Leveraging Local and Global Features for Enhanced Segmentation of Brain Metastatic Tumors in Magnetic Resonance Imaging

Mojtaba Mansouri Nejad[1, 4 ,5], Habib Rostami[2, 3, *], Ahmad Keshavarz[1, 4], Hojat Ghimatgar[1, 5], Mohamad Saleh Rayani[1, 4, 5], Leila Gonbadi[2,3]

[1]Department of Electrical Engineering, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, Bushehr 7516913817, Iran

[2]Department of Computer Engineering, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, Bushehr 7516913817, Iran

[3]Artificial Intelligence and Intelligent Healthcare Lab, Artificial Intelligence and Data Mining Research Group, ICT Research Institute, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, 7516913817, Bushehr, Iran.

[4]IoT and Signal Processing Research Group, ICT Research Institute, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, 7516913817, Bushehr, Iran.

[5] Computational Neuroscience Lab. ICT Research Institute, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, 7516913817, Bushehr, Iran.

* Author in Correspondence (habib@pgu.ac.ir)

Keywords: CNN, MRI, VIT

Abstract: Metastatic brain tumors present significant challenges in diagnosis and treatment, contributing to high mortality rates worldwide. Magnetic Resonance Imaging (MRI) is a pivotal diagnostic tool for identifying and assessing these tumors. However, accurate segmentation of MRI images remains critical for effective treatment planning and prognosis determination. Traditional segmentation methods, including threshold-based algorithms, often struggle with precisely delineating tumor boundaries, especially in three-dimensional (3D) images. This paper introduces a 3D segmentation framework that combines Swin Transformers and 3D U-Net architectures, leveraging the complementary strengths of these models to improve segmentation accuracy and generalizability for metastatic brain tumors. We train multiple 3D U-Net and Swin-U-Net models, selecting the best-performing architectures for segmenting tumor voxels. The outputs of these networks are then combined using various strategies, such as logical operations and stacking the outputs with the original images, to guide the training of a third model. Our method employs an innovative ensemble approach, integrating these outputs into a unified prediction model to enhance performance reliability. Experimental analysis on a newly released metastasis brain tumor dataset, which to the best of our knowledge has been tested for the first time using our models, yielded an impressive accuracy of 73.47 %, validating the effectiveness of the proposed architectures.

1. Introduction

Frequently, tumors identified in the brain originate elsewhere in the body before disseminating to the brain. [1] These are referred to as metastatic brain tumors or brain metastases. In the central nervous system (CNS), brain tumors account for 85% to 90% of primary cancer cases. [2]. In 2023, the US reported 24,810 new cases of brain tumors and other cancers of the nervous system. From 2015 to 2019, the Surveillance, Epidemiology, and End Results (SEER) Program database provided statistics on the cumulative incidence of brain and other CNS malignancies in the United States, which came to 6.3 per 100,000 individuals annually. An estimated 4.4 deaths per 100,000 people were expected to occur annually, based on data collected between 2016 and 2020 [2].

The doctor first evaluates the patient's symptoms and medical history to determine whether an MRI is necessary and which specific area or organ needs to be investigated before beginning

the MRI diagnosis process. After the scan, a computer processes the raw data to create detailed photographs of the region that is being scanned. After that, a radiologist—a specialist in deciphering medical images—looks over the images and assesses the features, dimensions, and strengths of the structures. After the MRI images have been interpreted, a diagnosis is made and a thorough report is created for the patient and the referring doctor to review [3]. The radiologist's segment lesions to identity the border of the lesions and estimate their volume before planning the treatment.

In computer vision, semantic image segmentation is the process of classifying individual pixels in an image into instances, each of which is assigned a class. This task is in scene comprehension, or more precisely, explaining an image entire context. Applications of image segmentation include radiation, image-guided surgery, and improved radiological diagnostics in the field of medical image analysis [4].

Image segmentation is a crucial part of medical image analysis. Whether the target tissues are diseased or not, image segmentation can automatically identify their architecture, providing medical professionals with the information they need for a follow-up diagnostic and treatment plan. Most of the early methods for segmenting medical images were threshold-based segmentation algorithms [4]. However, recently, deep learning-based methods have been employed for segmentation, demonstrating higher accuracy compared to previous approaches[5]. In segmentation methods for three-dimensional images such as MRI, two-dimensional convolution methods applied to individual slices. This simplification leads to the loss of spatial information, reducing segmentation accuracy and limiting their applicability for complex, high-resolution medical imaging tasks [6]**.**

To address these limitations, this paper introduces a 3D segmentation method that uniquely combines the complementary strengths of 3D transformers and 3D convolutional networks in a hybrid ensemble approach. This framework enhances stability, accuracy, and generalizability in MRI image segmentation for metastatic brain tumors. Experimental analysis shows that our proposed method surpasses a simple U-Net or the newly purposed Swin-U-Net models in terms of various evaluation metrics. The ensemble strategy further refines these results, marking a significant advancement in 3D medical image segmentation techniques with promising implications for the future of tumor detection and treatment.

The main contributions of this paper can be summarized as follows:

1. **Novel Hybrid Approach:** We propose a hybrid ensemble framework that leverages the complementary strengths of convolutional neural networks (CNNs) and Swin Transformers for accurate brain metastasis tumor segmentation, addressing limitations in representing 3D spatial dependencies.
2. **Benchmarking on a New Dataset:** We present the first application of our ensemble method on a newly released metastasis tumor segmentation dataset, achieving state-of-the-art accuracy and providing a baseline for future studies.

3. **Optimization via Fusion Strategies:** Experiments were conducted to analyze various fusion strategies, revealing that the integration of logical fusion with additional learning via Swin U-Net3D is the most effective and computationally efficient approach.

The remainder of the paper is organized as follows: Section 2 reviews the related work. Section 3 details the proposed method. Section 4 presents the experimental results. Section 5 discusses the findings and their implications. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related work

Recent studies have highlighted the innovative use of neural networks and machine learning tools in the diagnosis and segmentation of brain metastases. In this section, we review the most relevant works that employ deep learning techniques to segment brain tumors and metastatic brain tumors in MRI images.

Seo et al. [7] reviewed deep learning architectures, including recurrent neural networks (RNNs), CNNs, and artificial neural networks (ANNs), as well as traditional machine learning algorithms. The review underscores the advancements in deep learning for neuro-oncology and its potential to impact brain metastasis diagnosis significantly. However, while this review establishes the utility of deep learning, it lacks a discussion of 3D segmentation or hybrid models, leaving a gap in addressing the specific challenges of metastatic segmentation.

Jie Xue et al. [8] proposed a fully cascaded 3D convolutional network for detecting and segmenting brain metastases using 3D MRI images. While effective for segmenting large lesions, their method struggles with small and scattered metastases, a common challenge in real-world datasets. Furthermore, the reliance on data from a single hospital limits the generalizability of their approach.

Grøvik et al. [9] employed a modified GoogLeNet 2.5D architecture for detecting and segmenting brain metastases using multi-sequence MRI data. Despite promising results, the 2.5D approach fails to fully utilize 3D spatial information, leading to suboptimal segmentation accuracy for more complex cases.

Yu et al. [10] introduced an ensemble approach combining 3D U-Net and DeepMedic for brain metastasis detection. Although this method improves performance, its reliance on private datasets and the absence of a robust framework for fusing local and global features limit its scalability and generalizability to public datasets.

Yoo et al. [11] proposed a 2.5D overlapping patch technique to enhance 2D U-Net's performance for small lesion detection. While their approach improves sensitivity for smaller metastases, it sacrifices efficiency and fails to capture comprehensive 3D spatial relationships, which are critical for accurate segmentation.

Abu Saleh et al. [12] developed a hybrid network combining 3D U-Net, VNet, and transformers for brain tumor segmentation. This model effectively leverages transformers for

global context but focuses on non-metastatic brain tumors using BraTS 2020 data, leaving metastatic segmentation unexplored.

Zhang et al. [13] introduced custom modules like Enblock and InitConv for BraTS 2021 data. However, their work is centered on general brain tumor segmentation, with no application to metastatic cases. Similarly, Al-Khalil et al. [14] employed GANs for data augmentation and Fourier adaptations, focusing on BraTS 2020 without addressing the challenges of metastatic brain tumor segmentation.

Limitations of Existing Methods:

Most existing methods rely on private datasets, making reproducibility and generalization challenging. Furthermore, approaches using 2.5D or 2D slices compromise the full utilization of 3D spatial information, leading to reduced segmentation accuracy for intricate metastases. Ensemble methods, while effective, often lack a robust strategy for feature fusion, which is critical for capturing both local and global tumor characteristics.

Proposed Method:

This study uses a recently published public dataset [15] comprising complex MRI images of brain metastases, which include small and scattered lesions across multiple brain layers. We propose a hybrid ensemble approach that uniquely combines Swin Transformers and 3D U-Net models, addressing the limitations of prior methods by preserving 3D spatial relationships and employing innovative fusion strategies to integrate local and global features. Experimental results demonstrate that this method outperforms state-of-the-art models, including standalone U-Net and Swin-U-Net, in terms of segmentation accuracy, especially for small metastases.

A summary of related work and a comparison of existing methods is provided in Table 1, focusing on differences in approach (2D vs. 3D), dataset accessibility (public vs. private), and the inclusion of brain metastases. In contrast to our approach, other studies use either a private dataset, don't include metastasis tumors, or don't utilize 3D models. This underscores both the novelty and the significance of our work in addressing the challenges of metastatic brain tumor segmentation.

Table 1: Summary of related work.

	Approach	Database	Metastatic	Dimension
Huang [7]	Ensemble(DeepMedic+)	Private data	Yes	3
Jie Xue [8]	Cascaded 3D Fully Convolutional Network	Private data	Yes	3
Grøvik [9]	Modified GoogLe Net	Private data	Yes	3
Yeu [10]	Ensemble(3D U-Net, DeepMedic)	Private data	Yes	3
Sang Kyun Yoo [11]	Deep-Learning-Based Automatic Detection	Private data	Yes	2.5
Aboussaleh [12]	Ensemble (VNet, 3D U-Net, Transformer)	BraTS 2020	No	3

Zhang [13]	Enblock, InitConv, DeUp_Cat, DeBlock	BraTS 2021	No	3
Al-Khalil [14]	GAN, Fourier Adaptation	BraTS 2020	No	3
Proposed method	Ensemble (3D Swin, 3D U-Net, 3D Swin + U-Net)	B. Ocaña-Tienda *et al* [15]	YES	3

3. Materials and Methods

In this paper, we propose an ensemble model architecture, leveraging the strengths of different models for segmentation of brain metastasis tumors. The neural network used in our proposed method is a 3D U-Net architecture with a mixture of convolutional blocks and Swin transformer blocks in two parallel paths as it's feature extractors. To explain the proposed architecture and methods, initially, we provide a review of the Swin Transformer [16], U-Net [17] and their corresponding 3D versions. Then, we outline the architecture in detail.

3.1. Swin

Vaswani et al.'s [18] introduction of transformers has revolutionized the field of natural language processing and beyond. Unlike conventional sequence-to-sequence models, transformers analyze input data in parallel using self-attention mechanisms enabling effective handling of sequential dependencies. The architecture's success lies in its ability to capture long-range dependencies, enhancing performance in tasks such as sentiment analysis, image recognition, and language translation. This adaptability across various domains underscores transformers' significance and has significantly influenced advancements in artificial intelligence, propelling the field of deep learning forward.

The Vision Transformer (ViT) processes images by partitioning them into patches, which are subsequently treated as tokens akin to words in natural language processing. These patches undergo linear embedding and are then inputted into a standard transformer encoder, initially intended for sequential data such as sentences. This methodology enables the ViT to adeptly capture spatial relationships and features within the image, thus facilitating exceptional accuracy and efficiency in image recognition tasks [19].

The Swin Transformer, an advancement that refines the self-attention mechanism, has emerged as an extension of the pioneering transformer architecture. Introduced by Liu et al. [16], the Swin Transformer employs a hierarchical processing approach. It partitions the input data into non-overlapping patches and processes each patch independently before integrating information across different scales. This method enhances computational efficiency and facilitates scalability, overcoming challenges associated with processing lengthy images and sequences. The Swin

Transformer has showcased effectiveness in handling complex visual data and has exhibited remarkable performance in various computer vision applications, including object recognition and image categorization, amidst ongoing advancements in the industry. The overall Swin transformer architected can be observed in Figure 1.

The Swin Transformer is assembled by substituting the traditional Multi-Head Self-attention (MSA) module in one transformer block with a shifted window-based module, while retaining the same configuration for the remaining layers. As illustrated in Figure 2, a two-layer MLP with a GELU nonlinearity intervenes between a shifted window-based MSA module within a Swin Transformer block. Layer normalization (LN) is applied before each MSA module, each MLP, and each residual connection. The equations of the Swin block are as shown in (1).

$$
\begin{aligned}
\hat{\boldsymbol{z}}^{\boldsymbol{l}} &= \boldsymbol{W-MSA}\left(\boldsymbol{LN}\left(\boldsymbol{z}^{\boldsymbol{l-1}}\right)\right) + \boldsymbol{z}^{\boldsymbol{l-1}} \\
\boldsymbol{z}^{\boldsymbol{l}} &= \boldsymbol{MLP}\left(\boldsymbol{LN}\left(\hat{\boldsymbol{z}}^{\boldsymbol{l}}\right)\right) + \hat{\boldsymbol{z}}^{\boldsymbol{l}} \\
\hat{\boldsymbol{z}}^{\boldsymbol{l+1}} &= \boldsymbol{SW-MSA}\left(\boldsymbol{LN}\left(\boldsymbol{z}^{\boldsymbol{l}}\right)\right) + \boldsymbol{z}^{\boldsymbol{l}} \\
\boldsymbol{z}^{\boldsymbol{l+1}} &= \boldsymbol{MLP}\left(\boldsymbol{LN}\left(\hat{\boldsymbol{z}}^{\boldsymbol{l+1}}\right)\right) + \hat{\boldsymbol{z}}^{\boldsymbol{l+1}},
\end{aligned}
\qquad (1)
$$

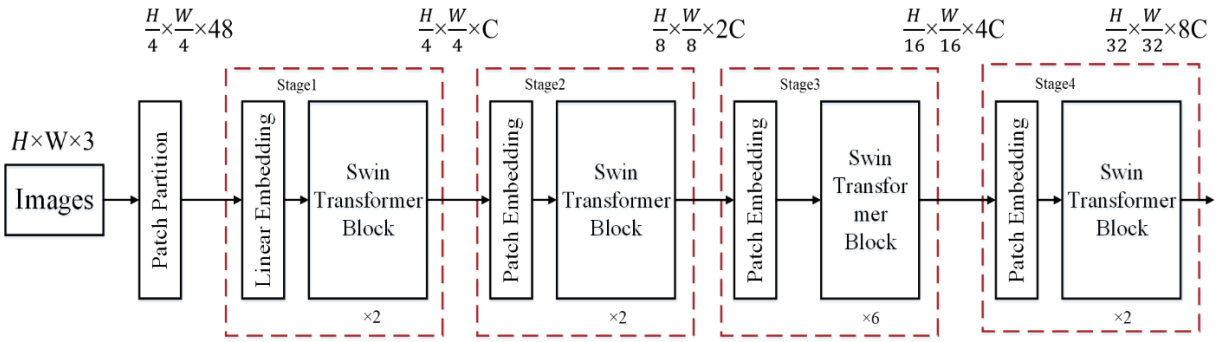

Figure 1: Overall structure of a Swin Transformer (Swin-T).

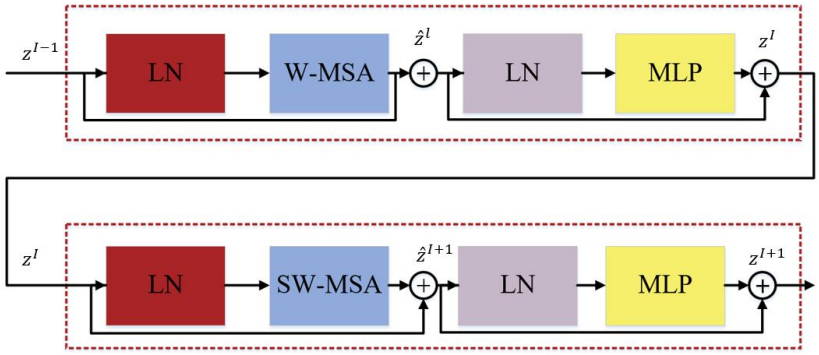

Figure 2: Structure of a Swin block.

3.2. U-Net

The U-Net architecture [17] comprises a contracting path and a symmetric expanding path. The contracting path functions as a typical convolutional network, capturing contextual information, while the expanding path facilitates accurate localization. In the contracting path, a sequence of convolutional and pooling operations is applied to decrease the spatial resolution of the input image while increasing the number of feature channels. Subsequently, the expanding path employs a series of up-convolutions and concatenations with feature maps from the contracting path to restore the original spatial resolution of the image. This enables the network to capture both local and global contextual information while maintaining precise localization. Moreover, the U-Net incorporates skip connections between the contracting and expanding paths to retain fine-grained details and enhance segmentation accuracy.

In Figure 3, a schematic representation of the network is depicted, specifically tailored for input data consisting of 224 * 224 pixels. This model configuration has been designed to enhance the understanding of the network's architecture when processing datasets with a 224 * 224 pixel dimension.

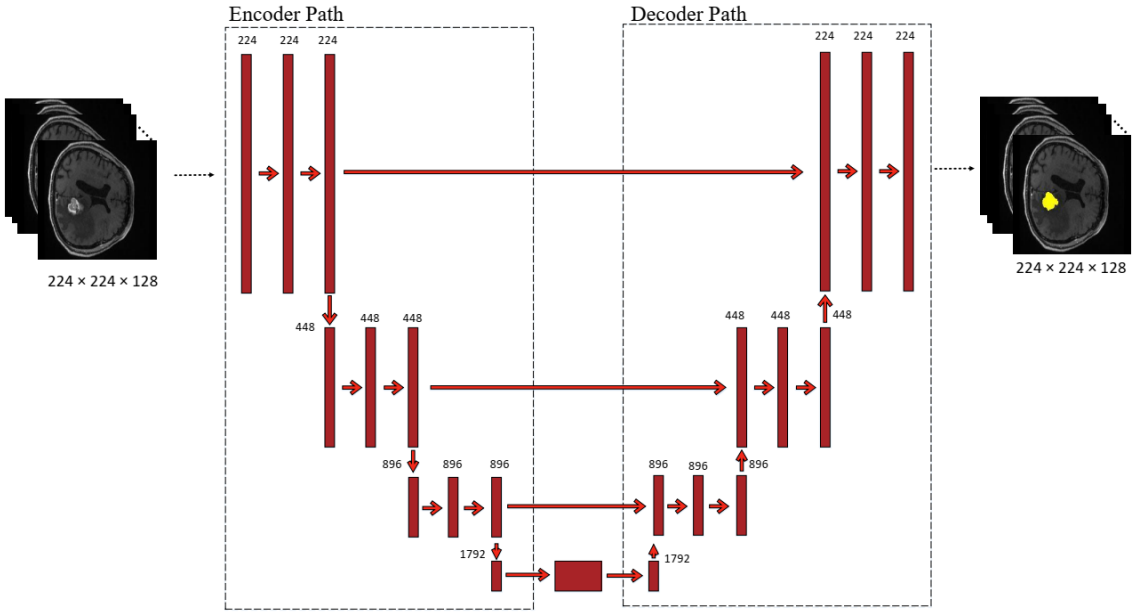

Figure 3: Depiction of the U-Net3D encoder-decoder architecture for a brain MRI image of size 224*224 with 128 slices.

3.3. Swin U-Net3D

Swin U-Net3D is a cutting-edge model designed to address the challenges of 3D medical image analysis by combining the strengths of CNNs and ViTs. While CNNs excel at capturing local spatial features, ViTs are effective in modeling long-range dependencies within images. This hybrid approach enables Swin U-Net3D to deliver precise voxel-based segmentation of complex medical images, such as MRI and CT scans, aiding in the identification and analysis of critical

regions like tumors or lesions. The model's innovative architecture is optimized for extracting both local and global contextual information, making it a powerful tool for advanced medical imaging tasks.

CNN and ViT are combined in the Swin U-Net 3D model to overcome the shortcomings of previous models in representing long- and short-distance dependent information in images. The model is intended for use in the voxel segmentation of 3D medical images.[20]

The encoder, decoder, and leap connection comprise the architecture of Swin Unet3D. Patch Merging3D, Swin Transformer Block3D, and Conv Block3D modules are utilized by the model for feature extraction, image downsampling, and feature extraction correspondingly. Whereas the Conv Block3D acquires knowledge of local dependencies, the Swin Block3D is in charge of learning long-range dependencies inside the image. Feature fusion of the image features extracted by these two modules is carried out by the model at the end of each decoder (Figure 4).

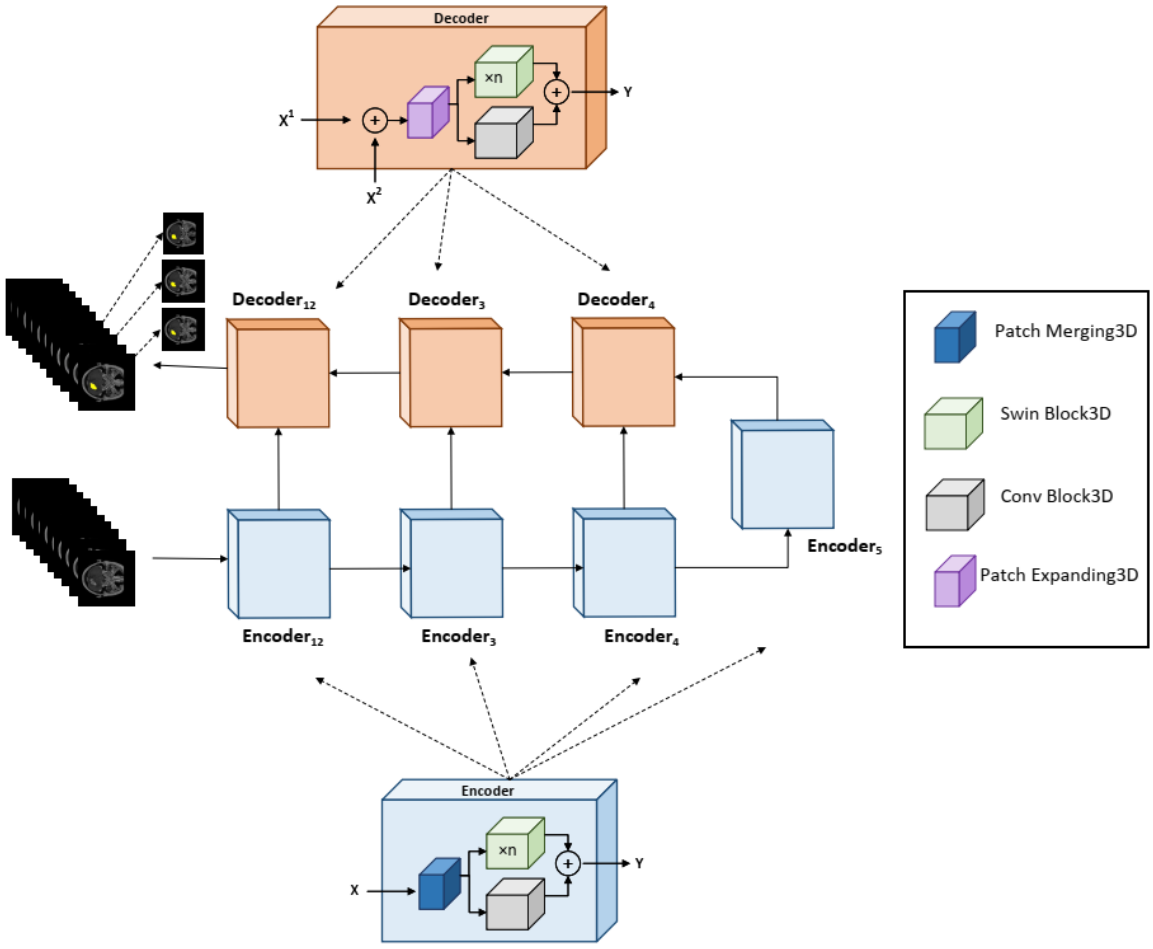

Figure 4: Depiction of the Swin-U-Net3D architecture and the blocks of the encoder and decoder path.

A 3D medical image is split into non-overlapping voxel patches by the Swin U-Net 3D model, which are then flattened and encoded as tokens. To extract features from images, these tokens are passed into an encoder that is transformer-based. The encoders collect features from the image, which are then transmitted to the decoder for up-sampling. The decoder recovers the image's spatial resolution and gradually fuses it with the encoders' extracted features to finish the image's semantic segmentation using a jump connection.

3.4. Proposed Methods

In this section, we present three proposed methods to segment metastatic tumors in MRI images. In the first method, we use a 3D U-Net and a 3D-Swin U-Net to extract local and long-range features from the images respectively. Then a logical fusion module concludes the final segmentation area. In the second method, like the first method a U-Net and Swin U-Net extract initial segmentations, and the resultant output of the models are stacked and fed, as an initial guess, to another Swin U-Net3D to draw the final segmentation. The third method is a combination of the first and the second method. In this method, the results yielded by the first method are stacked with the original image and fed to a Swin U-Net3D. In the rest of this section, after reviewing the preprocessing of the images, we explore the methods in detail.

3.4.1. Preprocessing

In the first step of our proposed approach, all three-dimensional MRI volumes in the dataset are resized to a fixed dimension of 224×224×128 voxels, where 128 corresponds to the number of slices. This choice serves two primary purposes. First, it ensures that the model inputs fulfill the multiple-of-32 requirement in each dimension, as stipulated by Swin U-Net3D [20]. Specifically, adhering to multiples of 32 is crucial for the Swin U-Net3D architecture, as it facilitates proper downsampling and upsampling within the model’s encoder-decoder structure. Second, it strikes an effective balance between retaining essential features and preventing excessive memory usage. Min-Max normalization was also applied on the dataset to bring values in the 0 to 1 range.

3.4.2. The First Method

The architecture of the first method contains a U-Net3D for drawing segmentation area based on local features extracted by convolutional operations and a Swin U-Net 3D whose power is to extract long-range features to segment the images. Then the segmentation results of the architectures are fused logically to determine the final segmentation. To do so, we input three-dimensional data into both the U-Net3D model and the Swin U-Net3D model separately. We continue training these models independently until convergence, where further improvement ceases. Subsequently, we individually pass all images to these models and fuse the resulting outputs using logical AND and OR. When employing logical fusion AND, the resultant mask image delineates the region agreed upon by both models. Conversely, in logical fusion OR, each voxel predicted by the models is considered independently, and each voxel is considered a tumor if only one of the models predicts it as such. The framework for the OR operation is depicted in Figure 5 and is formulated in the equation(2):

$$\boldsymbol{O} = \max\left(\boldsymbol{f}_{\boldsymbol{\theta_U}}(\boldsymbol{I}), \boldsymbol{g}_{\boldsymbol{\theta_S}}(\boldsymbol{I})\right) \quad (2)$$

Where O is the output, I is the input, f and g are the U-Net3D and Swin U-Net3D networks and θ_U and θ_S are their respective parameters.

The primary advantage of this method lies in its simplicity and computational efficiency, as it does not require additional training. However, its main limitation is the potential underperformance in handling complex and highly diverse datasets.

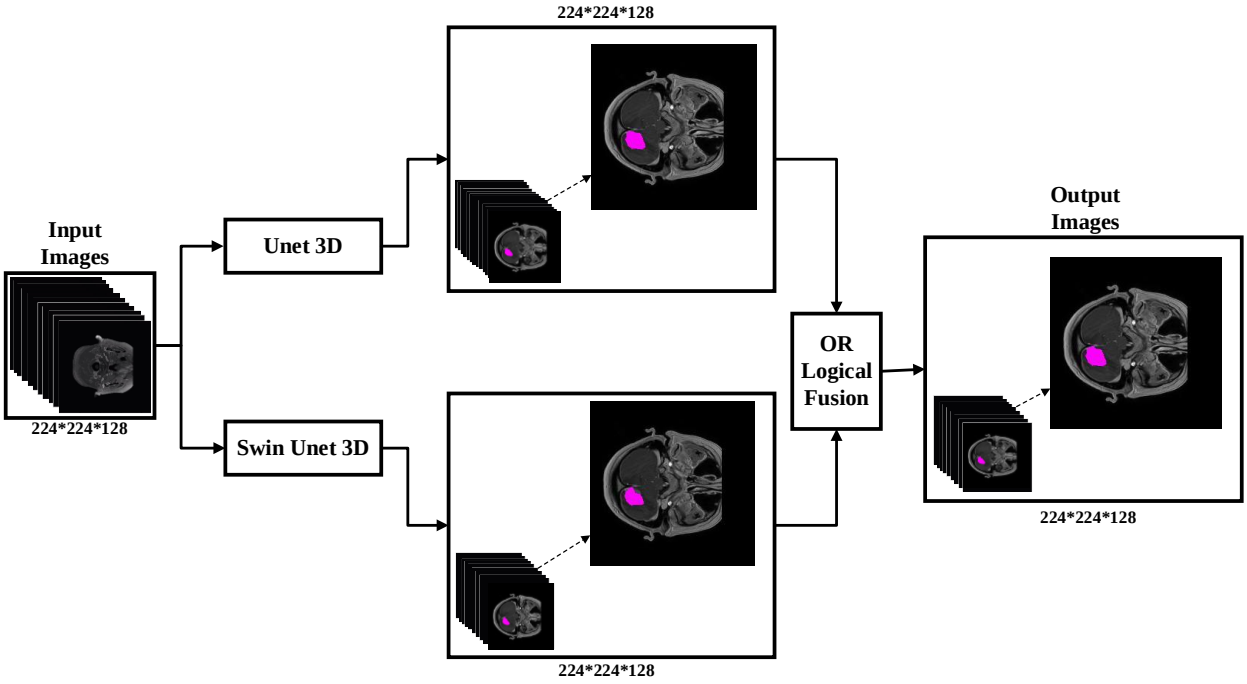

Figure 5: Flowchart of the first method, the OR operation.

3.4.3. The Second Method

Despite the first method where the fusion is a simple logical operation and not trainable; In this method, the segmentation output of the U-Net3D and Swin U-Net3D models are used as an initial guide to the final model which is another Swin U-Net3D.

To do so, we stack the output of the U-Net3D and Swin U-Net3D with the original image to obtain a 3*224*224*128 image. We call this 'stack fusion', and the results are input into a new Swin U-Net3D model. To reduce a tendency for the model to allocate more attention to the mask channels while potentially neglecting the original images, we experimented with random inverting a specific

percentage of pixels of the added channels—0%, 5%, 10%, 15%. This is shown as the 'Inversion' block in Figure 6. The method is formulated as equation(3):

$$\boldsymbol{O} = \boldsymbol{g}_{\boldsymbol{\theta}'_{\boldsymbol{S}}}\left(\left[\boldsymbol{I}, \mathbf{invert}\left(\boldsymbol{f}_{\boldsymbol{\theta}_{\boldsymbol{U}}}(\boldsymbol{I}), \boldsymbol{p}\right), \mathbf{invert}\left(\boldsymbol{g}_{\boldsymbol{\theta}_{\boldsymbol{S}}}(\boldsymbol{I}), \boldsymbol{p}\right)\right]\right) \quad (\mathbf{3})$$

Where O is the output, I is the input, f and g are the U-Net3D and Swin U-Net3D networks and θ_U and θ_S are their respective parameters and the θ'_S are the parameters of the 2nd Swin U-Net3D. p is the inversion percentage and $[\cdot]$ denotes the stacking operation. Algorithm 1 represents the pseudo-code for training with this method.

Algorithm 1 3-Channel Stack Fusion Training

Input: f_{θ_U}, g_{θ_S} $I, mask, p$
Output: $g_{\theta'_S}$
Randomly initialize θ'_S
$M_1 \leftarrow f_{\theta_U}(I)$
$M_2 \leftarrow g_{\theta_S}(I)$
if $p \neq 0$ ***then***
    Randomly invert p% of pixels in M_1 and M_2
end if
$S_{input} \leftarrow \text{Stack}(M_1, M_2, I)$
Train $g_{\theta'_S}$ using gradient descent with input S_{input} and targets mask

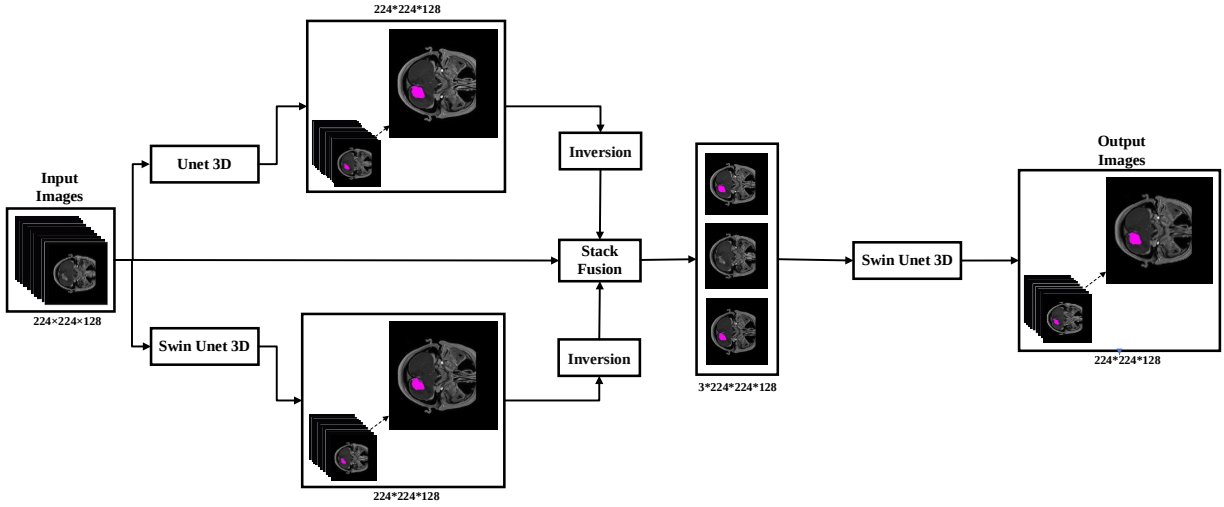

Figure 6: Flowchart of the second method, stacking the results of each network on the original images.

This method excels in improving accuracy by leveraging 'stack fusion,' which enhances the interaction between the models and the original image. However, it introduces higher computational complexity compared to the first method.

3.4.4. The Third Method

This method is a combination of the first and the second methods. Following the logical OR fusion of the outputs from the two models, we stacked the output with the input image and fed the results to a Swin U-Net3D. Analogous to the prior approach, aimed at bolstering the model's generalization capacity and mitigating the risk of excessive emphasis on particular channels, we invert a portion of voxels of the logical fusion output. The inversion percentage aligns with that utilized in the second method. The choice to utilize the OR operation rather than the AND operation stemmed from the intention to utilize predictions from both models. This ultimate approach is visualized in Figure 7. This method follows this equation:

$$\boldsymbol{O} = \boldsymbol{g}_{\boldsymbol{\theta}'_S}\left(\left[\boldsymbol{I}, \text{invert}\left(\textbf{max}\left(\boldsymbol{f}_{\boldsymbol{\theta}_U}(\boldsymbol{I}), \boldsymbol{g}_{\boldsymbol{\theta}_S}(\boldsymbol{I})\right), \text{p}\right)\right]\right) \quad (4)$$

Where O is the output, I is the input, f and g are the U-Net3D and Swin U-Net3D networks and θ_U and θ_S are their respective parameters and the θ'_S are the parameters of the 2nd Swin U-Net3D.[20] Denotes the stacking operation. Algorithm 2 represents the pseudo-code for training with this method.

Algorithm 2 2-Channel Stack Fusion Training

Input: $f_{\theta_U}, g_{\theta_S}, I, mask, p$
Output: $g_{\theta'_S}$
Randomly initialize θ'_S
$M_1 \leftarrow f_{\theta_U}(I)$
$M_2 \leftarrow g_{\theta_S}(I)$
$M \leftarrow M_1 \bigvee M_2$
if $p \neq 0$ ***then***
    Randomly invert p% of pixels in M
end if
$S_{input} \leftarrow \text{Stack}(M, I)$
Train $g_{\theta'_S}$ using gradient descent with input S_{input} and targets mask

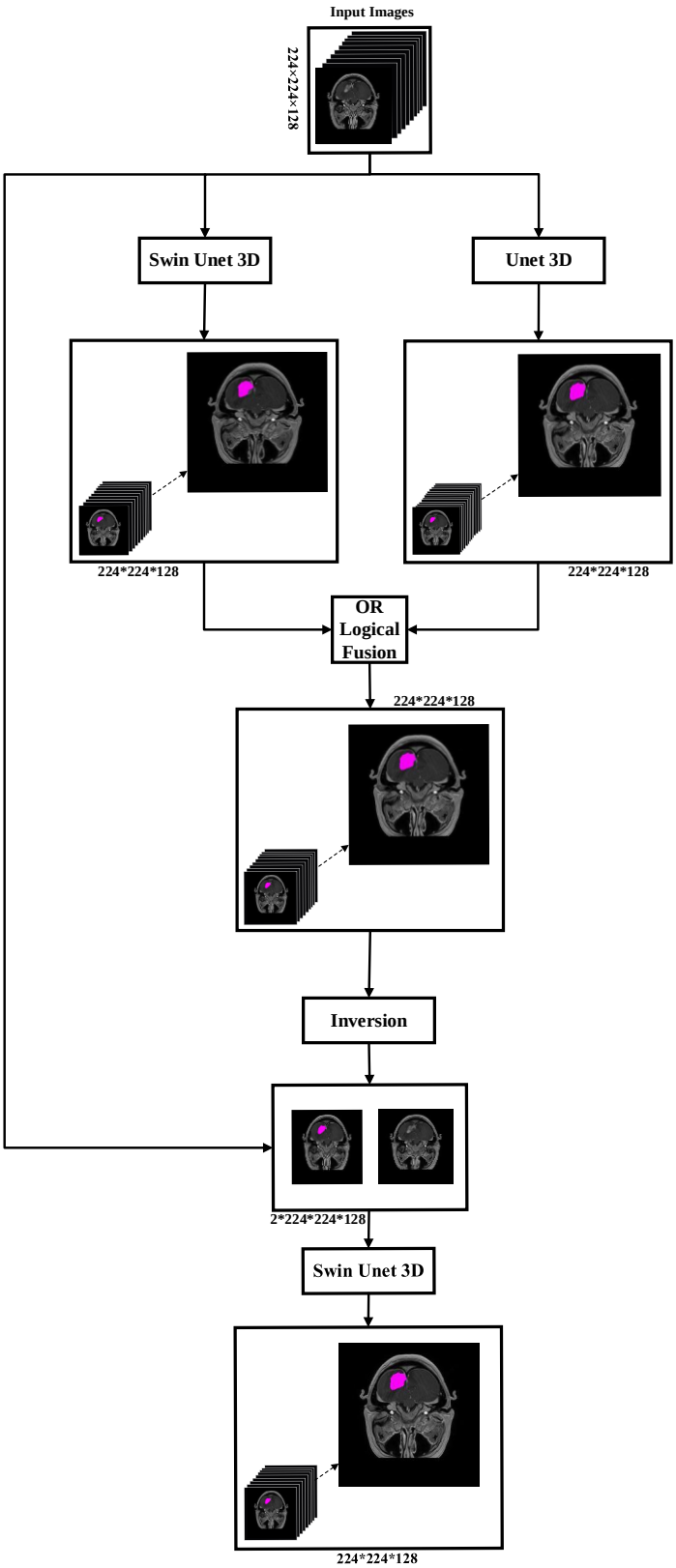

Figure 7: Flowchart of the third method: stacking the OR result of the two networks on the original images.

The third method combines the strengths of the previous two, providing a more robust predictive capability by integrating logical fusion with learning-based strategies. However, it demands more computational resources and precise configuration.

4. Experimental Results

4.1. Setup

The models in this paper were implemented using Python 3.8.1, PyTorch 2.0.1, and MONAI 1.3.1. The system utilized for these implementations is equipped with an RTX4090 graphics card, running the Windows 11 operating system, and powered by an Intel Core i9-13300 CPU with 32 GB RAM and an SSD drive. We used the AdamW. [21] optimizer with parameters: *lr* = 3e-4, *eps*=1e-7 and *weight_decay* =1e-5.

4.2. Dataset

A majority of brain tumor datasets include sequences such as T1, T2, Fluid-Attenuated Inversion Recovery (FLAIR), and Diffusion-Weighted Imaging (DWI). The dataset used in this study is a comprehensive collection of longitudinal MRI studies, focusing on patients diagnosed with Brain Metastasis (BM). The dataset encompasses a total of 75 patients, harboring 260 BM lesions, and includes 637 imaging studies [15]. The dataset comprises contrast-enhanced T1-weighted sequences. The dataset is further enriched by the inclusion of semi-automatic segmentations of 154 different BMs, amounting to a total of 593 T1-weighted segmentations. In addition to the imaging data, the dataset also incorporates extensive clinical data.

This includes patient information, details about the primary tumor, treatment details, and the date of the patient's death. A set of morphological and radiomic-based features, obtained from the segmentations, are also included with the imaging data. The data were collected from five different medical institutions. The inclusion criteria for patients were defined as deceased adult patients with a pathologically confirmed diagnosis of BM between January 1, 2005, and December 31, 2021. The availability of imaging studies with at least the post-contrast T1-weighted high-resolution sequence (pixel spacing $\leq$ 2 mm., slice thickness $\leq$ 2 mm., no gap between slices), absence of noise or artifacts in the images, and availability of basic clinical data (age at diagnosis, sex, treatment schemes followed, survival, etc.) were also part of the inclusion criteria. The dataset is publicly available on the Figshare [22] repository and on a dedicated webpage.

In Figure 8, the distributions of demographics, MRI scanner types, vendors, and tumor types are illustrated. Figure 8. A depicts the gender distribution, with a majority of male patients (61.33%). Figure 8.B shows that the majority of scans were performed on 1.5-T MRI scanners, which accounted for 93.3% of the scans, while higher-field (3.0-T) and lower-field (1.0-T) scanners were used less frequently. Figure 8.C highlights the distribution of MRI vendors, with General Electric being the most commonly used vendor. Figure 8.D presents the distribution of primary tumor types, with non-small cell lung cancer (NSCLC) and breast cancer being the most prevalent.

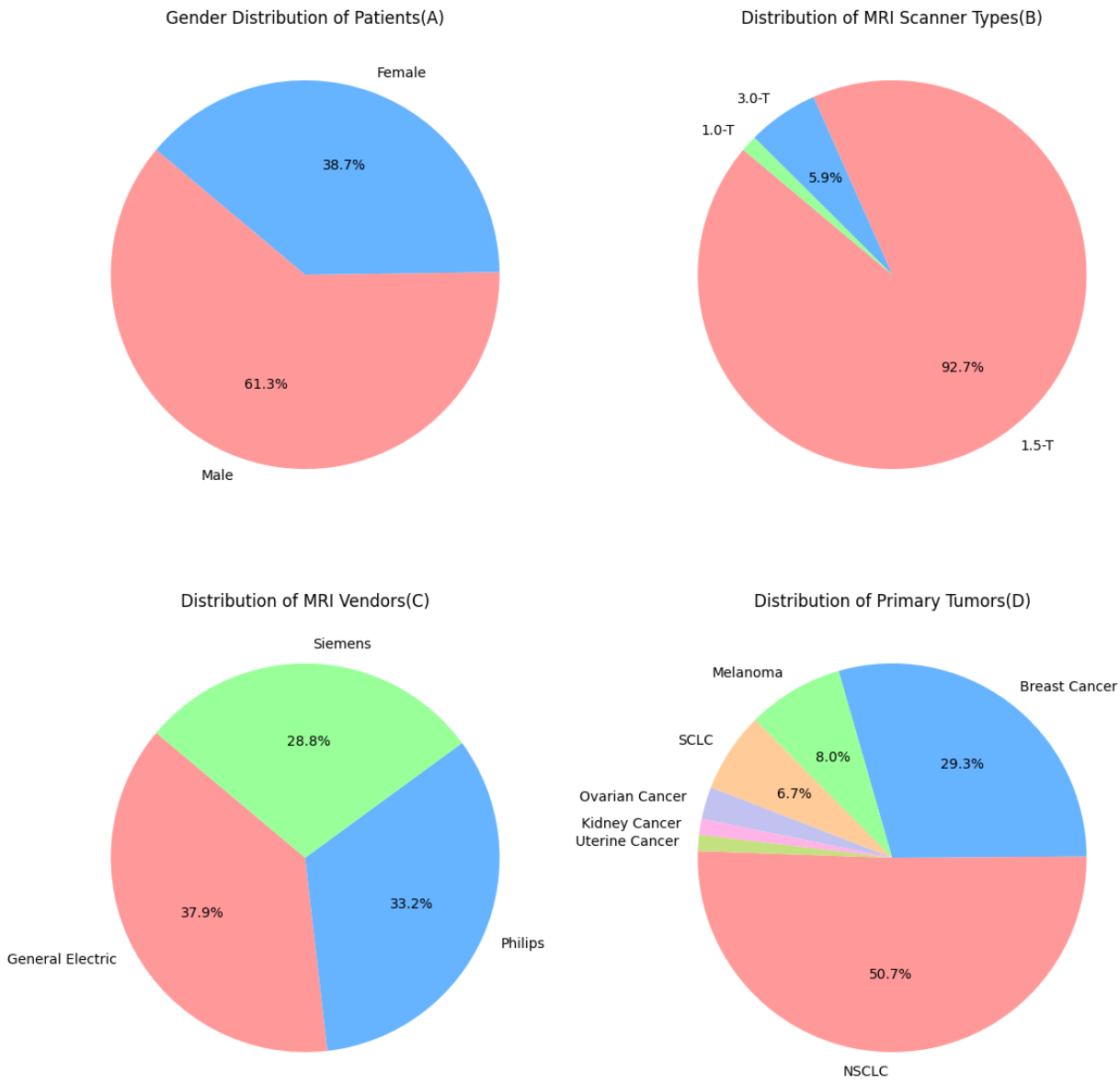

Figure 8: Overview of patient demographics, MRI scanner types, MRI vendors, and primary tumor types. (a) Gender distribution of patients. (b) Distribution of MRI scanner types. (c) MRI vendor distribution. (d) Distribution of primary tumor types.

4.3. Metrics

In this research, we employed a diverse set of criteria to evaluate various aspects of the problem and enhance the accuracy of our results. This section provides an explanation of these criteria. The Dice score is a prevalent metric in domains such as computer vision and medical image segmentation. This measure quantifies the congruence between two binary segmentation masks: the predicted segmentation (prediction) and the annotated segmentation (ground truth). The Dice score is computed as the double intersection area of the two masks, divided by the sum of their

areas. In other words, it is defined as twice the area of overlap between sets A and B, normalized by the sum of the areas of A and B. Mathematically, the Dice score can be expressed as (5) [23]:

$$\boldsymbol{Dice} = \frac{\boldsymbol{2|A \cap B|}}{\boldsymbol{|A| + |B|}} = \frac{\boldsymbol{2TP}}{\boldsymbol{2TP + FP + FN}} \quad \textit{(5)}$$

where TP represents True Positives, FP represents False Positives, and FN represents False Negatives This value ranges from zero (indicating no overlap) to one (indicating perfect agreement). Despite the existence of other metrics, the Dice score is often preferred due to its balance between precision and recall, and its robustness to class imbalance [23].

The Hausdorff distance (HD) quantifies the distance between the surface of the actual area and the predicted area. It is particularly sensitive to the segmented boundary, as defined by the following equation (6):

$$\boldsymbol{d_H(G,S)} = \mathbf{max}[\sup_{\boldsymbol{g \in G}} \inf_{\boldsymbol{s \in S}} \boldsymbol{d(g,s)}, \sup_{\boldsymbol{s \in S}} \inf_{\boldsymbol{g \in G}} \boldsymbol{d(g,s)}] \quad (\mathbf{6})$$

Here, (G) represents the predicted segmentation, (S) denotes the manually segmented output label, and (d) represents the distance metric in the space [12]. We use HD95 which uses the 95th percentile. This is more robust to outliers due to the potential irregular boundaries of tumor predictions.

Additionally, equations (7) and (8) show the formulas of precision and sensitivity:

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}} \quad (\mathbf{7})$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TN}}{\boldsymbol{TN + FP}} \quad (\mathbf{8})$$

4.4. Result

Initially, the Swin U-Net3D model underwent training for 1500 epochs, resulting in a Dice score of 68.29%. Subsequently, the U-Net3D model was trained for the same duration, achieving a Dice score of 64.37%. The training progress of both models is illustrated in Figure 9.

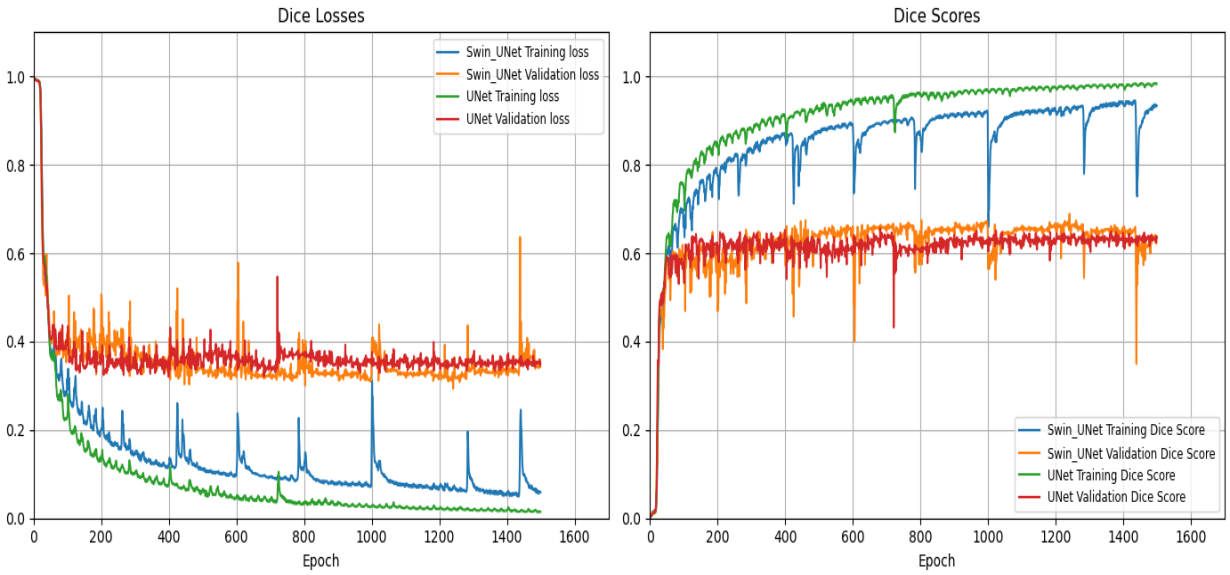

Figure 9: Training curves for Swin U-Net3D and U-Net3D models.

Following this, the fusion method was implemented, utilizing both "and" and "or" operations. The Dice score reached 61.25% with "and" and 70.90% with "or".

Next, the stack fusion method was executed, producing Dice scores of 71.33% with 0% inversion, 69.49% with 5% inversion, 70.28% with 10% inversion, and 70.83% with 15% inversion. The corresponding curves are displayed in Figure 10, with the highest score of 71.33% obtained from stack fusion with 0% inversion.

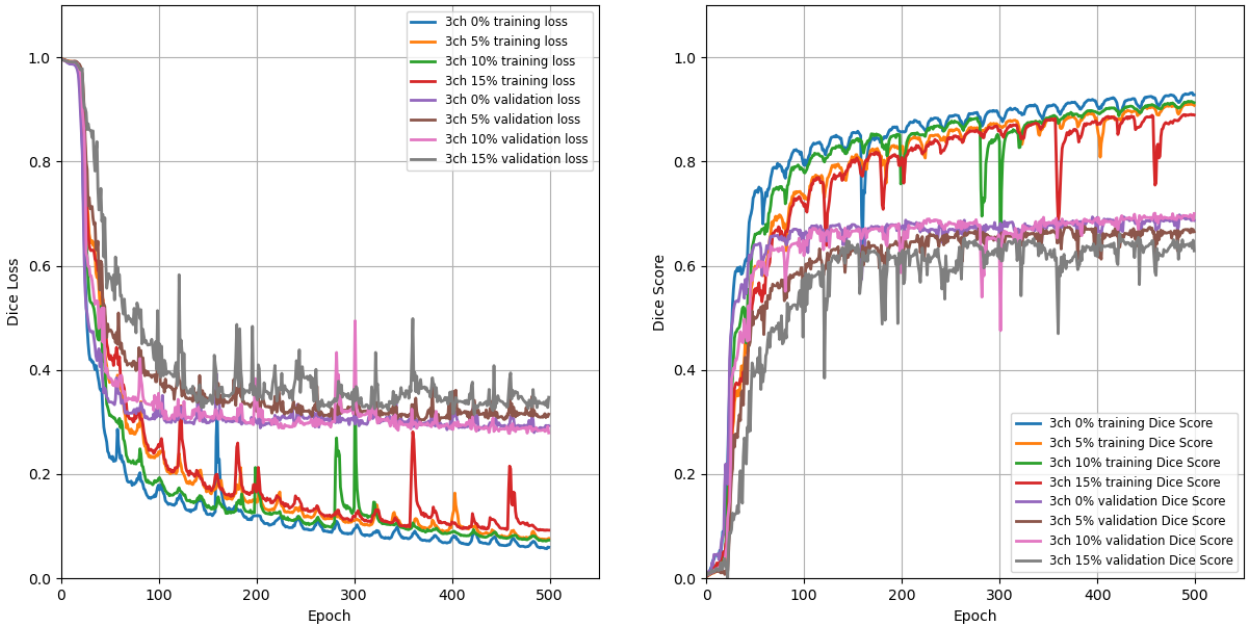

Figure 10: Training curves for Swin U-Net3D in Stack Fusion models.

In the third method, the Dice scores were 73.47% with 0% inversion, 71.34% with 5% inversion, 71.33% with 10% inversion, and 71.42% with 15% inversion. The training progress is depicted in Figure 11, and the impact of mask inversions on the Dice score is presented in Figure 12, with the highest score of 73.47% achieved with 0% inversion. The results are listed in Table 2.

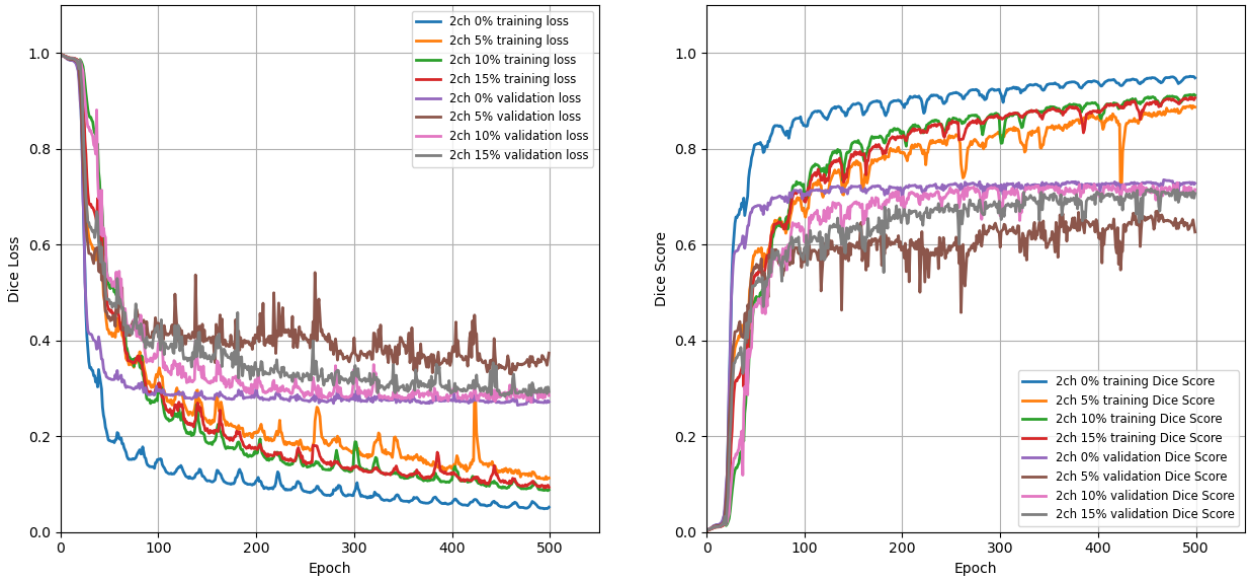

Figure 11: Training curves for Swin U-Net3D in 2-channel Stack Fusion models.

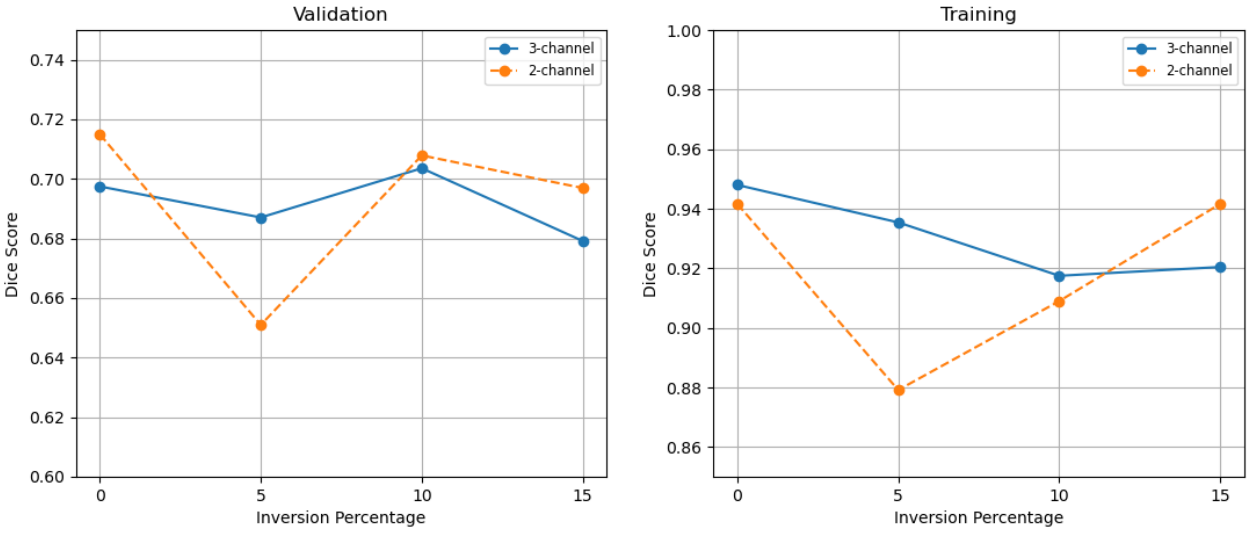

Figure 12: Dice scores of different inversion percentages in the stack fusion methods.

Table 2: result evaluation metrics.

Method	Model	Dice-score	HD95	Sensitivity	Precision
	Swin U-Net3D	68.29 ± 2.35	22.72 ± 2.90	66.93 ± 2.95	76.86 ± 2.07
	U-Net3D	64.37 ± 3.33	22.73 ± 2.92	60.86 ± 4.37	79.02 ± 2.43
1	**AND Logical Fusion**	61.25 ± 3.30	23.33 ± 2.48	55.20 ± 3.84	$\mathbf{83.11 \pm 2.10}$
	OR Logical Fusion	70.90 ± 2.49	22.85 ± 3.21	72.60 ± 3.24	74.17 ± 2.14
2	**Stack Fusion 3 Channel 0% inverse**	71.33 ± 1.90	27.51 ± 4.27	72.80 ± 1.73	74.58 ± 3.83
	Stack Fusion 3 Channel 5% inverse	69.49 ± 1.40	24.16 ± 4.44	68.01 ± 1.63	79.75 ± 3.94
	Stack Fusion 3 Channel 10% inverse	70.28 ± 3.99	23.26 ± 5.13	69.93 ± 4.47	77.00 ± 2.28
	Stack Fusion 3 Channel 15% inverse	70.83 ± 2.59	24.42 ± 2.63	69.68 ± 3.40	77.87 ± 2.05
3	**Stack Fusion 2 Channel 0% inverse**	$\mathbf{73.47 \pm 1.84}$	24.91 ± 3.44	$\mathbf{73.63 \pm 1.53}$	77.22 ± 3.57
	Stack Fusion 2 Channel 5% inverse	71.34 ± 3.95	25.51 ± 4.91	71.38 ± 4.20	76.03 ± 3.57
	Stack Fusion 2 Channel 10% inverse	71.33 ± 3.50	22.69 ± 4.10	72.34 ± 4.25	76.86 ± 2.90
	Stack Fusion 2 Channel 15% inverse	71.42 ± 2.47	$\mathbf{22.04 \pm 2.97}$	71.73 ± 2.90	78.57 ± 2.32

5. Discussion

We have introduced a methodology for integrating a suite of 3D convolutional networks using various fusion techniques to yield an enhanced model. We have validated the efficacy of our proposed approach. By evaluating diverse models and combining them, particularly given their three-dimensional nature, we gain deeper insights into image examination compared to traditional two-dimensional models, thereby enabling better analysis of spatial characteristics.

During the training process of the models, we observed instances where they yielded a significantly low Dice score or even zero percent for certain images. Upon closer examination, we identified that in some samples, one or more tumors were not annotated in the Ground Truth masks provided with the dataset, as illustrated in Figure 13. It appears that these tumors were erroneously labeled. Despite the model accurately learning the tumors from correctly annotated training samples and segmenting them accordingly, the Dice score for these cases remained at zero. In

future work, automated outlier detection methods could be incorporated during training to identify and address such annotation errors early in the process. Additionally, incorporating uncertainty estimation techniques could help improve the robustness of predictions on mislabeled data. It is worth noting that each epoch time was around 4 minutes to 4 minutes and 30 seconds, achieved under the same hardware configurations outlined in the Setup section.

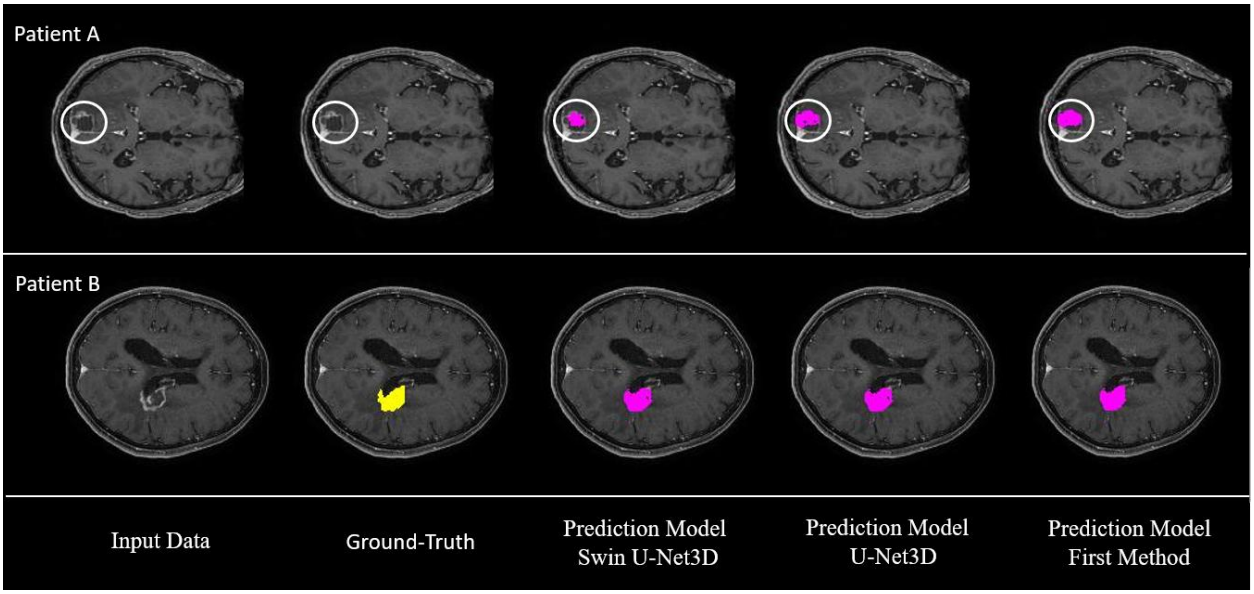

Figure 13: segmentation examples of the dataset. The first row depicts a scenario where the tumor was missed and not labeled in the ground truth.

The results obtained from our study demonstrate that the third method, which integrates logical fusion with additional learning via Swin U-Net3D, achieves the best overall performance. This approach combines the strengths of the first two methods, leveraging the robustness of logical OR fusion and the advanced learning capabilities of the Swin U-Net3D architecture. In this context, the inversion strategy did not yield significant improvements. Nevertheless, including an additional training step, coupled with stacking the fused output with the original image, further enriches the input data, allowing Swin U-Net3D to refine predictions more effectively. Notably, this method yielded the highest Dice score of 73.47%, significantly improving the accurate identification of critical regions such as brain metastases.

Comparatively, the first method demonstrated the advantage of simplicity and reduced computational requirements but fell short of leveraging the full predictive power of the combined models. The second method, despite utilizing multi-channel inputs, did not surpass the third method in terms of accuracy, likely due to the Swin U-Net3D's inability to fully exploit the additional input channels. The logical OR operation plays a pivotal role in the success of the third method, ensuring that predictions from both models contribute to the final output. This synergy effectively addresses the challenge of capturing small and localized brain metastases. However, testing the model on other datasets revealed potential limitations, particularly when encountering tumors with larger sizes, which may demand further adaptation of the methodology.

To address this limitation, future research could explore adaptive model scaling techniques or data augmentation strategies tailored to handle variations in tumor sizes across datasets. For instance, implementing 'tumor size-normalization' involves rescaling the images or tumor regions based on their measured size range, thereby placing all tumors within a standardized dimensional range so that the model handles them more consistently during training. Such preprocessing or multi-scale training approaches could allow the models to generalize better to diverse datasets.

The third method provides a balanced trade-off between computational complexity and predictive performance. Its robust design and adaptability underscore its potential for broader applications, particularly in domains requiring precise segmentation of small, localized regions. This disparity highlights a notable difference between brain metastases and primary brain tumors. Additionally, the scarcity of data on brain metastases poses a significant challenge, necessitating the use of data from various sources, including common datasets reviewed in the literature. Although data availability remains a challenge, the quality of data used for training purposes is satisfactory.

6. Conclusion and Future Work:

This study demonstrated that combining 3D U-Net and Swin U-Net3D architectures enhances the segmentation of metastatic brain tumors in MRI, particularly for small and scattered lesions. Despite these advances, issues such as mislabeled data, variations in imaging protocols, and limited availability of annotated datasets continue to pose challenges.

Future improvements can focus on mitigating these limitations. Automated outlier detection and uncertainty estimation can refine the training process by identifying and excluding mislabeled or unreliable data, reducing errors in model predictions. Incorporating multi-modal imaging data, such as T1-CE, T2, and FLAIR, could provide richer contextual information, enabling better differentiation of tumor boundaries and reducing false positives. Leveraging self-supervised or unsupervised learning techniques would make better use of unlabeled data, extracting robust representations that improve performance in scenarios with limited labeled examples. Additionally, harmonizing data from different scanners and employing adaptive, multi-scale model architectures can address variations in tumor sizes and imaging protocols, ensuring better generalizability and reliability across diverse datasets and clinical settings. These enhancements would further optimize the framework for real-world applications in tumor detection and treatment planning.

References

[1] "Board pate.advances in brain and spinal cord tumor research: https://www.cancer.gov/types/brain/research."

[2] "Board PATE. Adult central nervous system tumors treatment (PDQ®): Health Professional Version. Website. 2022. https://www.cancer.gov/types/brain/hp/adult-brain-treatment-pdq.."

[3] "https://www.mayoclinic.org/tests-procedures/mri/about/pac-20384768."

[4] K. Bhargavi and S. Jyothi, "A survey on threshold based segmentation technique in image processing," *International Journal of Innovative Research and Development*, vol. 3, no. 12, pp. 234-239, 2014.

[5] R. Archana and P. E. Jeevaraj, "Deep learning models for digital image processing: a review," *Artificial Intelligence Review*, vol. 57, no. 1, p. 11, 2024.
[6] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243-1267, 2022.
[7] H. Seo *et al.*, "Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications," *Medical physics*, vol. 47, no. 5, pp. e148-e167, 2020.
[8] J. Xue *et al.*, "Deep learning–based detection and segmentation-assisted management of brain metastases," *Neuro-oncology*, vol. 22, no. 4, pp. 505-514, 2020.
[9] E. Grøvik, D. Yi, M. Iv, E. Tong, D. Rubin, and G. Zaharchuk, "Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI," *Journal of Magnetic Resonance Imaging*, vol. 51, no. 1, pp. 175-182, 2020.
[10] S.-Y. Hu *et al.*, "Multimodal volume-aware detection and segmentation for brain metastases radiosurgery," in *Artificial Intelligence in Radiation Therapy: First International Workshop, AIRT 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 1*, 2019: Springer, pp. 61-69.
[11] S. K. Yoo *et al.*, "Deep-learning-based automatic detection and segmentation of brain metastases with small volume for stereotactic ablative radiotherapy," *Cancers*, vol. 14, no. 10, p. 2555, 2022.
[12] I. Aboussaleh, J. Riffi, K. el Fazazy, A. M. Mahraz, and H. Tairi, "3DUV-NetR+: A 3D hybrid Semantic Architecture using Transformers for Brain Tumor Segmentation with MultiModal MR Images," *Results in Engineering*, p. 101892, 2024.
[13] G. Zhang, J. Zhou, G. He, and H. Zhu, "Deep fusion of multi-modal features for brain tumor image segmentation," *Heliyon*, vol. 9, no. 8, 2023.
[14] Y. Al Khalil, A. Ayaz, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer, "Multi-modal brain tumor segmentation via conditional synthesis with Fourier domain adaptation," *Computerized Medical Imaging and Graphics*, vol. 112, p. 102332, 2024.
[15] B. Ocaña-Tienda *et al.*, "A comprehensive dataset of annotated brain metastasis MR images with clinical and radiomic data," *Scientific data*, vol. 10, no. 1, p. 208, 2023.
[16] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012-10022.
[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 2015: Springer, pp. 234-241.
[18] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[19] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[20] Y. Cai *et al.*, "Swin Unet3D: a three-dimensional medical image segmentation network combining vision transformer and convolution," *BMC medical informatics and decision making*, vol. 23, no. 1, p. 33, 2023.
[21] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[22] "Tienda, Beatriz Ocaña; Pérez-Beteta, Julián; Romero-Rosales, José; Molina-García, David; Suter, Yannick; Asenjo, Beatriz; et al. (2023). A comprehensive dataset of annotated brain metastasis images with clinical and radiomic data. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.6194104.v1."
[23] M. Brusco, J. D. Cradit, and D. Steinley, "A comparison of 71 binary similarity coefficients: The effect of base rates," *Plos one*, vol. 16, no. 4, p. e0247751, 2021.