Transfer Learning Approaches for Brain Metastases Screenings

Minh Sao Khue Luu *, Bair N. Tuchinov, Victor Suvorov, Roman M. Kenzhin, Evgeniya V. Amelina, Andrey Yu. Letyagin

Posted Date: 24 October 2024

doi: 10.20944/preprints202410.1738.v1

Keywords: transfer learning; brain metastases; segmentation

Article

Transfer Learning Approaches for Brain Metastases Screenings

Minh Sao Khue Luu [1,*], Bair N. Tuchinov [1], Victor Suvorov [1], Roman M. Kenzhin [1], Evgeniya V. Amelina [1,3] and Andrey Yu. Letyagin [1,2]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Research Institute of Clinical and Experimental Lymphology, Branch of IC&G SB RAS, Novosibirsk, Russia
[3] FSBI Federal Neurosurgical Center, Novosibirsk, Russia
* Correspondence: khue.luu@g.nsu.ru

Abstract: Brain metastasis is a severe and complicated cancer that requires timely screening and follow-up approaches using precise segmentation techniques for effective surgical intervention, radiation therapy, or monitoring the disease progression. In this study, we explore the efficacy of transfer learning in improving the automatic segmentation of brain metastases on Magnetic Resonance Imaging scans with the subsequent possibility of using it in clinical practice for preventive examinations and remote diagnostics. We train three deep learning models on a public dataset from the Brain Tumor Segmentation 2024 Challenge, then fine-tune these pretrained models on a small private dataset and compare them to models trained on private data from scratch. The results indicate that models utilizing transfer learning achieve superior accuracy and generalization in segmenting brain metastases. We also observed that the custom loss function significantly enhances performance compared to the default configuration. This study highlights the importance of leveraging transfer learning in medical imaging to address challenges associated with small, specialized datasets.

Keywords: transfer learning; brain metastases; segmentation

1. Introduction

Brain metastasis (BM) happens when cancer cells spread to the brain, creating secondary tumors that interfere with normal brain function [1,2]. These tumors usually indicate an advanced stage and high malignancy of cancer, making treatment more difficult and lowering the chances of long-term survival. BM is serious because even small brain tumors can be deadly by pressing inside the skull or causing serious problems if they are in critical areas [3]. Survival rates are generally low, especially without early detection and intervention. Even with treatment, the prognosis can vary depending on tumor characteristics (number and type of tumors), patient factors (age), and treatment options (surgery, radiation, and chemotherapy) [4].

Diagnosis and treatment of BMs often requires advanced imaging techniques like magnetic resonance imaging (MRI) to identify the size, location, and number of tumors [5]. MRI is especially useful because it provides clear and detailed images with excellent soft tissue contrast, allowing doctors to detect small tumors, assess swelling around them, identify perinodular bleeding within the brain, and understand how they affect important parts of the brain. Moreover, MRI does not use ionizing radiation, making it safer for patients who need multiple scans [6]. Common MRI sequences, such as T1-weighted and T2-weighted scans, reveal the brain's structure, while FLAIR (Fluid-Attenuated Inversion Recovery) highlights areas of swelling and lesions near cerebrospinal fluid spaces. Advanced MRI techniques like diffusion-weighted imaging and perfusion imaging offer additional insights into tumor characteristics, such as cellularity and blood flow [6]. By using these sequences, MRI helps distinguish between healthy and abnormal tissues, guiding early personalized diagnosis and effective treatment planning for brain metastases. Using AI to automatically measure

 doi:10.20944/preprints202410.1738.v1

and track BM volumes following SRS treatment, this study showed a strong correlation between AI-driven measurements and the current clinically used method: manual axial diameter measurements [7].

Precise segmentation of BMs helps doctors accurately assess tumor size, location, and spread, which is essential improving of the results for surgery and radiation therapy [8]. By clearly defining tumor boundaries, treatments can be targeted more effectively while reducing harm to healthy brain tissue. Segmentation also allows doctors to track tumor changes over time, helping to evaluate how well treatments are working. However, manual segmentation is time-consuming, prone to errors, and can vary between doctors, leading to inconsistent results.

Deep learning (DL) [9] has made significant strides in medical images analysis by enabling automated segmentation of brain MRI [10–14]. A recent review demonstrated the role of machine learning and deep learning methods in lesion detection, diagnosis, and anatomical segmentation of various types of brain tumors, including metastases [15]. Building on these advancements, several novel approaches have been proposed to further improve the accuracy and reliability of brain metastasis segmentation. Grøvik *et al.* used a fully convolutional neural network based on a 2.5D GoogLeNet architecture to automatically detect and segment brain metastases from multisequence MRI, achieving an overall AUC of 0.98 and a Dice score of 0.79, with an average false-positive rate of 8.3 per patient [16]. Huang *et al.* proposed the DeepMedic+ network with a custom volume-level sensitivity-specificity loss, which significantly improved brain metastasis detection by increasing sensitivity from 0.853 to 0.975 and precision from 0.691 to 0.987, while reducing false positives by 0.444 [17]. Highlighting the importance of strategic modality selection and multi-stage processing, [18] proposed a two-stage detection and segmentation model using T1c, T1, and FLAIR modalities, which significantly improved brain metastasis segmentation accuracy compared to single-pass models. Yang *et al.* trained the 3D-TransUNet model for brain metastases segmentation, exploring both Encoder-only and Decoder-only configurations with Transformer self-attention, and found that the Encoder-only version with Masked-Autoencoder pre-training achieved a lesion-wise Dice score of 59.8% [19]. In addition to these model-based improvements, new tools have been developed to bring advanced segmentation methods into clinical practice. An open-source software, Raidionics, was also developed to perform preoperative segmentation of major brain tumor types, allowing standardized clinical reports to be generated in about ten minutes. It achieved an average Dice of 0.85 with 0.95 recall and precision for preoperative segmentation, while postoperative performance was lower, with an average Dice of 0.41 [20].

Despite these advances, DL faces challenges, particularly the limited availability of high-quality annotated medical images. Transfer learning (TL) [21] can overcome these limitations by allowing models trained on large datasets to be adapted for specific tasks with less training data, based on the idea that many basic features are shared across different image types. This includes various techniques like instance-based, network-based, and adversarial-based transfer [22]. TL has been applied to a wide range of medical imaging tasks—such as segmentation, object identification, disease classification, and severity grading [23].

Several recent studies have introduced advanced deep learning techniques for brain tumor segmentation, using pre-trained models and improved architectures to boost performance. Wacker *et al.* enhanced the AlbuNet architecture for brain tumor segmentation using a 3D U-Net-based model with a ResNet34 encoder, pretrained on ImageNet, and extended convolutional layers to process volumetric MRI data. The 3D model with pretraining achieved higher Dice scores and more stable training than the 2D model, but its performance was less consistent on a new clinical dataset due to data differences [24]. Tataei Sarshar *et al.* introduced a deep learning pipeline for brain tumor segmentation using a pretrained ResNet50 model with a multi-modality approach (T1, T1-c, T2, FLAIR), combined with cascade feature extraction, inception, and new mReLU blocks to enhance learning. The proposed method achieved mean Dice scores of 0.9211 for the tumor core, 0.8993 for the whole tumor, and 0.9223 for the enhancing region [25].

Further extending these innovations, recent study by Messaoudi *et al.* [26] proposed embedding pre-trained 2D networks into higher-dimensional U-Nets for effective segmentation of 2D and 3D

 doi:10.20944/preprints202410.1738.v1

medical images, utilizing weight and dimensional transfer techniques, and demonstrated superior performance in benchmarks. Huang *et al.* presented a federated learning framework using Learning Without Forgetting to train deep learning models for brain metastasis segmentation across multiple medical centers without sharing raw data, achieving strong performance on diverse MRI datasets [27]. Pani and Chawla introduced a hybrid approach that integrates transfer learning and self-supervised learning within a 3D UNET architecture to improve brain tumor segmentation in MRI scans, effectively reducing the need for extensive annotated data while achieving high accuracy with a Dice score of 90.15 [28].

While deep learning models have shown promise in automatically segmenting brain metastases, their application in clinical settings is often limited by a lack of alignment with radiologist expertise. Most existing studies evaluate model performance based solely on quantitative metrics but do not considering feedback from medical experts, who are key to judging how useful and accurate the results are in practice. Our research addresses this by including radiologists and physicians feedback in evaluating fine-tuned models, ensuring the results are both accurate and practical for clinical use.

2. Materials and Methods

2.1. Overall Process

We started by training three deep learning models on a public dataset of brain metastases MRI, using 5-fold cross-validation. The private dataset with smaller number of images was divided into a training set (26 images) and a testing set (10 images). We fine-tuned the selected models on the private training set using pre-trained weights on the public dataset and also trained them from scratch on the private data to compare performance. Both the fine-tuned and scratch-trained models were tested on 10 unseen sample (a test set). We employed explainable techniques to visualize the segmentations and gathered detailed feedback from medical experts on the results, focusing on their clinical relevance and accuracy. This helped us assess how well the models performed in segmenting brain metastases after fine-tuning compared to training from scratch.

2.2. Data Sources and Preparation

In this study, we employed two distinct datasets. The first dataset, ASNR-MICCAI Brain Metastasis Challenge 2024 (BraTS) [29], included MRI scans of untreated brain metastases collected from multiple institutions under standard clinical conditions. The dataset consisted of pre-contrast T1-weighted (T1W), post-contrast T1-weighted (T1C), T2-weighted (T2W), and T2-weighted Fluid Attenuated Inversion Recovery (FLAIR) sequences. Initial segmentation was performed using automated algorithms, followed by manual refinement by neuroradiologists and final approval by board-certified specialists. The BraTS dataset utilized a three-label segmentation system, which included non-enhancing tumor core (NETC), surrounding FLAIR hyperintensity (SNFH), and enhancing tumor (ET).

The second dataset, the Siberian Brain Tumors (SBT), was sourced from the Federal Neurosurgical Center in Novosibirsk, Russia. This dataset included clinical data and MRI scans (T1W, T1C, T2W, FLAIR) from 496 patients, covering a range of diagnoses such as glioblastomas, astrocytomas, neurinomas, meningiomas, and metastases. For this study, only the metastasis cases (36 in total) were utilized. MRI scans were predominantly acquired using a 1.5T Siemens Magnetom Avanto machine, with some cases imaged at 3T Philips Ingenia system. Annotations for this dataset included peritumoral edema, non-enhancing tumor regions, GD-enhancing tumor regions, and necrotic tumor cores. These annotations were manually conducted by two board-certified neuroradiologists following a strict protocol, with an independent expert reviewing and correcting any errors. Ground truth labels were cross verified with patient clinical histories, as well as histological and immunohistochemical data. All cases were formatted in NIFTI, with 192 slices of 512×432 pixels, co-registered to a standard template and resampled to a 1x0.5x0.5 mm^3 resolution. The SBT data were divided into 26 training cases and 10 test cases.

 doi:10.20944/preprints202410.1738.v1

The key feature is the availability of information about the source of the oncological process and medical imaging of metastases in the brain.

All models used the default nnU-Net preprocessing to ensure the input data was prepared consistently and effectively. This preprocessing included adaptive resampling, which adjusts depending on the dataset's characteristics, and dynamic intensity normalization, which changes the normalization method based on data variability. Foreground cropping was improved to better focus on the important areas, reducing extra computation. Padding and patch size were made flexible, adjusting to the GPU memory for maximum efficiency. Adaptive patch spacing was used to optimize the resolution during training, balancing efficiency and detail. Lastly, dynamic data sampling helped manage class imbalance, especially useful for datasets with different tumor sizes.

2.3. Model Architecture and Configuration

We employed the nnU-Net framework [30] for all model implementations, leveraging its self-configuring U-Net architecture. nnU-Net was chosen because of its ability to adapt automatically to different datasets by optimizing key architectural parameters such as depth, number of layers, and patch sizes, which are crucial in medical image segmentation tasks where data variability is high. The encoder-decoder structure, along with skip connections, ensures that spatial information is preserved, which is critical for precise tumor boundary detection.

For this study, we employed three models: (1) the default nnU-Net configuration (Default), (2) nnU-Net with a custom combined loss function (TverskyBCE), and (3) SegResNet [31].

The **Default** model was used without modifications, dynamically adapting to the dataset by optimizing patch size, input modalities, and network depth. Its self-adjusting capability allowed it to achieve optimal performance across different datasets, justifying its use as a baseline. The initial learning rate was set to 1e-2, and the optimizer employed was stochastic gradient descent (SGD) with Nesterov momentum (momentum = 0.99). A polynomial learning rate scheduler was applied to gradually reduce the learning rate as training progressed, ensuring smooth convergence and preventing overshooting.

The **TverskyBCE** configuration modified the nnU-Net architecture by employing a combined loss function that merges Tversky loss [32] with Binary Cross-Entropy (BCE) loss. Tversky loss, designed to handle class imbalance, prioritized false negative reduction—critical for detecting small lesions. The Tversky loss parameters were set at $\alpha = 0.3$ and $\beta = 0.7$, emphasizing recall over precision. To further enhance sensitivity to small tumors, BCE loss was applied with a positive class weight (pos_w = 10), focusing more on correctly identifying tumor regions. The two losses were combined at a 1:1 ratio to optimize both small tumor detection and overall segmentation quality. The optimizer and learning rate scheduler used for the TverskyBCE model were identical to those of the Default model.

SegResNet was implemented with a U-Net-like architecture, enhanced by residual connections to improve gradient flow and prevent vanishing gradients, ensuring stability during the training of deeper networks. The encoder consisted of 4 stages, with residual blocks (1, 2, 2, 4) capturing progressively complex features, while the decoder used 1 residual block per stage for reconstruction. The model started with 32 filters, which increased in deeper layers to capture finer details. Input and output channels were configured based on the number of input modalities and segmentation classes. The Adam optimizer with a learning rate of 1e-4 and weight decay of 1e-5 was used to ensure efficient parameter updates. A polynomial learning rate scheduler was employed to gradually reduce the learning rate, preventing overfitting and ensuring smooth convergence.

2.4. Training and Fine-Tuning

All models were trained using GPU acceleration with the 3d_fullres configuration from the nnU-Net framework. Training spanned 200 epochs for pre-training and 100 epochs for fine-tuning. To enhance generalization and robustness, data augmentation included spatial transformations (such as rotations and scaling), intensity adjustments (including brightness, contrast, and gamma

 doi:10.20944/preprints202410.1738.v1

modifications), noise injections (e.g., Gaussian noise, Gaussian blur), and techniques like low-resolution simulation and mirroring.

For pre-training, models were trained on the BraTS dataset using training plans aligned with those of the SBT dataset, ensuring consistency and facilitating smoother transfer learning. During fine-tuning, models initialized with pre-trained weights from BraTS were further trained on the SBT dataset. Additionally, we trained models on the SBT dataset from scratch for 100 epochs to evaluate the benefits of fine-tuning compared to training from scratch.

All training was conducted on three workstations with Quadro RTX 8000, NVIDIA GeForce GTX TITAN X, and NVIDIA RTX A4500 GPUs, using Python programming language.

2.5. Evaluation and Validation

To evaluate the performance of our segmentation models, we used several common metrics in medical image analysis, including Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), Sensitivity (SEN), and Specificity (SPE) [33]. DSC measures overlap between predicted and ground-truth segmentations (0 to 1, higher is better). HD assesses boundary quality (lower is better). SEN and SPE measure the ability to correctly identify true positives and true negatives (both range from 0 to 1, higher is better). We also calculated true and predicted volumes to evaluate how well the model captures lesion size. For testing, we ensembled the predictions from all fold models and used a separate hold-out set of 10 images for an unbiased evaluation.

For the quantitative evaluation, we conducted a comparative analysis of the models using descriptive statistics, including the mean, standard deviation, and 95% confidence intervals. These statistical measures provided insights into the average performance, consistency, and reliability of the fine-tuned models relative to the scratch-trained models. To determine whether the data met the assumptions for parametric testing, we first applied the Shapiro-Wilk test [34] to assess whether the distribution of metrics for all models followed a normal distribution. Based on the results of the normality test, we observed that some of the metrics did not follow a normal distribution. As a result, we employed a one-sided Mann-Whitney U test [35], a non-parametric statistical test, to evaluate whether the performance scores of the fine-tuned models were stochastically greater than those of the scratch-trained models. The alternative hypothesis (H_1) posited that the fine-tuned models would outperform the scratch-trained models in terms of evaluation metrics. We reject the H_1 if the p-value from the Mann-Whitney U test is greater than or equal to the significance level of 0.05, indicating that there is insufficient evidence to support that the fine-tuned models outperform the scratch-trained models.

For the qualitative evaluation, we selected the two best models from our quantitative analysis and sent their predictions on 10 test cases to our medical experts (radiologists and physicians). The segmentation results were evaluated by medical experts using 3D Slicer [36] software to visually assess the quality and accuracy of the predicted tumor boundaries. After their independent assessment, a follow-up meeting was held to discuss the models' strengths and weaknesses, particularly in handling complex tumor characteristics. This discussion helped identify areas for improvement to enhance clinical use.

2.6. Data and Code Availability

The BraTS dataset presented in the study are openly available at https://www.synapse.org/Synapse:syn59059764.

The SBT dataset presented in this article is not readily available because it contains private health data collected in our Federal neurosurgical center and sharing it publicly would violate patient privacy and confidentiality agreements in accordance with ethical and regulatory standards.

The code to reproduce this study is available at https://github.com/luumsk/BrainMetaSeg.git.

2.7. Manuscript Preparation

 doi:10.20944/preprints202410.1738.v1

An AI-assisted tool, ChatGPT (OpenAI), was used to assist in refining the text of the manuscript. The tool was employed exclusively to improve language clarity and grammar, without altering the scientific content or interpretations provided by the authors.

3. Results

3.1. Quantitative Analysis of Model Performance

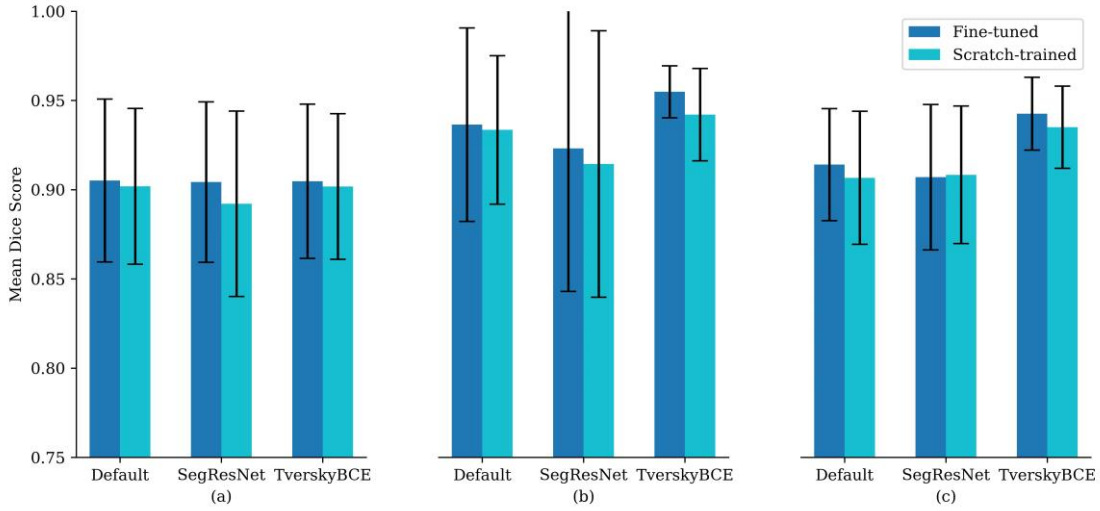
3.1.1. Descriptive Statistics

Our analysis showed that fine-tuned models generally performed better than scratch-trained models in key areas such as segmentation accuracy, sensitivity, and boundary localization. Fine-tuned models achieved higher DSC, lower HD for enhancing tumor and tumor core, and better SEN. They also demonstrated more consistent and reliable predictions, indicated by narrower confidence intervals and lower standard deviations. Although scratch-trained models had a slight advantage in specificity and boundary delineation for the whole tumor, fine-tuning proved to be more effective in the most important aspects of segmentation.

Specifically, fine-tuned models had higher DSC across all tumor regions: 0.905 compared to 0.902 for enhancing tumor, 0.936 compared to 0.934 for tumor core, and 0.914 compared to 0.907 for whole tumor. Figure 1 supports this trend, showing that fine-tuned models consistently achieved higher mean DSC across different configurations, with similar or narrower error bars, suggesting improved accuracy and stability. Although the differences were modest, the results suggest that fine-tuning offers clear benefits for segmentation accuracy and consistency, particularly for enhancing tumor and tumor core regions.

Figure 2 further shows that fine-tuned models generally provided better boundary localization than scratch-trained models, especially for enhancing tumor and tumor core. The fine-tuned model had a mean HD of 2.76 mm for enhancing tumor and 1.87 mm for tumor core, compared to 2.78 mm and 5.50 mm, respectively, for the scratch-trained model. For the whole tumor, scratch-trained models performed slightly better, with a mean HD of 3.23 mm versus 3.36 mm for fine-tuned models. Despite this, fine-tuned models showed lower variability for enhancing tumor and tumor core, indicating more consistent results.

Among the transfer learning approach with fine-tuned model variants, TverskyBCE demonstrated the best overall performance for brain metastasis segmentation. It achieved higher DSC and superior boundary localization compared to the Default model, showcasing a clear advantage in segmentation quality. While SegResNet also performed well, it lacked consistency, particularly in capturing fine boundary details. TverskyBCE offered the optimal balance of accuracy and stability, making it the preferred choice for enhancing segmentation quality in this context.



 doi:10.20944/preprints202410.1738.v1

Figure 1. Comparison of mean Dice score between fine-tuned and scratch-trained models on three brain tumor sub-regions: (a) enhancing tumor, (b) tumor core, (c) whole tumor. The error bars indicate the 95% confidence interval.

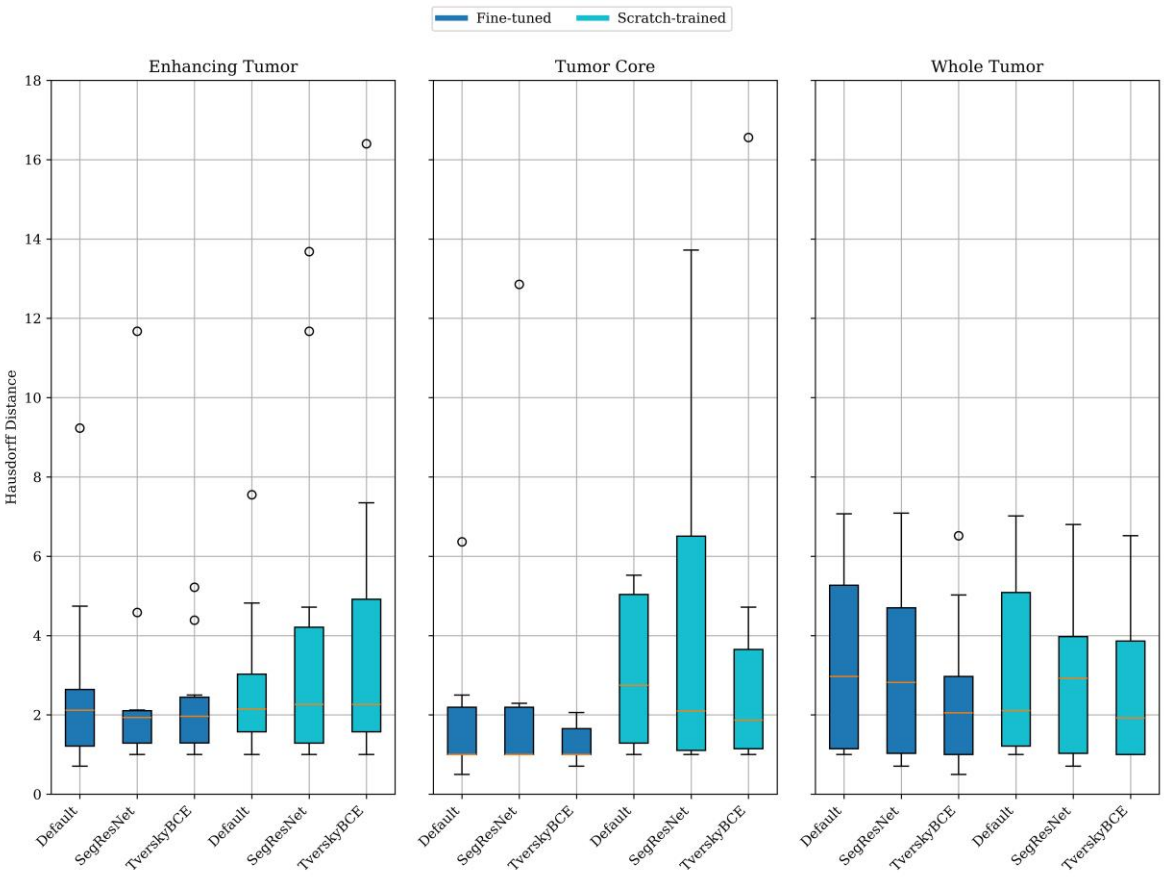


Figure 2. Comparison of mean Hausdorff distance between fine-tuned and scratch-trained models on three brain tumor sub-regions.

3.1.2. Statistical Testing for Significance

The statistical test results indicated no significant difference between the fine-tuned and scratch-trained models across all metrics. However, it is important to note that the sample size in this study is relatively small, consisting of only 10 cases approved by the histological and immunohistochemical evaluations. This small sample size reduces the statistical power of the test, increasing the likelihood of failing to detect real performance differences that may exist between the models. Despite the lack of statistical significance, the fine-tuned models consistently showed slightly higher performance across all metrics, with improvements in DSC, lower HD, and better SEN and SPE. Also, all experiments were focused to evaluating a clinical significance and application some explainable techniques. These slight improvements suggest that fine-tuning does contribute positively to segmentation performance, particularly in complex tasks such as brain tumor segmentation. In practical applications, where even small improvements in accuracy and boundary precision can be clinically meaningful, these results are still relevant and important meaning for screening process.

To further compared model variants across sub-regions, we run the Mann-Whitney U test on pairs of models, with the alternative hypothesis that the scores of the first model were significantly greater than those of the second. The results indicated that the TverskyBCE model outperformed both SegResNet ($p = 0.04$) and Default ($p = 0.03$) in whole tumor segmentation. Additionally, TverskyBCE exhibited statistically significant higher SEN across all sub-regions compared to both SegResNet ($p = 0.02$) and Default ($p = 0.01$). These findings suggested that TverskyBCE offered superior performance in whole tumor segmentation and SEN metric.

 doi:10.20944/preprints202410.1738.v1

3.2. Individual Case Analysis and Visual Comparison

For each model and metric, we identified the cases with the lowest DSC, SEN, SPE, and the highest HD, labeling them as challenging cases. To find the test samples that most models struggled with, we counted how often each sample was marked as a challenging case. Finally, we selected the top three most challenging cases across all models for further visual analysis.

Test sample 733 was identified as the most challenging case, having been marked 28 times across models. The tumor is small, spans only a few slices, and is centrally located in the brain. Most models, including the TverskyBCE scratch-trained model, failed to correctly segment the tumor core. However, the TverskyBCE Fine-tuned model performed significantly better, as shown in Figure 3, achieving the highest DSC and the lowest HD for this case. The Tversky loss function helped the model focus on reducing false negatives, which is essential for capturing small tumor regions. However, the full potential of this loss function became evident only after fine-tuning. Fine-tuning allowed the model to further refine its weights, adapting more closely to the specific dataset and learning critical patterns necessary for detecting small, subtle lesions. The combination of the Tversky loss and fine-tuning enabled TverskyBCE Fine-tuned to successfully segment the tumor core, which was missed by the scratch-trained version and other models. This highlights the complementary roles of the Tversky loss function and fine-tuning in enhancing SEN to challenging cases.

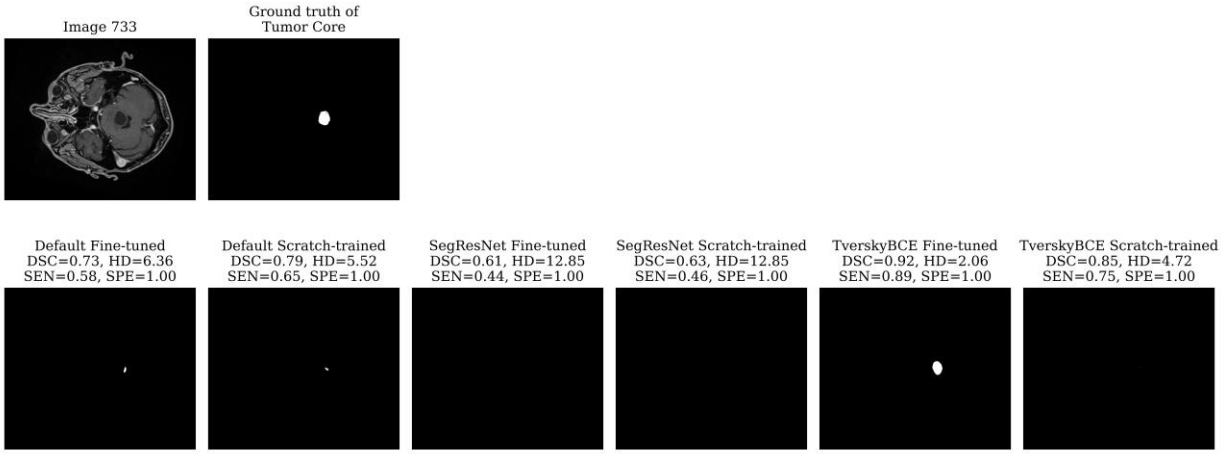


Figure 3. Segmentation results for Test Sample 733. TverskyBCE Fine-tuned achieved the highest DSC and lowest HD, successfully capturing the small tumor core missed by other models.

Test sample 731 was identified as the second most challenging case, marked 11 times across models. The tumor is large, located on one side of the brain with significant surrounding non-enhancing FLAIR hyperintensity. While all models detected the tumor boundary well, as indicated by the low HD for both the tumor core and whole tumor, most models struggled with the enhancing tumor, missing the middle region, which resulted in low SEN. The TverskyBCE Fine-tuned model performed best, with the smallest amount of missing tumor tissue. Although the DSC and SEN scores did not fully reflect the visual difference, the TverskyBCE Fine-tuned model clearly showed fewer missing regions in the images, as shown in Figure 4. This demonstrates that the combination of Tversky loss and fine-tuning was particularly effective at reducing false negatives and capturing more of the enhancing tumor.

 doi:10.20944/preprints202410.1738.v1

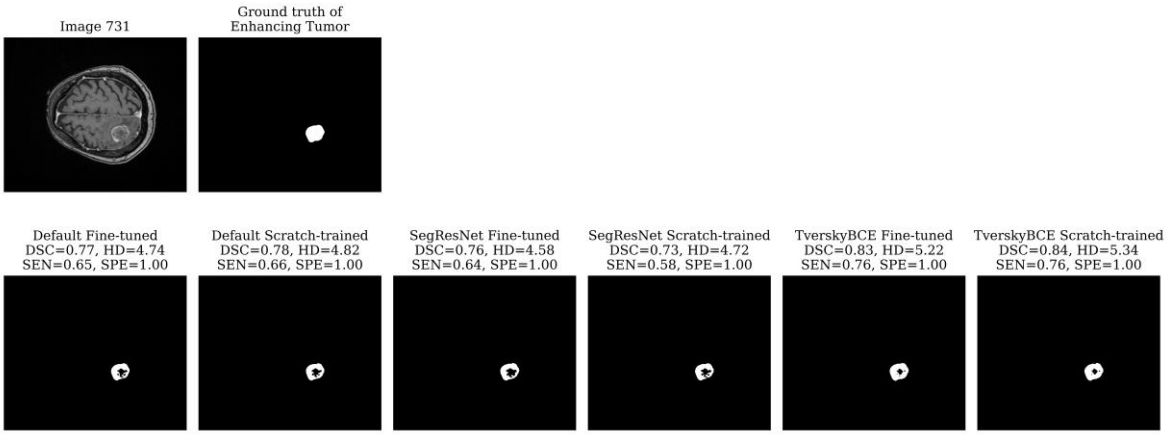


Figure 4. Segmentation results for Test Sample 731. TverskyBCE Fine-tuned had the fewest missed regions for the enhancing tumor, with the highest SEN.

Test sample 739 was marked 11 times as a challenging case. The tumor is very large, covering a significant portion of the brain and spanning multiple slices. While most models segmented the boundary reasonably well for both the enhancing tumor and the tumor core, they struggled with the segmentation of the whole tumor, resulting in a high HD. The whole tumor's irregular, heterogeneous shape, with jagged edges and asymmetrical extensions, as shown in Figure 5, added to the difficulty. The TverskyBCE Fine-tuned model performed best in this case, achieving the highest DSC and SEN, along with the lowest HD for the whole tumor.

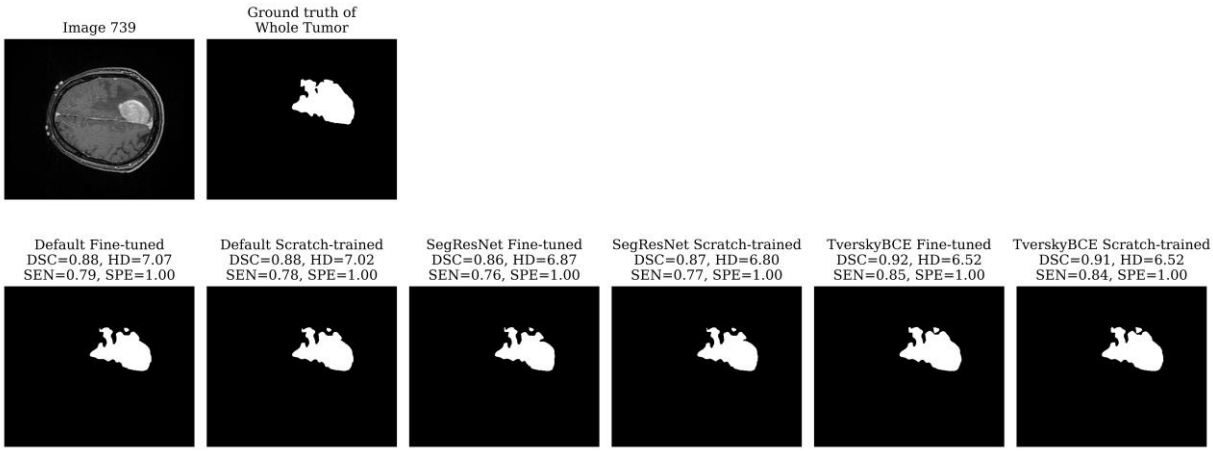


Figure 5. Segmentation results for Test Sample 739. TverskyBCE Fine-tuned captured the large, irregular tumor more effectively, with the highest DSC and SEN.

The analysis showed that the TverskyBCE Fine-tuned model outperformed other models, particularly in challenging cases with small or irregular tumors. The combination of Tversky loss and fine-tuning was crucial in reducing false negatives and improving segmentation accuracy. Fine-tuning enabled the model to better adapt to the dataset, allowing it to capture complex tumor boundaries more effectively.

3.3. Expert Feedback and Clinical Relevance

After evaluating the predictions of the TverskyBCE Fine-tuned and SegResNet Fine-tuned models on 10 test cases, medical experts provided feedback based on medical visualization (evaluation) and comparisons with ground truth annotations. Overall, both models performed well. In cases with large, straightforward lesions, such as neuroendocrine metastases (Sample 730) and

 doi:10.20944/preprints202410.1738.v1

undifferentiated cancer metastases (Samples 734 and 739), there were minimal differences between the models' results. However, both models struggled with recognizing additional small tumors, often missing tiny lesions in more complex cases. Metastasis are grown from cancerous cells which spread to the brain from the affected area by blood. The imbalanced objects and regions issue affects the data-driven learning algorithm as the extracted features may be highly influenced by large tumors and additional small parts. For example, to the regions - the necrotic/non-enhancing tumor core region is much smaller than other regions; to the objects - a characteristic feature of metastases is their multicentrality, they can appear in several areas of the brain with different consequences (size of lesions).

In test sample 728 (breast metastases), displayed in Figure 6, both models achieved high quantitative metrics, including a high DSC (0.90 – 0.92) and low HD (0.8 – 2.0). However, upon review by medical experts, both models failed to detect tiny tumors, which are critical for accurate diagnosis and treatment. Missing these small tumors is especially concerning in metastatic breast cancer, as even tiny lesions can indicate early-stage metastasis or tumor progression, which directly impacts treatment decisions and prognosis. This discrepancy highlights the limitations of relying solely on quantitative metrics and indicates that comparative metrics may not fully capture the models' performance in detecting small, clinically important lesions.

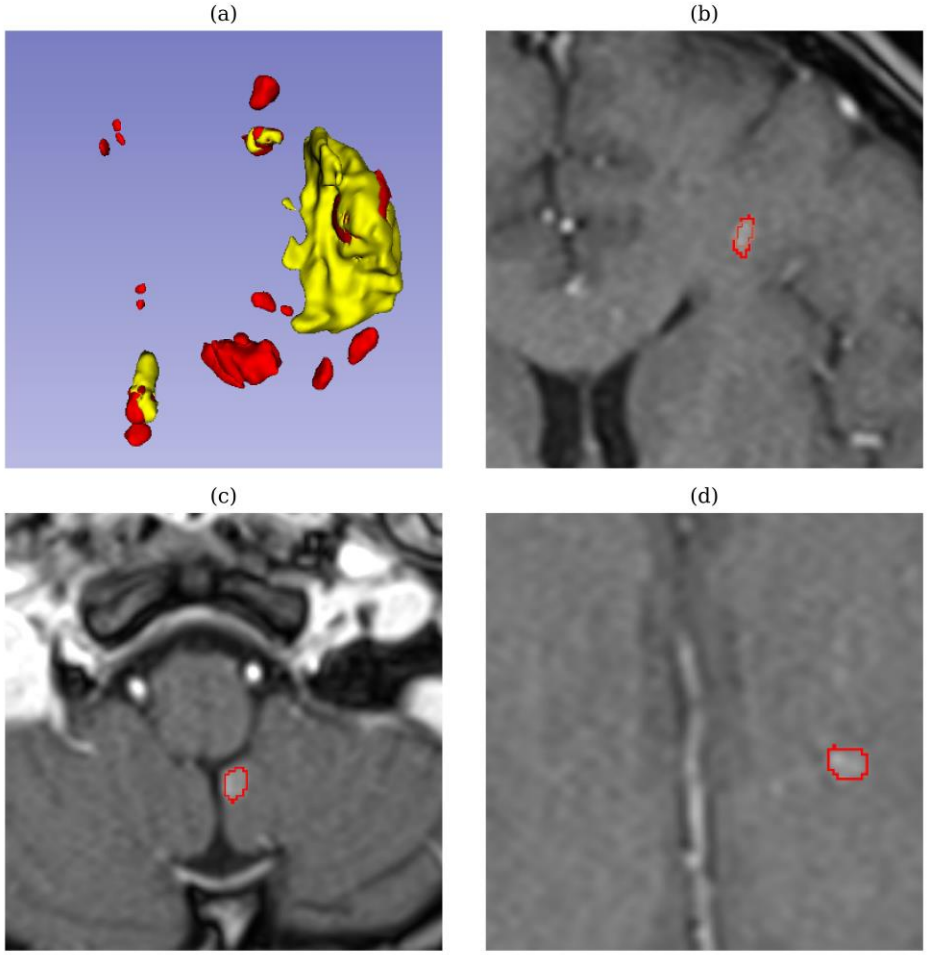


Figure 6. Segmentation results for Case 728 (breast metastasis), (a) ground truth showing multiple tiny and scattered tumors, (b) SegResNet correctly identified one small tumor, (c) and (d) are tiny tumor regions that both models completely missed.

In test sample 729 (adenocarcinoma metastases), TverskyBCE outperformed SegResNet in segmenting a cerebellar tumor. While SegResNet showed less edema in its segmentation, TverskyBCE's result was closer to the ground truth, capturing the tumor's characteristics more

 doi:10.20944/preprints202410.1738.v1

accurately. Similarly, in test sample 733 (low-grade cancer metastases), TverskyBCE successfully detected the necrotic tumor core, a critical sub-region that SegResNet entirely missed. Necrosis within a tumor is crucial for diagnosis and prognostic evaluation. This comparison is visualized in Figure 7, where the differences between the models are shown.

In summary, expert feedback indicated that while both models performed reliably in simpler cases, TverskyBCE demonstrated a clear advantage in handling more complex scenarios, particularly in recognizing necrotic core and accurately managing intricate tumor features. These findings highlight the strengths of each model and point to areas for improvement, especially in detecting small tumors and complex tissue structures.

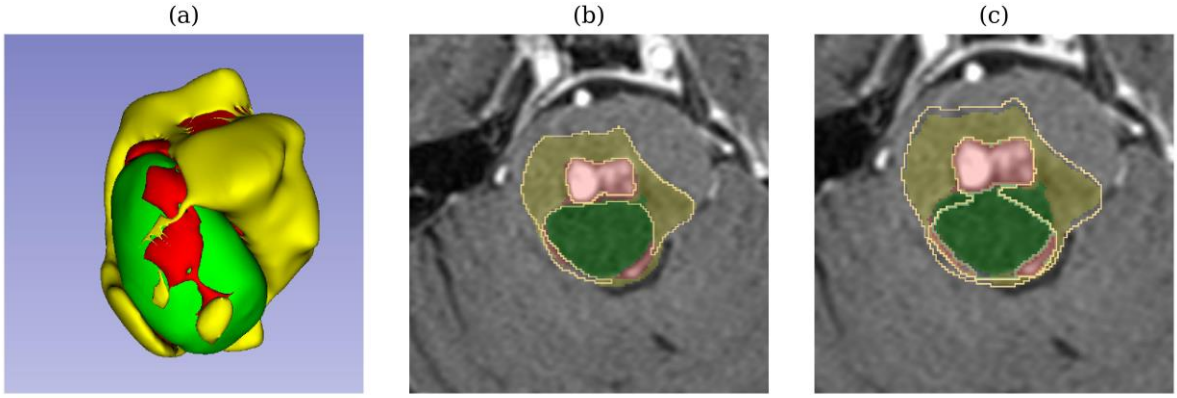


Figure 7. Visual comparison of segmentation results for case 729 (adenocarcinoma metastases) showing (a) the ground truth annotated by medical experts, (b) SegResNet segmentation where the necrotic tumor core (the green area) was missed, and (c) TverskyBCE segmentation successfully detecting the necrotic core. Solid-colored regions represent the ground truth annotations, while the outlined regions depict the model predictions.

4. Conclusions and Discussion

In this study, we assessed the performance of transfer learning models for brain metastasis segmentation. We pre-trained nnU-Net (with default settings and with Tversky and BCE loss) and SegResNet on the BraTS Metastases 2024 dataset, then fine-tuned these models on our private SBT dataset. We compared the fine-tuned models with those trained from scratch, using both quantitative and qualitative analyses, and to get feedback from medical experts on the segmentations produced. This expert input was crucial for evaluating clinical accuracy and ensuring the models were fit for practical use.

Our main objective was to develop methodologies for diagnosing and screening metastases. Brain metastases occur when cancer cells spread to the brain from the affected area. Any cancer can spread to the brain, but lung cancer, breast cancer, colon cancer, kidney cancer, and melanoma are the most likely to cause brain metastases. Tumors that experts call high-grade gliomas are tumors of the central nervous system (CNS). They are solid tumors and appear due to a mutation of brain or spinal cord cells. Since these tumors grow in the central nervous system, they are also called primary CNS tumors. That is, they are not metastases from other malignant tumors that have grown in other organs and their cancer cells have penetrated the central nervous system. An additional feature is that the lesions can be small - difficult to detect with the human eye. The diagnosis is highly dependent on the quality of the image, the thickness of the slices, the scanning parameters and the expertise of the doctor.

Throughout this study, several key challenges were identified in the context of brain metastasis segmentation:

- Location and morphological uncertainty: Metastases are grown from cancers cells which spread to the brain from the affected area by blood. Due to the wide spatial distribution of cancers cells, either lung cancer, breast cancer, colon cancer, kidney cancer, and melanoma may appear at

 doi:10.20944/preprints202410.1738.v1

any location inside the brain. The shape and size of different brain tumors varies with large morphology uncertainty. Each sub-regions of a meta may also vary in shape and size.

- Low contrast: High resolution and high contrast images are expected to contain diverse image information. Due to the image projection and tomography process, MRI images may be of low quality and low contrast. The boundary between biological tissues tends to be blurred and hard to detect.

- Annotation bias: Manual annotation highly depends on individual experience, which can introduce an annotation bias during data labeling. The annotation biases have a huge impact on the AI algorithm during the learning process

- Imbalanced region and object issues: The imbalanced issue affects the data-driven learning algorithm as the extracted features may be highly influenced by large tumors region. For example, to the regions - the necrotic/non-enhancing tumor core (NCR/ECT) region is much smaller than another region; to the objects - a characteristic feature of metastases is their multicentrality, they can appear in several areas of the brain with different consequences (size of lesions).

Our results showed that fine-tuning on the SBT dataset led to a slight improvement in metrics compared to training from scratch. This suggests that transfer learning is helpful when labeled data is limited, as pre-trained models use features learned from larger datasets. However, the performance differences were not statistically significant, likely due to the small test size (10 samples). Despite this, qualitative analysis showed that the fine-tuned models, especially TverskyBCE, performed better in challenging cases. TverskyBCE detected small or irregularly shaped tumors and did not miss the tumor core, unlike other models. This suggests that a customized loss function like Tversky with Binary Cross-Entropy helps handle class imbalance and improves sensitivity for smaller lesions. Medical experts also confirmed that TverskyBCE detected critical tumors in some cases where SegResNet did not.

However, all models, including TverskyBCE and SegResNet, struggled with detecting tiny tumors. Experts emphasized the need to improve detection of smaller lesions and boundary accuracy, despite the models achieving good overall segmentation quality. Even with high quantitative metrics, both models failed to detect small tumors, revealing that metrics alone are insufficient to capture clinically important issues. The involvement of medical experts is crucial in identifying these gaps, as missing tiny tumors can significantly impact treatment decisions and prognosis. Our findings not only demonstrate the potential of transfer learning for brain metastasis segmentation but also highlight the importance of medical expert input in developing AI algorithms. Their insights help bridge the gap between quantitative performance and real-world clinical needs, ensuring AI models are refined to detect smaller tumors and improve boundary segmentation for more accurate treatment planning.

Looking ahead, future research could address these limitations by applying advanced post-processing techniques to improve boundary delineation and using synthetic data generation or data augmentation to enhance tiny tumor detection. Incorporating multi-modal imaging data, such as combining MRI with CT or PET, could also provide more comprehensive information and boost model performance. Additionally, exploring new model architectures or hybrids could further improve segmentation, particularly for small or complex lesions. Another important direction would be to consider including edema in segmentation, as brain edema often surrounds metastases and could provide critical context for more accurate tumor delineation. Finally, ongoing collaboration with medical experts will be key to refining models and ensuring they meet clinical standards, with expert feedback incorporated into the model development process to continuously improve performance in real-world applications.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on : https://github.com/luumsk/BrainMetaSeg.git.

Author Contributions: Conceptualization, Bair N. Tuchinov and Andrey Yu. Letyagin; methodology, Andrey Yu. Letyagin; software, Victor Suvorov and Minh Sao Khue Luu; validation, Evgeniya V. Amelina, Andrey Yu. Letyagin, and Bair N. Tuchinov; formal analysis, Minh Sao Khue Luu; investigation, Minh Sao Khue Luu; resources, Bair N. Tuchinov; data curation, Evgeniya V. Amelina; writing—original draft preparation, Minh Sao Khue Luu and Roman M. Kenzhin; writing—review and editing, Bair N. Tuchinov and Andrey Yu. Letyagin;

visualization, Roman M. Kenzhin; supervision, Andrey Yu. Letyagin; project administration, Bair N. Tuchinov; funding acquisition, Bair N. Tuchinov. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730324P540002) and the agreement with the Novosibirsk State University dated December 27, 2023, No. 70-2023-001318.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Our partner FSBI Federal Neurosurgical Center specializing in neurosurgical operations (including brain cancers) - collect all documents including Consent Statement from the patients.

Data Availability Statement: The BraTS Metastases 2024 dataset presented in this study is openly available at https://www.synapse.org/Synapse:syn59059764. The Siberian Brain Tumor dataset presented in this article is not readily available due to privacy and ethical restrictions, as it contains private health data collected from our Federal Neurosurgical Center. Sharing it publicly would violate patient privacy and confidentiality agreements in accordance with ethical and regulatory standards. Requests for access to this dataset may be considered on a case-by-case basis, contingent upon appropriate ethical review and agreements.

Acknowledgments: We acknowledge the use of ChatGPT (OpenAI) for editorial support in language enhancement during the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Achrol, A.S.; Rennert, R.C.; Anders, C.; Soffietti, R.; Ahluwalia, M.S.; Nayak, L.; Peters, S.; Arvold, N.D.; Harsh, G.R.; Steeg, P.S.; et al. Brain Metastases. *Nat. Rev. Dis. Primer* **2019**, *5*, 5, doi:10.1038/s41572-018-0055-y.
2. Gavrilovic, I.T.; Posner, J.B. Brain Metastases: Epidemiology and Pathophysiology. *J. Neurooncol.* **2005**, *75*, 5–14, doi:10.1007/s11060-004-8093-6.
3. Lassman, A.B.; DeAngelis, L.M. Brain Metastases. *Neurol. Clin.* **2003**, *21*, 1–23, doi:10.1016/S0733-8619(02)00035-X.
4. Hall, W.; Djalilian, H.; Nussbaum, E.; Cho, K. Long-Term Survival with Metastatic Cancer to the Brain. *Med. Oncol.* **2000**, *17*, 279–286, doi:10.1007/BF02782192.
5. Lin, X.; DeAngelis, L.M. Treatment of Brain Metastases. *J. Clin. Oncol.* **2015**, *33*, 3475–3484, doi:10.1200/JCO.2015.60.9503.
6. Derks, S.H.A.E.; Van Der Veldt, A.A.M.; Smits, M. Brain Metastases: The Role of Clinical Imaging. *Br. J. Radiol.* **2022**, *95*, 20210944, doi:10.1259/bjr.20210944.
7. Prezelski, K.; Hsu, D.G.; Del Balzo, L.; Heller, E.; Ma, J.; Pike, L.R.G.; Ballangrud, Å.; Aristophanous, M. Artificial-Intelligence-Driven Measurements of Brain Metastases' Response to SRS Compare Favorably with Current Manual Standards of Assessment. *Neuro-Oncol. Adv.* **2024**, *6*, vdae015, doi:10.1093/noajnl/vdae015.
8. Nabors, L.B.; Portnow, J.; Ahluwalia, M.; Baehring, J.; Brem, H.; Brem, S.; Butowski, N.; Campian, J.L.; Clark, S.W.; Fabiano, A.J.; et al. Central Nervous System Cancers, Version 3.2020, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Canc. Netw.* **2020**, *18*, 1537–1570, doi:10.6004/jnccn.2020.0052.
9. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444, doi:10.1038/nature14539.
10. Akkus, Z.; Galimzianova, A.; Hoogi, A.; Rubin, D.L.; Erickson, B.J. Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. *J. Digit. Imaging* **2017**, *30*, 449–459, doi:10.1007/s10278-017-9983-4.
11. Zhao, F.; Wu, Z.; Li, G. Deep Learning in Cortical Surface-Based Neuroimage Analysis: A Systematic Review. *Intell. Med.* **2023**, *3*, 46–58, doi:10.1016/j.imed.2022.06.002.
12. Jyothi, P.; Singh, A.R. Deep Learning Models and Traditional Automated Techniques for Brain Tumor Segmentation in MRI: A Review. *Artif. Intell. Rev.* **2023**, *56*, 2923–2969, doi:10.1007/s10462-022-10245-x.
13. Singh, S.P.; Wang, L.; Gupta, S.; Goli, H.; Padmanabhan, P.; Gulyás, B. 3D Deep Learning on Medical Images: A Review. *Sensors* **2020**, *20*, 5097, doi:10.3390/s20185097.
14. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.W.M.; Van Ginneken, B.; Sánchez, C.I. A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* **2017**, *42*, 60–88, doi:10.1016/j.media.2017.07.005.
15. Chukwujindu, E.; Faiz, H.; AI-Douri, S.; Faiz, K.; De Sequeira, A. Role of Artificial Intelligence in Brain Tumour Imaging. *Eur. J. Radiol.* **2024**, *176*, 111509, doi:10.1016/j.ejrad.2024.111509.
16. Grøvik, E.; Yi, D.; Iv, M.; Tong, E.; Rubin, D.; Zaharchuk, G. Deep Learning Enables Automatic Detection and Segmentation of Brain Metastases on Multisequence MRI. *J. Magn. Reson. Imaging* **2020**, *51*, 175–182, doi:10.1002/jmri.26766.

17. Huang, Y.; Bert, C.; Sommer, P.; Frey, B.; Gaipl, U.; Distel, L.V.; Weissmann, T.; Uder, M.; Schmidt, M.A.; Dörfler, A.; et al. Deep Learning for Brain Metastasis Detection and Segmentation in Longitudinal MRI Data. *Med. Phys.* **2022**, *49*, 5773–5786, doi:10.1002/mp.15863.
18. Sadegheih, Y.; Merhof, D. Segmentation of Brain Metastases in MRI: A Two-Stage Deep Learning Approach with Modality Impact Study 2024.
19. Yang, S.; Li, X.; Mei, J.; Chen, J.; Xie, C.; Zhou, Y. 3D-TransUNet for Brain Metastases Segmentation in the BraTS2023 Challenge 2024.
20. Bouget, D.; Alsinan, D.; Gaitan, V.; Helland, R.H.; Pedersen, A.; Solheim, O.; Reinertsen, I. Raidionics: An Open Software for Pre- and Postoperative Central Nervous System Tumor Segmentation and Standardized Reporting. *Sci. Rep.* **2023**, *13*, 15570, doi:10.1038/s41598-023-42048-7.
21. Bozinovski, S. Reminder of the First Paper on Transfer Learning in Neural Networks, 1976. *Informatica* **2020**, *44*, doi:10.31449/inf.v44i3.2828.
22. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A Survey on Deep Transfer Learning. In *Artificial Neural Networks and Machine Learning – ICANN 2018*; Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2018; Vol. 11141, pp. 270–279 ISBN 978-3-030-01423-0.
23. Kora, P.; Ooi, C.P.; Faust, O.; Raghavendra, U.; Gudigar, A.; Chan, W.Y.; Meenakshi, K.; Swaraja, K.; Plawiak, P.; Rajendra Acharya, U. Transfer Learning Techniques for Medical Image Analysis: A Review. *Biocybern. Biomed. Eng.* **2022**, *42*, 79–107, doi:10.1016/j.bbe.2021.11.004.
24. Wacker, J.; Ladeira, M.; Nascimento, J.E.V. Transfer Learning for Brain Tumor Segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*; Crimi, A., Bakas, S., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2021; Vol. 12658, pp. 241–251 ISBN 978-3-030-72083-4.
25. Tataei Sarshar, N.; Ranjbarzadeh, R.; Jafarzadeh Ghoushchi, S.; De Oliveira, G.G.; Anari, S.; Parhizkar, M.; Bendechache, M. Glioma Brain Tumor Segmentation in Four MRI Modalities Using a Convolutional Neural Network and Based on a Transfer Learning Method. In *Proceedings of the 7th Brazilian Technology Symposium (BTSym'21)*; Iano, Y., Saotome, O., Kemper Vásquez, G.L., Cotrim Pezzuto, C., Arthur, R., Gomes De Oliveira, G., Eds.; Smart Innovation, Systems and Technologies; Springer International Publishing: Cham, 2023; Vol. 207, pp. 386–402 ISBN 978-3-031-04434-2.
26. Messaoudi, H.; Belaid, A.; Ben Salem, D.; Conze, P.-H. Cross-Dimensional Transfer Learning in Medical Image Segmentation with Deep Learning. *Med. Image Anal.* **2023**, *88*, 102868, doi:10.1016/j.media.2023.102868.
27. Huang, Y.; Khodabakhshi, Z.; Gomaa, A.; Schmidt, M.; Fietkau, R.; Guckenberger, M.; Andratschke, N.; Bert, C.; Tanadini-Lang, S.; Putz, F. Multicenter Privacy-Preserving Model Training for Deep Learning Brain Metastases Autosegmentation. **2024**, doi:10.48550/ARXIV.2405.10870.
28. Pani, K.; Chawla, I. A Hybrid Approach for Multi Modal Brain Tumor Segmentation Using Two Phase Transfer Learning, SSL and a Hybrid 3DUNET. *Comput. Electr. Eng.* **2024**, *118*, 109418, doi:10.1016/j.compeleceng.2024.109418.
29. de Verdier, M.C.; Saluja, R.; Gagnon, L.; LaBella, D.; Baid, U.; Tahon, N.H.; Foltyn-Dumitru, M.; Zhang, J.; Alafif, M.; Baig, S.; et al. The 2024 Brain Tumor Segmentation (BraTS) Challenge: Glioma Segmentation on Post-Treatment MRI 2024.
30. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nat. Methods* **2021**, *18*, 203–211, doi:10.1038/s41592-020-01008-z.
31. Myronenko, A. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization 2018.
32. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks 2017.
33. Taha, A.A.; Hanbury, A. Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool. *BMC Med. Imaging* **2015**, *15*, 29, doi:10.1186/s12880-015-0068-x.
34. Shapiro, S.S.; Wilk, M.B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **1965**, *52*, 591–611, doi:10.1093/biomet/52.3-4.591.
35. Mann, H.B.; Whitney, D.R. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50–60, doi:10.1214/aoms/1177730491.
36. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.-C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; et al. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **2012**, *30*, 1323–1341, doi:10.1016/j.mri.2012.05.001.