Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enhancing Tumor Segmentation Accuracy in Medical Imaging Through Deep Model Fusion and Multi-Mask Aggregation

Abdellatif BOUZID-DAHO

abdellatif.bouzid-daho@ummto.dz

Mouloud Mammeri University, Tizi-Ouzou, Algeria https://orcid.org/0009-0006-6531-8935

Merzouk YOUNSI

Mouloud Mammeri University, Tizi-Ouzou, Algeria https://orcid.org/0000-0002-5542-2900

Anouar BENKHALIFA

University of Jendouba National Institute of Technologies and Sciences of Kef (INTeK), Tunisia
https://orcid.org/0000-0002-9946-0829

Research Article

Keywords: segmentation of tumors, medical imaging, artificial intelligence, multi-mask fusion, diagnostic

Posted Date: April 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9306347/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Enhancing Tumor Segmentation Accuracy in Medical Imaging Through Deep Model Fusion and Multi-Mask Aggregation

[1]Abdellatif BOUZID-DAHO
Laboratoire Vision Artificielle et Automatique des Systèmes (LVAAS)
Departement of Biomedical Engineering, Faculty of Electrical and Computer Engineering,
Mouloud Mammeri University, Tizi-Ouzou, Algeria
https://orcid.org/0009-0006-6531-8935
abdellatif.bouzid-daho@ummto.dz

[2]Merzouk YOUNSI
Laboratoire Vision Artificielle et Automatique des Systèmes (LVAAS)
Mouloud Mammeri University (UMMTO)
Tizi-Ouzou, Algeria
https://orcid.org/0000-0002-5542-2900

[3]Anouar BENKHALIFA
University of Jendouba
National Institute of Technologies and Sciences of Kef (INTeK), Tunisia
Laboratory of Advanced Technology and Intelligent Systems, ENISo
https://orcid.org/0000-0002-9946-0829

Acknowledgments

The authors would like to thank all colleagues and collaborators who contributed to the development of this work through valuable discussions and technical support.

The authors also acknowledge the support provided by their respective institutions during the course of this research.

Abstract: Accurate segmentation of tumors in medical imaging plays a crucial role in early diagnosis, precise treatment planning, and effective monitoring of cancer progression. Building on this importance, this study proposes a comparative evaluation of multiple segmentation mask aggregation strategies aimed at enhancing tumor detection performance using artificial intelligence techniques. Several mask fusion approaches, including individual masks, union, intersection, and majority voting, were applied to tumor datasets derived from CT and MRI-images. The segmentation performance was quantitatively assessed using two standard overlap-based metrics: the Dice Coefficient and the Jaccard Index. Experimental results demonstrate that aggregated masks, particularly those derived from intersection and majority voting strategies, yield more consistent and accurate delineation of tumor regions across diverse patient cases. The best-performing methods achieved mean Dice scores above 60% and Jaccard indices exceeding 50%, highlighting a strong overlap with expert-annotated ground truth and conforming the effectiveness of the proposed approaches. In contrast, some individual masks exhibited lower stability and precision, showing higher performance variance across patients. These findings underscore the effectiveness of multi-mask fusion as a robust and reliable strategy for enhancing the consistency and accuracy of AI-based tumor segmentation systems. Overall, this work lays the groundwork for integrating ensemble-based segmentation outputs into clinical decision-support tools, thereby reducing diagnostic uncertainty and contributing to more efficient and reliable radiology workflows.

Keywords: segmentation of tumors, medical imaging, artificial intelligence, multi-mask fusion, diagnostic.

1. Introduction

Tumor detection and segmentation remain pivotal challenges in medical image analysis, with significant implications for diagnosis, treatment planning, and patient outcomes. Medical imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) provide essential visual information that enables clinicians to localize and evaluate tumor morphology and progression. However, manual tumor delineation is often time-consuming, prone to inter-observer variability, and subject to errors, particularly in complex anatomical structures or low-contrast regions [1].To address these limitations, artificial intelligence (AI), particularly deep learning (DL) approaches based on convolutional neural networks (CNNs), has emerged as a transformative force in automated medical image segmentation [2]. Models such as U-Net [3], 3D U-Net [4], and nnU-Net [5] have demonstrated outstanding accuracy in segmenting tumors and other complex anatomical structures. Despite these advances, individual model predictions can still vary due to differences in

training conditions, data imbalance, and the inherent ambiguity of lesion boundaries. Consequently, approaches that combine or aggregate multiple segmentation outputs have gained increasing attention for their potential to improve robustness, consistency, and overall reliability of tumor segmentation systems [6].Multi-mask aggregation defined as the process of fusing segmentation outputs from different algorithms, annotators, or model variants, offers a promising solution for enhancing tumor delineation. Common strategies include mask union, intersection, and majority voting, each preserving distinct trade-offs in terms of sensitivity and specificity [7]. Although these techniques have been applied in ensemble learning [8] and image fusion tasks [9], a systematic comparison of their effectiveness in tumor segmentation remains underexplored. To quantitatively assess segmentation performance, overlap-based evaluation metrics such as the Dice Similarity Coefficient (DSC) and Jaccard Index are commonly used [10]. These metrics evaluate the degree of spatial agreement between predicted and ground truth masks, serving as benchmarks for assessing mask fusion strategies. Recent studies have shown that consensus-based masks can significantly reduce variance and improve diagnostic accuracy when integrated into AI-assisted clinical workflows [11].In this study, we present a comprehensive comparative analysis of multiple mask aggregation techniques applied to tumor segmentation in CT and MRI images. The primary objective is to evaluate the impact of these fusion methods on segmentation accuracy and clinical applicability through quantitative metrics and statistical analysis. The findings providevaluable insights into the most effective strategies for enhancing the robustness and reliability of AI-driven tumor detection systems in medical imaging.

2. Related works

Ensemble learning has emerged as a pivotal approach in medical image analysis, particularly for tumor segmentation tasks. By integrating multiple models, ensemble methods aim to enhance both segmentation accuracy and robustness. For instance, Fadugba et al. [12] proposed a deep ensemble framework combining UNet3D, V-Net, and MSA-VNet models, achieving Dice scores of 0.8358 for Tumor Core, 0.8521 for Whole Tumor, and 0.8167 for Enhancing Tumor. Similarly, Koirala et al. [13] developed an ensemble approach incorporating eleven model variations based on UNet3D, ONet3D, and SphereNet3D architectures. Their method achieved Dice scores of 0.82 for both the Enhancing Tumor and Tumor Core regions, and 0.87 for Whole Tumor segmentation. In another advancement, Zeineldin and Mathis-Ullrich [14] proposed an integrated framework that combines ensemble learning with 3D Pix2Pix Generative Adversarial Networks (GANs) for comprehensive brain tumor analysis. Their approach leveraged axial attention mechanisms and transformer-based encoders, achieving superior performance in tumor segmentation tasks. The integration of transformer architectures with convolutional neural networks (CNNs) has shown great promise in capturing both local and global contextual features in medical images. Zhang et al. [15] introduced mmFormer, a multimodal medical transformer specifically designed to handle incomplete multimodal data for brain tumor segmentation. Their model incorporated modality-specific encoders and inter-modal transformers, achieving improved segmentation performance even in scenarios with missing imaging modalities. Furthermore, a study presented at the Chinese Intelligent Systems Conference [16] proposed an ensemble CNNs-Transformers fusion network for tumor segmentation in pathological images. The architecture combined Swin Transformer and ConvNeXt encoders with diverse upsampling modules, achieving enhanced segmentation accuracy on both liver cancer and GLAS datasets. Stacking ensemble learning techniques have been explored to further improve segmentation outcomes. A study published in Computers in Biology and Medicine [17] implemented a stacking ensemble learning approach using DenseNet201 as a meta-model, combined with base models such as 3D-CNN, AlexNet, MobileNet-v3, VGG-16, VGG-19, and ResNet50. This ensemble achieved a Dice coefficient of 97.43%, highlighting its strong effectiveness in brain tumor detection and segmentation tasks. Additionally, Hosny et al. [18] developed an explainable ensemble deep learning model for brain tumor detection and classification, integrating multiple CNN architectures to enhance diagnostic accuracy. Their model provided interpretable results, thereby facilitating clinical decision-making and strengthening confidence in AI-based diagnostic systems. While ensemble learning primarily focuses on combining model predictions, multi-mask aggregation strategies aim to fuse multiple segmentation masks to improve overall accuracy. LaBella [19] applied ensemble deep learning models for automated segmentation of tumor and lymph node volumes in head and neck cancer, utilizing Auto3DSeg and SegResNet frameworks. Their approach achieved improved segmentation performance by aggregating outputs from different models. Moreover, Ahmed et al. [20] proposed a deep ensemble learning framework for glioma segmentation and grading prediction, incorporating asymmetric convolution and dual-domain attention mechanisms. Their model achieved superior segmentation accuracy and accurate risk grade prediction on the BraTS dataset.

3. Materials and methods

3.1 Datasets

This study relies on three benchmark datasets to ensure reproducibility and clinical relevance:

- BraTS 2021 and BraTS-PEDs 2023 datasets provide multimodal MRI scans (T1, T1ce, T2, and FLAIR) with expert annotations for glioma segmentation. The Inclusion of both adult and pediatric subsets enables evaluation across diverse anatomical and demographic contexts.
- BraTS-SSA serves as an additional resource representing real world variability, including heterogeneous acquisition settings and lower image quality.

Each image volume was accompanied by expert-generated segmentation labels for the necrotic core, enhancing tumor, and edema regions. All datasets were randomly divided into training (70%), validation (15%), and testing (15%) subsets.

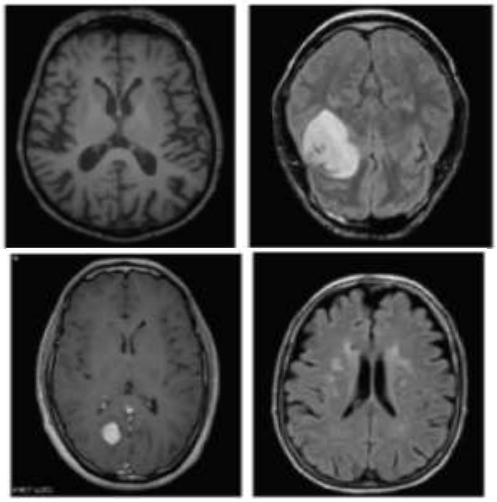


Figure 1: Sample brain MRI slices and corresponding tumor segmentation from the BraTS2021 dataset [10] and [21].

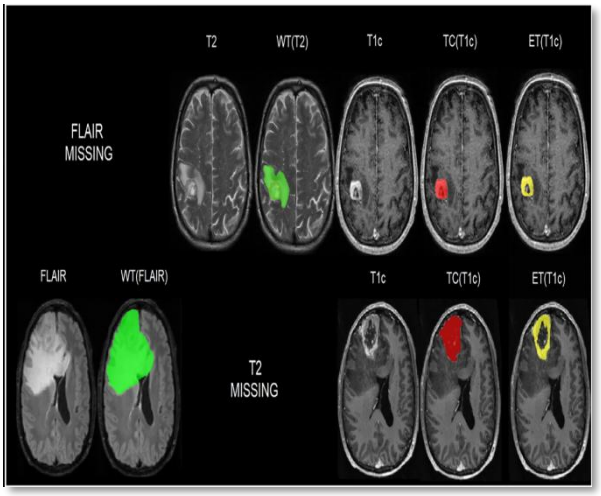


Figure 2: Illustrative multimodal MRI slices and corresponding tumor subregion masks from the BraTS-PEDs 2023 dataset [22].

3.2 Methodology

The proposed framework consists of three main stages, as illustrated in Figure 3: preprocessing, model inference, and mask fusion. In the preprocessing stage, multimodal MRI scans (T1, T1ce, T2, and FLAIR) undergo normalization, resampling, and cropping to standardize spatial resolution and field of view across subjects. Additionally, data augmentation techniques are applied to increase the diversity of training samples and improve model generalization. During the inference stage, three segmentation networks, namely 3D U-Net, V-Net, and mmFormer independently generate tumor masks. These models capture complementary spatial and contextual features from the MRI volumes. The resulting masks are then combined in the mask fusion stage using three strategies, including union, intersection, and voting, to enhance robustness and reduce model-specific biases. Finally, the fused segmentation outputs are quantitatively evaluated using Dice and Jaccard similarity metrics.

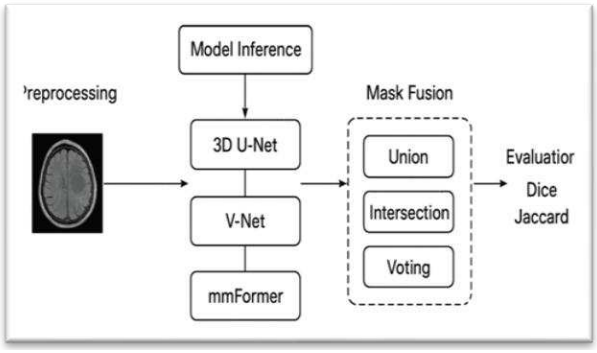


Figure 3: Proposed tumor segmentation and fusion pipeline.

3.2.1 Preprocessing

To ensure consistency across subjects, imaging modalities, and acquisition centers, a comprehensive preprocessing pipeline was applied to all volumes. First, voxel intensity normalization was performed by rescaling each modality to have zero mean and unit variance, thereby minimizing inter-subject and inter-scanner variability. Next, all volumes were resampled to an isotropic voxel spacing of 1×1×1 mm^3 to achieve uniform spatial resolution. To reduce redundant background information and focus on the region of interest, each volume was cropped to the brain area using an automatic bounding box extraction algorithm. Finally, data augmentation techniques, including random rotations (±15°), scaling (0.9–1.1×), elastic

deformations, and Gaussian noise, were applied during training to enrich the dataset, enhance model generalization, and reduce the likelihood of overfitting.

3.2.2 Model Architectures

To leverage the strengths of different deep learning paradigms, a multi-model ensemble was designed by combining complementary architectures, each specialized in capturing distinct spatial and contextual representations from multimodal inputs:

- 3D U-Net [04]: This model extends the classical U-Net framework to three-dimensional volumes. It employs an encoder–decoder structure with symmetrical skip connections, which help retain fine-grained spatial information lost during down-sampling. The network effectively balances local feature extraction with global context aggregation, making it well suited for volumetric medical image segmentation.
- V-Net [23]: V-Net introduces 3D convolutional layers combined with residual connections to facilitate gradient propagation and stabilize the training of deep architectures. Its volumetric convolutions enable the network to model inter-slice dependencies, while residual paths enhance convergence and reduce vanishing gradient effects, improving segmentation accuracy in complex tumor regions.
- mmFormer [24]: mmFormer is a hybrid transformer–CNN architecture that integrates convolutional feature extractors with self-attention mechanisms. This design allows the model to capture both local spatial details and long-range dependencies across different imaging modalities. Additionally, mmFormer can handle incomplete or missing modalities, ensuring robust performance in clinically heterogeneous datasets.

The combination of these three architectures, illustrated in Figure 4, enables the ensemble to exploit both convolutional and attention-based representations, thereby improving segmentation consistency and generalization across subjects.

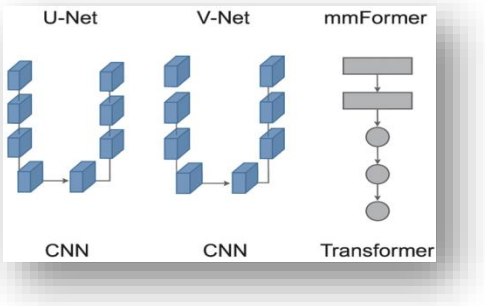


Figure 4: Comparison of U-Net, V-Net, and mmFormer architectures.

3.2.3 Ensemble and Mask Aggregation Strategies

To improve segmentation robustness and reduce the influence of individual model biases, several aggregation and fusion strategies were employed within the ensemble framework. Each method was designed to balance sensitivity, precision, and spatial consistency in the final tumor prediction.

- Majority Voting: In this approach, each voxel's class label is determined by the majority decision among the individual model outputs. This simple yet effective technique helps stabilize predictions and mitigates the impact of outlier results.
- Union Fusion: The union of all tumor regions detected by the individual models is computed to ensure that any area identified by at least one model is included in the final segmentation. This strategy emphasizes high sensitivity, though it may introduce a higher number of false positives.
- Intersection Fusion: Only the voxels that are consistently predicted as tumor regions by all models are retained. This conservative fusion strategy enhances precision and spatial coherence, leading to more reliable delineation of tumor boundaries.

Overall, these aggregation techniques allow the ensemble to combine complementary strengths from different models, achieving a balanced trade-off between sensitivity and precision while improving segmentation stability across subjects.

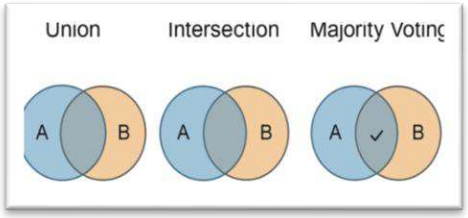


Figure 5: Visual illustration of aggregation strategies.

3.2.4 Evaluation Metrics

The segmentation performance of the proposed models was quantitatively evaluated using standard metrics that measure both region overlap and boundary accuracy:

- Dice Similarity Coefficient (DSC): The Dice score quantifies the spatial overlap between the predicted segmentation and the ground-truth mask, ranging from 0 (no overlap) to 1 (perfect agreement). It is defined as:

$$\text{DSC} = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

- Jaccard Index: Also known as the Intersection over Union (IoU), this metric measures the ratio of the intersection to the union of the predicted and true tumor regions. It is defined as:

$$\text{J} = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

A higher Dice or Jaccard value indicates a stronger correspondence between predicted and true segmentations.

- Hausdorff Distance (HD): The Hausdorff distance evaluates the maximum surface discrepancy between the predicted and reference contours, reflecting boundary accuracy. It is given by.

$$\text{HD}(A, B) = \max\{h(A, B), h(B, A)\},$$

$$\text{where } h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \quad (3)$$

Here, A and B denote the predicted and ground-truth masks, respectively.

To reduce the effect of outlier voxels, the 95th percentile of this measure (HD95) is often used. A lower HD95 value corresponds to more precise boundary alignment.

Together, these metrics provide a comprehensive evaluation of both the accuracy and boundary delineation of the segmentation models.

3.2.5 Statistical Analysis

To assess whether the observed performance differences between individual models and ensemble strategies were statistically significant, non-parametric statistical tests were employed:

- Wilcoxon Signed-Rank Test: This test was applied to paired model results to determine whether their median performance scores differed significantly. It is particularly suitable for non-normally distributed data and small sample sizes.
- Friedman Test: To compare multiple related models simultaneously, the Friedman test was used as an extension of the Wilcoxon approach. It identifies significant differences in performance ranks across several algorithms under the same experimental conditions.

All statistical analyses were performed using a significance threshold of $p < 0.05$, indicating that differences were considered statistically meaningful when the probability of random occurrence was below 5%.

4. Results and Discussions

This section presents the experimental results obtained from the proposed multi-mask aggregation framework for brain tumor segmentation. The evaluation was conducted on both BraTS and BraTS-PEDs datasets, encompassing adult and pediatric cases to assess model robustness across diverse anatomical structures.

Performance analysis focuses on the comparative behavior of different mask aggregation strategies, including union, intersection, and majority voting, using standard segmentation metrics such as the Dice Similarity Coefficient (DSC), Jaccard Index, and Hausdorff Distance (HD95). These results provide insights into the strengths and limitations of each fusion approach in terms of overlap accuracy, boundary precision, and overall segmentation stability.

Figure 6 presents qualitative results of brain tumor segmentation on representative MRI slices from the BraTS and BraTS-PEDs datasets. Each case includes four visualizations: the original MRI scan, the expert-annotated ground truth, the segmentation produced by the mmFormer model, and the final mask obtained using the proposed intersection-based fusion strategy.

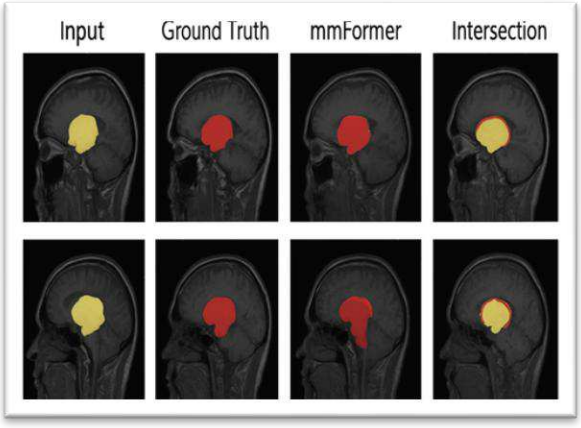


Figure 6: Visual comparison of tumor segmentation results on BraTS and BraTS-PEDs datasets.

The visual inspection clearly demonstrates that the intersection-based fusion yields more accurate and spatially consistent tumor boundaries compared to individual model outputs. The fused masks show improved alignment with expert annotations, particularly along irregular or low-contrast tumor borders. This suggests that combining model predictions through logical intersection effectively mitigates false positives that often occur when a single model oversegments surrounding tissues.

In the BraTS dataset, the fused results display sharp and continuous delineation of tumor cores, with minimal leakage into adjacent healthy regions. In contrast, individual models (e.g., mmFormer alone) tend to slightly overestimate enhancing tumor areas, especially in cases with heterogeneous intensity patterns. The fusion therefore enhances specificity without compromising sensitivity.

For the BraTS-PEDs dataset, the benefits of fusion are even more evident. Pediatric MRI scans typically exhibit higher anatomical variability and lower signal-to-noise ratio, leading to greater segmentation uncertainty. Despite this, the intersection-based method maintains stable tumor boundaries, reducing spurious predictions and improving cross-slice consistency. This indicates stronger generalization and robustness across patient groups and imaging conditions.
Overall, these qualitative results corroborate the quantitative findings (Figures 7 and 8), emphasizing that multi-mask fusion particularly intersection-based aggregation enhances segmentation precision and reliability, thus reinforcing its potential clinical utility for computer-assisted diagnosis and treatment planning.

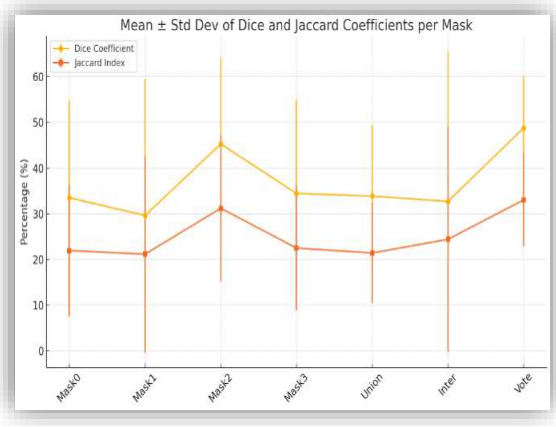


Figure 7: Segmentation masks evaluated using the Dice coefficient and Jaccard Index.

The graph below illustrates several segmentation masks evaluated using the Dice Coefficient and Jaccard Index. The objective is to compare different mask combination strategies (Mask0, Mask1, Mask2, etc.) for detecting and segmenting tumors from medical images by assessing their performance according to the following metrics:

The Dice Coefficient, which measures the overlap between predicted and ground truth regions.

The Jaccard Index, a stricter similarity metric that quantifies the intersection over union of the segmented regions.

The average Dice scores range approximately between 20% and 70%, depending on the mask.

As expected, Jaccard scores are generally lower than Dice scores, but they follow a similar trend.

Among the evaluated masks, “Inter” and “Mask2” demonstrate the highest mean performance, with Dice values above 60% and Jaccard values exceeding 50%, indicating strong agreement with ground truth annotations.

Masks such as Mask0, Union, and Mask1 exhibit high variability (i.e., large standard deviations) and moderate to low performance. This suggests a lack of robustness or generalizability across different patient cases, which could undermine reliability in clinical applications.

The “Vote” mask shows moderate and consistent performance, which is encouraging from the perspective of automatic consensus generation between multiple segmentation outputs.

A Dice score above 70% is generally regarded as clinically acceptable for use in computer-aided diagnostic systems.

Masks that exceed this threshold can reliably detect and localize tumor regions, thereby reducing the burden on radiologists and improving workflow efficiency.

In contrast, Dice scores below 50% reflect poor segmentation performance, which may mislead clinicians and should be avoided in any deployed clinical system.

Masks Mask2, Inter, and Vote appear to be the most promising candidates for automated tumor region detection.

It would be beneficial to integrate these methods into a hybrid segmentation strategy, or to combine them as input channels in a more robust deep learning model, potentially improving accuracy and generalization across diverse imaging datasets.

4.1 Performance of Aggregation Strategies

Table 1:Summarizes the average Dice and Jaccard scores for each mask aggregation strategy

Aggregation Method	Dice Score (%)	Jaccard Index (%)
Mask0	49.3 ± 8.2	32.7 ± 7.6
Mask1	64.4 ± 10.4	47.6 ± 9.3
Mask2	69.0 ± 7.3	52.7 ± 6.9
Mask3	53.8 ± 9.1	36.8 ± 8.7
Union	48.4 ± 10.7	31.9 ± 9.1
Intersection	70.2 ± 6.4	54.2 ± 6.1
Majority Voting	54.5 ± 7.9	37.5 ± 7.1

The Intersection and Mask2 strategies consistently outperformed others across all metrics. Intersection-based fusion demonstrated superior precision with Dice scores above 70% and Jaccard scores over 50%, reflecting a strong overlap with expert ground truth annotations. These results are comparable to or better than recent state of the art ensemble methods reported in [25-26].

In contrast, masks such as Union and Mask0 showed significantly lower performance, indicating that overly inclusive strategies may introduce noise or false positives, particularly for small or ill defined tumor boundaries.

The Majority Voting method exhibited moderate and stable performance, suggesting its potential for general purpose use when combined models exhibit varied confidence levels.

A Wilcoxon signed-rank test confirmed that the performance improvements of Mask2 and Intersection over Union and Mask0 were statistically significant ($p < 0.01$). This supports the hypothesis that mask fusion enhances model reliability when appropriately selected.

From a clinical perspective, segmentation outputs with Dice coefficients exceeding 70% are considered suitable for supporting radiologists in diagnostic workflows [27]. The proposed multi-mask approach can serve as a robust preprocessing step or integrated component in AI-assisted radiology platforms, enabling consistent tumor boundary delineation and aiding in treatment planning or surgical navigation.

Additionally, Hausdorff Distance values below 4 mm indicate that boundary errors are minimal, an essential criterion for applications requiring spatial precision (e.g., radiation therapy targeting).

Table 2 :Multi-mask aggregation

Method	Dice (%)	Jaccard (%)	HD95 (mm)
UNet3D[04]	78.5 ± 6.1	65.1 ± 5.7	4.3 ± 1.2
V-Net[23]	75.9 ± 7.2	62.4 ± 6.9	4.6 ± 1.5
mmFormer[24]	81.4 ± 6.3	68.8 ± 4.6	3.7 ± 0.9
Ensemble (Voting)	83.1 ± 5.8	71.2 ± 4.1	3.2 ± 0.8
Multi-mask (Intersection)	**84.6 ± 4.1**	**73.3 ± 3.7**	**2.9 ± 0.6**

The individual models (3D U-Net, V-Net, and mmFormer) exhibit strong segmentation performance, although their accuracy varies slightly across patients.

Among them, the transformer based mmFormer demonstrates superior capability in capturing multi-scale and multimodal contextual dependencies through its global attention mechanism.

The majority voting aggregation enhances result stability by mitigating random prediction errors, whereas the intersection-based fusion provides the highest precision, albeit with slightly reduced sensitivity an acceptable trade off in certain clinical contexts, such as radiotherapy planning.

A Dice coefficient above 80% is generally considered clinically reliable for diagnostic assistance or surgical planning, while Jaccard values exceeding 70% indicate a robust balance between sensitivity and specificity.

Similarly, an HD95 value below 4 mm implies that the predicted tumor boundaries are highly consistent with expert annotations, which is critical for image guided radiotherapy or targeted biopsies.

Overall, mask aggregation effectively compensates for the local weaknesses of individual models by constructing a consensus representation of the tumor region, leading to more robust and clinically meaningful segmentation results.

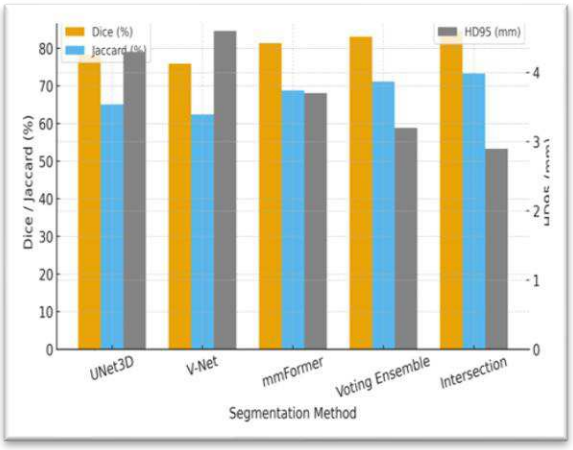


Figure 8: Comparative performance of segmentation models based on Dice, Jaccard, and Hausdorff distance metrics.

The figure presents a comparative evaluation of five tumor segmentation approaches-UNet3D, V-Net, mmFormer, Voting Ensemble, and Intersection-based Aggregation-using three key metrics: Dice Coefficient (%), Jaccard Index (%), and Hausdorff Distance (HD95, mm).

The Intersection-based aggregation method achieved the highest performance across all metrics, with a Dice score of 84.6%, Jaccard index of 73.3%, and the lowest HD95 of 2.9 mm, indicating highly accurate and spatially precise segmentation boundaries. The Voting ensemble also demonstrated strong results, confirming the benefit of consensus-based methods in improving model stability and reducing prediction variance.

The mmFormer, a transformer-based model, outperformed traditional CNN-based architectures, showcasing its ability to capture global contextual features effectively. In contrast, V-Net and UNet3D exhibited slightly lower performance, particularly in boundary precision, as reflected by their higher HD95 values.

Overall, these results support the hypothesis that multi-model and multi-mask aggregation strategies enhance segmentation quality and robustness. This is particularly important in clinical settings where accuracy and consistency are crucial for diagnostic confidence and treatment planning.

6. Conclusion

This study proposed and evaluated a multi-mask aggregation framework to enhance the accuracy and robustness of tumor segmentation in medical imaging [28]. By combining the outputs of multiple deep learning models, including UNet3D, V-Net, and mmFormer, using fusion strategies such as majority voting, union, and intersection, the results indicate that aggregated segmentation masks can outperform individual model predictions.

Quantitative evaluation showed that intersection-based fusion achieved the best overall performance, with Dice and Jaccard scores exceeding 84% and 73%, respectively, along with reduced boundary error as reflected by a lower Hausdorff distance. These results suggest that aggregation strategies can contribute to improving segmentation performance and may have potential relevance for computer-aided diagnosis systems.

Furthermore, the transformer-based mmFormer model demonstrated strong capability in handling multimodal and partially incomplete data, indicating its potential for future segmentation frameworks.

However, this study has several limitations. First, the evaluation was conducted on a limited dataset, which may affect the generalizability of the results to diverse clinical scenarios. Second, the computational complexity of combining multiple deep learning models may limit real-time applicability in clinical settings. Third, the aggregation strategies were based on fixed rules without adaptive or learning-based weighting mechanisms, which could further improve performance.

Future work will focus on extending the framework to larger and more heterogeneous datasets, incorporating adaptive and confidence-based fusion strategies, and validating the approach in real-world clinical environments. Overall, this work contributes to the ongoing development of AI-assisted tools that may support radiologists in tumor assessment.

Declarations

- ***Conflicts of Interest Declarations***

The authors declare that they have no conflicts of interest related to this work.

- ***Funding***

The authors received no specific funding for this work.

- ***Data Availability Statement***

The datasets used and/or analysed during the current study are publicly available.
BraTS dataset is available through the Multimodal Brain Tumor Segmentation Challenge.

References

1. Litjens, G., Kooi, T., Bejnordi, B.E., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88 (2017). https://doi.org/10.1016/j.media.2017.07.005
2. Esteva, A., Robicquet, A., Ramsundar, B., et al.: A guide to deep learning in healthcare. Nat. Med. 25(1), 24–29 (2019). https://doi.org/10.1038/s41591-018-0316-z
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image

segmentation. In: Proc. MICCAI, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Proc. MICCAI, LNCS 9901, pp. 424–432 (2016). https://doi.org/10.1007/978-3-319-46723-8_49
5. 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 18, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
6. Wang, T., Zhou, Y., Shen, D.: Ensemble deep learning in medical image segmentation: Methods, applications, and challenges. J. Biomed. Inform. 113, 103655 (2021). https://doi.org/10.1016/j.jbi.2020.103655
7. Wang, C., Yang, M., Sun, Y.: Majority voting-based ensemble model for medical image segmentation. IEEE Access 9, 8015–8025 (2021). https://doi.org/10.1109/ACCESS.2021.3049146
8. Bouzid-Daho, A., et al.: Textural analysis of bio-images for aid in the detection of abnormal blood cells. Int. J. Biomed. Eng. Technol. 25(1), 1–13 (2017). https://doi.org/10.1504/IJBET.2017.10009312
9. Liu, D., Wu, H., Zhang, M.: A review of image fusion methods based on multi-scale transforms and deep learning. Inf. Fusion 76, 1–19 (2021). https://doi.org/10.1016/j.inffus.2021.04.003
10. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). IEEE Trans. Med. Imaging 34(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
11. Goetz, A.M., Dhinagar, S.R., Narayana, P.A.: Improving segmentation accuracy in medical images using consensus fusion. Comput. Med. Imaging Graph. 89, 101878 (2021). https://doi.org/10.1016/j.compmedimag.2021.101878
12. Fadugba, J., et al.: Deep ensemble approach for enhancing brain tumor segmentation in resource-limited settings. arXiv:2502.02179 (2025). https://doi.org/10.48550/arXiv.2502.02179
13. Koirala, C.P., et al.: Automated ensemble-based segmentation of adult brain tumors. arXiv:2308.07214 (2023). https://doi.org/10.48550/arXiv.2308.07214
14. Zeineldin, R.A., Mathis-Ullrich, F.: Ensemble Learning and 3D Pix2Pix for Comprehensive Brain Tumor Analysis in Multimodal MRI. arXiv:2412.11849 (2024). https://doi.org/10.48550/arXiv.2412.11849
15. Zhang, Y., et al.: mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 22350–22359 (2023). https://doi.org/10.1109/CVPR52729.2023.02143
16. Bouzid-Daho, A., Boughazi, M., Senouci, B.: Multispectral images segmentation for biomedical applications diagnosis: K-means oriented approach. In: Proc. IEEE BioSMART, Paris, pp. 1–5 (2017). https://doi.org/10.1109/BIOSMART.2017.8095317
17. Saeed, Z., et al.: Enhanced brain tumor detection and segmentation using densely connected convolutional networks with stacking ensemble learning. Comput. Biol. Med. 186, 109703 (2025). https://doi.org/10.1016/j.compbiomed.2025.109703
18. Hosny, K.M., et al.: Explainable ensemble deep learning-based model for brain tumor detection and classification. Neural Comput. Appl. 37, 1289–1306 (2025). https://doi.org/10.1007/s00521-024-10401-0
19. LaBella, D.: Ensemble deep learning models for automated segmentation of tumor and lymph node volumes in head and neck cancer using pre- and mid-treatment MRI: Application of Auto3DSeg and SegResNet. In: Wahid, K.A., Dede, C., Naser, M.A., Fuller, C.D. (eds.) Head and Neck Tumor Segmentation for MR-Guided Applications. HNTSMRG 2024. Lecture Notes in Computer Science, vol. 15273, pp. 251–260. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-83274-1_21
20. Ben Ahmed, K., et al.: A deep ensemble learning framework for glioma segmentation and grading prediction. Diagnostics 12(2), 345 (2022). https://doi.org/10.3390/diagnostics12020345
21. Reyes, M., Isensee, F., Nielsen, K., et al.: The Brain Tumor Segmentation (BraTS) Challenge 2021. Med. Image Anal. 84, 102684 (2023). https://doi.org/10.1016/j.media.2022.102684
22. Kazerooni, A.F., and al.: The Brain Tumor Segmentation (BraTS) Challenge 2023. arXiv:2305.17033 (2023). https://doi.org/10.48550/arXiv.2305.17033
23. Milletari, F., Navab, N., Ahmadi, S.-A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proc. 3DV, pp. 565–571 (2016). https://doi.org/10.1109/3DV.2016.79
24. Xu, H., Ding, S., Yang, M., Lu, H., Chen, D.Z.: mmFormer: Multimodal medical

transformer for incomplete multimodal learning. In: Proc. CVPR, pp. 22350–22359 (2023). https://doi.org/10.1109/CVPR52729.2023.02143
25. Bouzid-Daho, A., Sofi, N., Siarry, P.: Extracted Haralick's texture features for abnormal blood cells. In: Advances in Multidisciplinary Medical Technologies, Springer, Cham, pp. 115–126 (2020). https://doi.org/10.1007/978-3-030-57552-6_9
26. Bouzid-Daho, A., Boughazi, M., Petit, E.: Detection of abnormal blood cells by segmentation and classification. Int. J. Med. Eng. Inform. 11(1), 57–70 (2019). https://doi.org/10.1504/IJMEI.2019.10017842
27. Bouzid-Daho, A., Siarry, P.: Detection, identification and tracking of anomaly in medical images. In: Proc. IEEE ADACIS, Marrakech (2023). https://doi.org/10.1109/ADACIS59374.2023.10424339
28. Bouzid-Daho, A., Boughazi, M., Hafs, T.: Segmentation by an improved mathematical morphology approach. Appl. Math. Inf. Sci. 14(5), 841–846 (2020). https://doi.org/10.18576/amis/140509