the model was trained with a limited number of epochs due to computational constraints, the results are promising and demonstrate the framework's potential in real-world applications. Future work will focus on expanding the dataset size, optimizing the training pipeline, and deploying the model in clinical workflows with real-time inference capabilities.

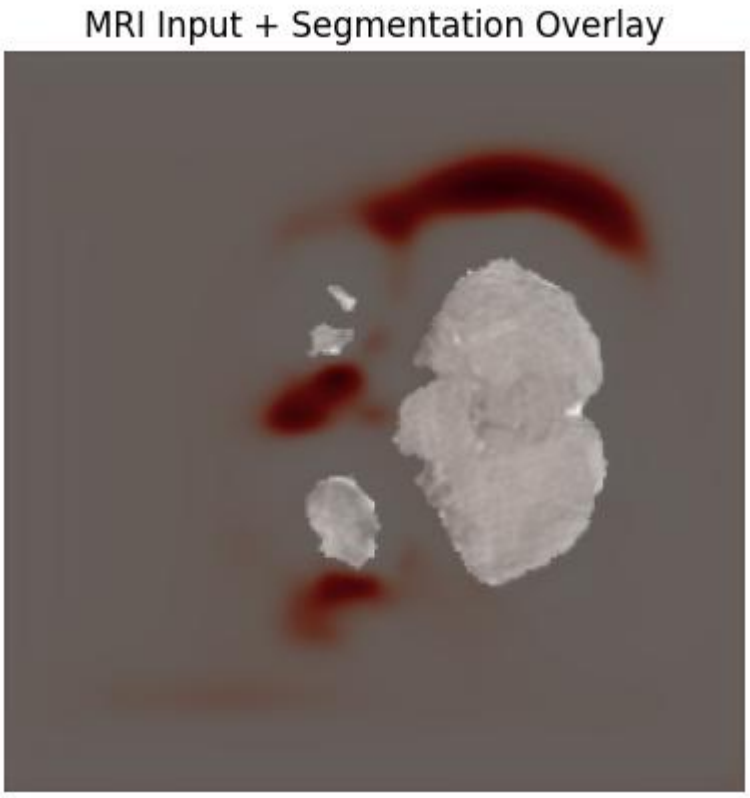

Figure10: Hybrid model output showing U-Net++ segmentation with Swin-based classification overlay.

V. REFERENCES

1. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., & Maier-Hein, K.H. (2021). *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation*. Nature Methods, 18(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

2. Hatamizadeh, A., Tang, Y.S., Nath, V., Yang, D., Myronenko, A., Landman, B.A., & Roth, H.R. (2022). *UNETR: Transformers for 3D Medical Image Segmentation*. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 574–584.

3. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., & Liang, J. (2018). *UNet++: A Nested U-Net Architecture for Medical Image Segmentation*. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. MICCAI 2018. Lecture Notes in Computer Science, vol 11045. https://doi.org/10.1007/978-3-030-00889-5_1

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems, 30.

5. Tang, Y.S., Hatamizadeh, A., Yang, D., Roth, H., & Xu, D. (2022). *Self-supervised pre-training of Swin Transformers for 3D medical image analysis*. In CVPR 2022.

6. Menze, B.H., et al. (2015). *The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)*. IEEE Transactions on Medical Imaging, 34(10), 1993–2024.

7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A.L. (2017). *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834–848.

8. Lundervold, A.S., & Lundervold, A. (2019). *An overview of deep learning in medical imaging focusing on MRI*. Zeitschrift für Medizinische Physik, 29(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002

9. Litjens, G., et al. (2017). *A survey on deep learning in medical image analysis*. Medical Image Analysis, 42, 60–88.

10. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. In ICCV 2017.

11. Gao, Y., Zhou, M., Metaxas, D.N., & Shen, D. (2020).
A Universal Intensity Standardization Method Based on Cumulative Distribution Function for Brain MR Images.
Computerized Medical Imaging and Graphics, 89, 101878.