Team work makes the dream work: An ensemble learning approach for segmentation with Convolutional Neural Networks and Vision Transformers

Anne Andresen[1, 2], Yasmin Lassen-Ramshad[2], Christian Rønn Hansen[4, 5, 6], Slavka Lucakova[1, 3], Nadine Vatterodt[1, 2], and Jesper Kallehauge[1, 2]

[1]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[2]Danish Center For Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[3]Department of Oncology, Aarhus University Hospital, Aarhus, Denmark
[4]Department of Oncology, Odense University Hospital, Odense, Denmark
[5]Laboratory of Radiation Physic, Odense University Hospital, Odense, Denmark

Abstract

Introduction
Recent developments in segmentation focus on enhancing accuracy and efficiency. Convolutional Neural Networks (CNN), especially the UNet architecture, has been applied for segmentation, however lately Vision Transformers ViT models have achieved high accuracy for segmentations. Therefore using a CNN UNet and ViT in ensemble learning could benefit from the different properties, to enhance a models' capabilities of handling variations in organ sizes, while providing high quality segmentations on smaller datasets.

Method and Materials
nnUNet and a ViT model were trained individually to perform segmentation of 8 organs at risk (OAR) in the brain. The dataset included 55 T1 contrast enhanced MRI scans split with 49 used for training and 6 for testing. Ensemble learning was applied to the models from fold 4 to generate the final segmentation. The evaluation compared the predicted segmentations and manual delineations provided by oncologists, using dice score, 1mm normalized surface dice and hausdorff distance 95th percentile.

Results
1mm normalized surface dice medians was in the range of 0.90 to 0.94, the dice in the range of 0.85-0.97 and hausdorff 95th percentile ranged from 1.0 to 2.0 across 8 organs at risk,

Discussion
The results suggest that ensemble learning of different small deep learning models could contribute to increasing accuracy when having a smaller dataset size and class imbalance however additional evaluation on a new dataset is needed.

1 Introduction

With the recent advancement in artificial intelligence, in particular in deep learning, more advanced models are being proposed to obtain highly accurate and clinically acceptable segmentations. Traditionally, this is achieved by increasing a single models' complexity, which creates a reliance on the availability of large datasets as the size of the dataset impacts the performance of the segmentation model.[1] This is a challenge in radiotherapy as data is often limited resulting in a small dataset sizes.

Additionally a challenge called class imbalance also arises, which can induce biases in model performance, when the model struggles to adequately learn the underrepresented class. So a limited dataset size coupled with class imbalance presents obstacles for acquiring a high accuracy and generalizability of segmentation models. To enhance model performance and address these challenges effectively, several techniques could be employed.

Therefore we apply ensemble learning for two distinct, separately trained, deep learning models: a Convolutional Neural Network (CNN) UNet and a Vision Transformer (ViT). We hypothesize that this will reduce the impact of the limited dataset size and class imbalance for multi organ segmentation on MRI scans of patients with brain cancer.

1.1 Related works

Significant advancements have been made in the field of medical image segmentation in recent years, to improve accuracy and efficiency of the models . CNNs have frequently been proposed for segmentation, with UNets as a widely employed architecture. CNN UNets has been successful in delineating organs at risk(OARs) in various areas of the body including in the brain. [2] [3] [4] [5] [6].

More recently ViT models have been used for segmentation. [1]. ViTs have been shown to provide highly accurate segmentations of small structures which have been attributed to their ability to capture global context and relationships within the images. [1] [7] [8]

These notable advancements observed in the individual models, prompted a suggestion to integrate the properties of CNNs and transformers. This could use the diverse model properties for capturing both local and global information, thereby enhancing the models' ability to handle variations in organ sizes [9] [10]. For brain structures such architectures have been employed, to enhance segmentation accuracy of OARs, including for brain structures where DSCs of 0.96 has been achieved for larger organs, however this on significantly larger datasets. [1]. More commonly DSCs of 0.70-0.89 are acquired on smaller datasets. [11] [12]

While advancements in medical image segmentation are evident, challenges such as limited dataset sizes and achieving consistently high segmentation accuracy independent of organ size, remain [13] [14]

2 Materials and Methods

2.1 Data

Patient demographics of the patients included in the study is presented in table 1

Median Age (range)	47.0 (35.5 - 59.5)
Gender	Male: 53 % , Female : 47 %
Tumor type:	glioma , meningioma

Table 1: Patient demographics

The available dataset for this study consists of 55 T1-weighted contrast-enhanced (T1w CE) Magnetic Resonance Imaging (MRI) scans with corresponding manually delineated structures of 8 OARs of patients with brain cancer. The 8 OARs manually delineated include

- Brainstem
- Chiasm
- Hippocampus L/R
- Optic nerves L/R
- Optic tract L/R

49 of the patients were allocated as training data with the remaining 6 patients for test.

2.2 Preprocessing

Initial preprocessing consisted of z-score normalization, to standardize the intensity values across all scans. Subsequently, cropping a central region with boundaries based on the label mask was performed to reduce the computational load.

2.3 Ensemble learning

Ensemble learning was applied to integrate the predictions of two individually trained weak models, to collectively improve quality and generalization via a strong model. This approach allows for fine-tuning of individual models, emphasizing their distinctive strengths of smaller compact models which are well-suited for optimization of the grounds of smaller datasets while enhancing accuracy.
The initial weak model in the ensemble is a UNet architecture using CNN, specifically nnUNet, while the second weak model in the ensemble adopts ViTs trained using an identical dataset to collectively create a strong model which provides final segmentation of the 8 OARs as presented in figure 1.

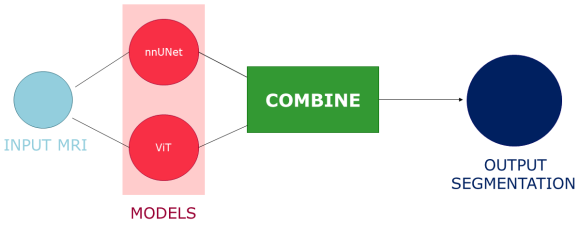

Figure 1: Representation of ensemble learning

2.4 Evaluation

For assessment of the proposed ensemble model, a set of evaluation metrics representing the volumetric and distance accuracy was employed. These metrics encompass the 1mm normalized surface Dice (1NsDSC), Dice score (DSC), and the 95th percentile of the Hausdorff distance (HD95). Using these metrics, we assess the similarity between the predicted segmentations from the ensemble model and the manually delineated OARs provided by an oncologist.

3 Results

3.1 Dice results

The median DSC values ranged from 0.85 to 0.97 across all segmented structures. The lowest median DSC was observed for the right optic tract see figure 2, while the highest obtained median DSC was for the brainstem.

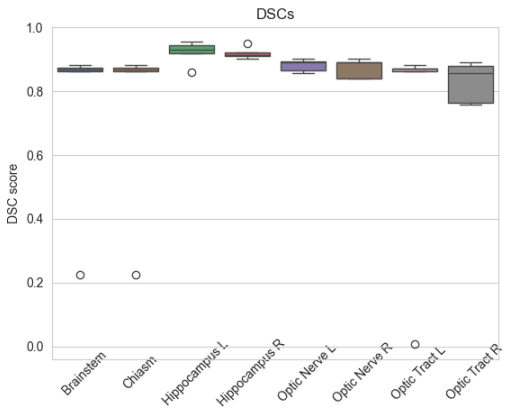

Figure 2: DSCs for the ensemble mode for 8 OARs

3.2 1 mm normalized surface dice results

The 1NsDSC, demonstrated median scores within the range of 0.90 to 0.94, with the lowest medians being for the left hippocampus and the highest being the brainstem

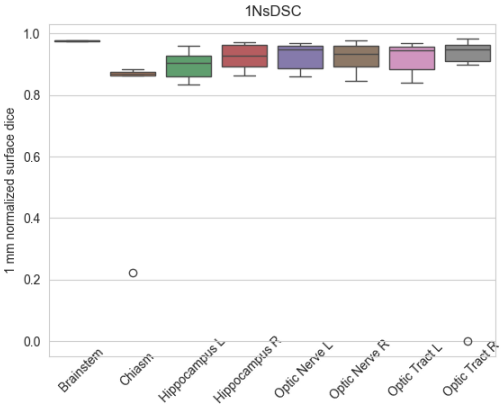

Figure 3: The ensemble models' 1 mm normalized surface dice for 8 OARs

3.3 HD95 distance

The HD95 distance metric, medians ranging from 1.0 to 2.0, with the right hippocampus displaying the lowest HD95 distance and the left optic tract showcasing the highest.

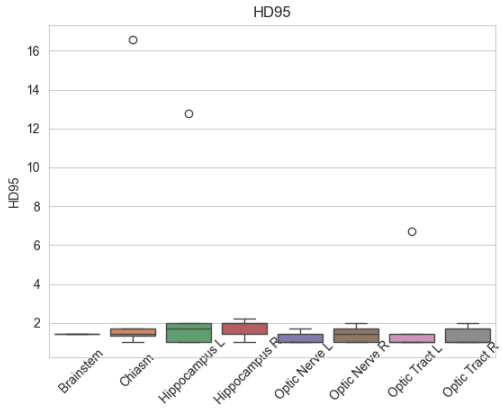

Figure 4: Hausdorff distance 95th percentile for the ensemble model across 8 OARs

4 Discussion

4.1 Results

The ensemble learning approach, combining the strengths of UNet and ViT models suggest a plausible approach when working on smaller datasets. The consistently high DSC and 1NsDSC, coupled low HD95 distances, suggest a high efficacy of the ensemble models' ability to achieve precise segmentations across the diverse set of OARs. Furthermore, these findings indicate a high capacity to localize small anatomical structures in challenging regions, which could indicate that combining various types of deep learning models in such way could contribute to mitigate the class imbalances otherwise present in the dataset.

Assessing the effectiveness of the ensemble approach can be done through a comparative assessment of the ensemble model and the weaker models contributing to the final strong model. Effectiveness is assessed by comparison of the evaluation metrics, specifically the 1NsDSC and HD95, which was used in the evaluation of the ensemble model performance. Examination of the 1NsDSC scores, for the nnUNet, medians lie in the range of 0.77 to 0.89 and 0.85-0.90 for the ViT. Compared to the medians of 0.90-0.94 for the ensemble, it suggests that the ensemble model's ability to define organ boundaries did improve when compared to its individual weak models. For the HD95, the ViT and nnUNet models showed medians in the range of 3.3-1.9 and 4.3-1.0, respectively, whereas the the ensemble model demonstrated distances within the range of 2.0- 1.0. This coupled with the DSC scores results suggest an improvement when utilizing the ensemble to create the final segmentations
The ensemble has notable enhancements in accuracy, as indicated by the evaluation metrics, which suggest a potential of ensemble modeling as a robust strategy for improving medical image segmentation tasks when having small datasets containing class imbalance.

4.2 Model

The decision to employ ensemble learning, combining the strengths of CNN UNet and ViT models, was motivated by their ability distinct features and patterns due to their different architectures. The UNet provided local information, while the ViT ensures a global oversight of the entire image. By using both models, we aimed to exploit the unique capabilities of each model type for more accurate segmentation. Training on the same dataset, the models were provided with a common understanding of the underlying patterns in the dataset which contributed to the consistency of the ensemble model. Meanwhile the likelihood of overfitting can be reduced by distribution of parameters between two models with individualized training sessions.

4.3 Clinical Significance

The consistently high performance across multiple metrics suggests a strong potential for clinical use. Accurate segmentation of OARs is important in treatment planning of patients with brain cancer, and the ensemble models are promising. However, additional clinical qualitative evaluation should be performed.

4.4 Limitations and Future Directions

Result obtained in this study rely on the quality of the training dataset. Future work could explore the impact of dataset variability and the generalization capacity of the models for a larger patient cohort in the test phase. Additionally, the proposed ensemble architecture opens the possibility for additional fine-tuning and optimization on smaller datasets for an enhancement performance on other modalities. Finally, a

clinical assessment to determine the quality of the segmentations in a clinical setting is needed.

5 Conclusion

In conclusion, the ensemble learning approach integrating UNet and ViT models seems promising for the segmentation of OARs in the brain. The obtained results underscore the potential clinical significance of the proposed methodology and contribute to the ongoing discourse on using diverse model architectures for enhanced medical image segmentation. As the field continues to evolve, this work sets the stage for further exploration, refinement, and application in real-world clinical scenarios.

References

[1] J. Ma, Y. He, F. Li, et al. "Segment anything in medical images". *Nature Communications* 15.1 (2024), p. 654.

[2] E. Orasanu, T. Brosch, C. Glide-Hurst, et al. "Organ-at-risk segmentation in brain MRI using model-based segmentation: benefits of deep learning-based boundary detectors". *International Workshop on Shape in Medical Imaging*. Springer. 2018, pp. 291–299.

[3] Z. Akkus, A. Galimzianova, A. Hoogi, et al. "Deep learning for brain MRI segmentation: state of the art and future directions". *Journal of digital imaging* 30 (2017), pp. 449–459.

[4] R. Azad, E. K. Aghdam, A. Rauland, et al. "Medical image segmentation review: The success of u-net". *arXiv preprint arXiv:2211.14830* (2022).

[5] N. Siddique, S. Paheding, C. P. Elkin, et al. "U-net and its variants for medical image segmentation: A review of theory and applications". *Ieee Access* 9 (2021), pp. 82031–82057.

[6] N. Salpea, P. Tzouveli, and D. Kollias. "Medical image segmentation: A review of modern architectures". *European Conference on Computer Vision*. Springer. 2022, pp. 691–708.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". *arXiv preprint arXiv:2010.11929* (2020).

[8] K. Han, Y. Wang, H. Chen, et al. "A survey on vision transformer". *IEEE transactions on pattern analysis and machine intelligence* 45.1 (2022), pp. 87–110.

[9] H. Tang, Y. Chen, T. Wang, et al. "HTC-Net: A hybrid CNN-transformer framework for medical image segmentation". *Biomedical Signal Processing and Control* 88 (2024), p. 105605.

[10] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghoushchi, et al. "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images". *Scientific Reports* 11.1 (2021), p. 10930.

[11] D. N. Greve, B. Billot, D. Cordero, et al. "A deep learning toolbox for automatic segmentation of subcortical limbic structures from MRI images". *Neuroimage* 244 (2021), p. 118610.

[12] B. Lee, N. Yamanakkanavar, and J. Y. Choi. "Automatic segmentation of brain MRI using a novel patch-wise U-net deep architecture". *Plos one* 15.8 (2020), e0236493.

[13] D. Zhang, Y. Lin, H. Chen, et al. "Deep learning for medical image segmentation: tricks, challenges and future directions". *arXiv preprint arXiv:2209.10307* (2022).

[14] R. Wang, T. Lei, R. Cui, et al. "Medical image segmentation using deep learning: A survey". *IET Image Processing* 16.5 (2022), pp. 1243–1267.