2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635469

ENHANCING GENERALIZABILITY IN BRAIN TUMOR SEGMENTATION: MODEL ENSEMBLE WITH ADAPTIVE POST-PROCESSING

Zhifan Jiang[1], Daniel Capellán-Martín[1,2]*, Abhijeet Parida[1,2]*, Xinyang Liu[1], María J. Ledesma-Carbayo[2], Syed Muhammad Anwar[1,3], Marius George Linguraru[1,3]*

[1] Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC, USA
[2] Biomedical Image Technologies, ETSI Telecomunicación, Universidad Politécnica de Madrid & CIBER-BBN, Madrid, Spain
[3] School of Medicine and Health Sciences, George Washington University, Washington, DC, USA

ABSTRACT

Segmentation of brain tumors in multi-parametric magnetic resonance imaging facilitates quantitative analysis crucial for clinical trials and personalized patient care. This significantly influences clinical decision-making, encompassing diagnosis and prognosis and enhancing patient outcomes. The brain tumor segmentation (BraTS) challenge, in its 2023 edition, extended to a cluster of competitions incorporating multiple tumor types. Now, in conjunction with IEEE ISBI 2024, BraTS organizes its Generalizability Across Tumors (BraTS-GoAT) challenge. In this paper, we introduce a deep-learning-based ensemble strategy involving three state-of-the-art segmentation models. Furthermore, we also introduce a novel adaptive post-processing method, based on a cross-validated tumor-specific threshold search, designed to output enhanced accurate segmentations, ensuring generalizability across various tumor types. The evaluation of our proposed method on validation cases resulted in lesion-wise Dice scores of 0.842, 0.854, 0.872 and lesion-wise 95th-percentile Hausdorff Distance scores of 29.46, 24.67, 25.22 for the enhancing tumor, tumor core, and whole tumor, respectively.

***Index Terms*—** Brain tumor segmentation, Deep learning, Generalizability, MRI, Unsupervised learning.

1. INTRODUCTION

Since 2012, the Brain Tumor Segmentation (BraTS) challenge has served as a key platform for benchmark environments and datasets for the segmentation of adult brain gliomas [1]. The BraTS 2023 challenge expanded to a cluster of competitions, incorporating tumor types such as adult glioma, Sub-Saharan African brain glioma, brain/intracranial meningioma, brain metastasis, and pediatric brain tumors [2, 3, 4, 5, 6, 7]. In conjunction with the International Symposium on Biomedical Imaging (ISBI) 2024, the BraTS Generalizability Across Tumors (BraTS-GoAT) Challenge provides an opportunity to address generalizability in brain tumor segmentation tasks.

Successful segmentation algorithms in previous BraTS challenges have relied on model ensembling and post-processing techniques. In this scenario, model ensembling involves combining predictions from state-of-the-art (SOTA) segmentation architectures, e.g., nnU-Net [8], Swin UNETR [9, 10], and SegResNet [11], which have demonstrated outstanding performance in previous BraTS editions. Various ensemble strategies can be adopted while averaging probabilities is one of the most prevalent approaches. Post-processing is crucial for improving segmentation performance, especially for small sub-regions, and it includes enhancing tumor and tumor core. A common post-processing step involves removing small disconnected regions that introduce spurious and noisy blobs into segmentation outcomes. However, this task can be challenging as detecting small lesions, especially in metastatic scenarios, is critical for patient prognosis, as missing even one lesion can lead to repeated interventions and treatment delays.

In this work, we introduce an ensemble method involving three SOTA deep learning models, followed by an adaptive post-processing algorithm designed to output enhanced accurate segmentations, ensuring generalizability across various tumor types.

2. MATERIALS AND METHODS

2.1. Data Description

The BraTS-GoAT challenge used a subset of data from the BraTS 2023 cluster of challenges, consisting of 2,251 training and 360 validation cases. Each case comprises four multi-parametric MRIs: T1-weighted (T1), contrast-enhanced T1 (T1CE), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). Manual annotations are provided to establish ground truth for three tumor regions: enhancing tumor (ET), peritumoral edematous/invaded tissues (ED), and necrotic tumor core (NCR). Segmentation performance is evaluated on combined tumor regions: ET, tumor core (TC, NCR, and ET), and whole tumor (WT, TC, and ED). All data underwent registration, resampling, and skull-stripping before annotation, resulting in isotropic 1 mm^3 images with dimensions of $240\times240\times155$ voxels.

The key difference of this dataset lies in the varying presence of each label across training, validation, and test data while maintaining consistent label values, which aims to assess the model's generalizability across lesion types, institutions, and demographics.

2.2. Deep Learning Models

Our proposed pipeline (Fig.1) has two main steps to enhance brain tumor MRI segmentation's generalizability: i) using a model ensemble technique with three SOTA models as the starting point: nnU-Net ("no new U-Net", winner of BraTS 2020) [8], Swin UNETR ("Swin U-Net transformers", top-performing model of BraTS 2021) [9, 10], and SegResNet (BraTS 2018 winning solution) [11]; ii) applying adaptive tumor-specific post-processing, which adjusts post-processing for each tumor type using unsupervised clustering on images.

nnU-Net: The nnU-Net is a self-configuring deep learning segmentation framework based on the U-Net architecture [12]. Using a five-fold cross-validation approach, we trained a full-resolution 3D nnU-Net (v2) model. The training followed a region-based approach, where each output channel corresponds to the WT, TC, and ET regions. We opted for larger patches ($128\times160\times112$ voxels) while ensuring they fit within the GPU's capacity [8]. We used a class-weight loss function

* These authors contributed equally.

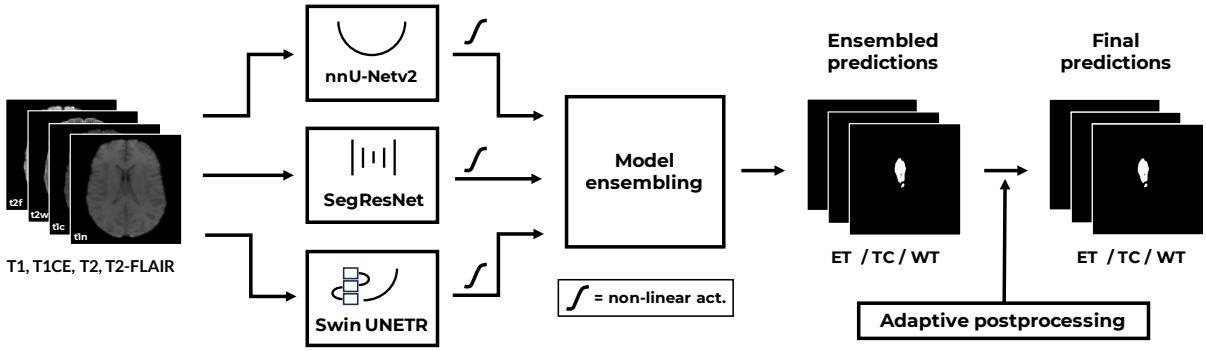

Fig. 1: Proposed pipeline: model ensemble and adaptive post-processing. Predictions are obtained from three SOTA deep learning models subjected to nonlinear activation functions and ensemble strategies. The ensemble predictions are then subjected to an adaptive tumor-specific post-processing step.

combining Dice loss and cross-entropy loss. Optimization was done using stochastic gradient descent (SGD) with Nesterov momentum, with an initial learning rate of 0.01, momentum of 0.99, and weight decay of 3e-05. Each fold underwent single-GPU training for 200 epochs on NVIDIA A100 40GB and NVIDIA V100 16GB GPUs. During inference, images were predicted using a sliding window approach matching the patch size used for training. The nnU-Net (v2) was implemented in https://github.com/MIC-DKFZ/nnUNet.

Swin UNETR: [9, 10, 13] utilizes a vision transformer (ViT)-based [14] hierarchical design. Employing a 3D Swin transformer encoder with window shifting to broaden its receptive fields, it connects seamlessly with a multi-scale residual U-Net-like decoder, tuned for applications like 3D medical image segmentation. We trained a 3D Swin UNETR model using a five-fold cross-validation and a region-based approach. Each batch contained four random patches of $128\times128\times128$ voxels, and the batch size was set to 1. The training employed a class-weight loss function combining Dice loss and focal loss, optimized using the AdamW optimizer with an initial learning rate of 0.0001, momentum of 0.99, and weight decay of 3e-05. Each fold was trained for 100 epochs on a cluster with 4 NVIDIA H100 80GB GPUs and 4 NVIDIA V100 32GB GPUs. For inference, we used a sliding window approach with the same patch size and overlapping of 0.625 for neighboring patches. The implementation integrated into the PyTorch-based MONAI framework (https://monai.io) was used.

SegResNet [11] is a U-Net style network with ResNet blocks with larger encoders but smaller decoders. We trained a 3D region-based SegResNet using the Auto3DSeg tool provided by MONAI. Depending on the hardware, the batch size ranged from 2 to 5, with each batch containing a patch of $224\times224\times144$ voxels. We used a combined loss using Dice loss and cross-entropy loss, optimized by AdamW with an initial learning rate of 0.0002 and weight decay of 1e-05. Each fold underwent training for 100 epochs on the same cluster as mentioned for the Swin UNETR training.

2.3. Model Ensembling

A model ensemble strategy is proposed to improve the segmentation model's accuracy and robustness. This method (Fig.1) harnesses the benefits inherent to both convolutional neural networks and vision transformers. We aggregated the outputs of each model trained on every fold using probability maps obtained by applying the respective non-linear activation functions. The ensemble was conducted label-wise, allowing each region to be ensembled with different folds and models based on the validation performance of individual models trained on a fold. In practice, we initially ensembled the five folds of each model, followed by the label-wise ensemble across models.

2.4. Adaptive Post-processing

We applied an adaptive post-processing to the ensembled predictions (Fig.2). First, we extracted 386 radiomic features using the PyRadiomics package [15] and implementation in [16]. The features are divided into 14 shape-based features and 93 intensity-based features for each MRI sequence. A k-means clustering algorithm was used to group cases into different clusters of tumors according to the radiomic features. An optimal number of clusters and features was determined using a grid search. This k-means algorithm was trained on the training set images along with the corresponding ground-truth masks provided. Second, we carried out an optimal threshold search to remove small disconnected components, reducing the number of false positives in the segmentation maps. Finally, we carried out another threshold search over these last refined segmentation maps for ET and ED label redefinition based on ET/WT and ED/WT ratios. In this last step, for example, if the ET/WT ratio fell below a certain threshold, the ET label would be redefined to either NCR or ED, corresponding to the TC region.

3. RESULTS AND DISCUSSION

Table 1 presents a detailed summary of the performance evaluation of our models on the validation dataset. This evaluation was conducted automatically through the challenge's digital platform, without access to the validation ground truth data. Fig. 4 depicts qualitative results for selected cases based on the median of lesion-wise Dice. The average inference time per case was 152 seconds on a single NVIDIA A5000 24GB GPU.

Overall, combining ensemble models and post-processing improved the generalizability of the models' performance. The trained SegResNet did not show good generalizability for ET and TC but is robust for WT. Hence, only the WT predicted by SegResNet was ensembled. Despite the consistent labeling of the three tumor sub-regions (ET, TC, WT) across all tasks, the morphology and spatial location of the lesions vary among subjects and tasks. Hence, implementing a robust post-processing step is crucial to enhance the accuracy of predictions and ensure the output segmentations are consistent. As demonstrated in section 3, the adaptive post-processing significantly improved lesion-wise metrics. This approach enabled tumor-specific post-processing, thereby improving segmentation for each tumor type whether present or not in the training data, aligning with the generalizability aspect of this challenge.

4. CONCLUSION

Generalizing tumor segmentation algorithms across diverse tumor types and age groups poses a significant challenge within the realm of deep

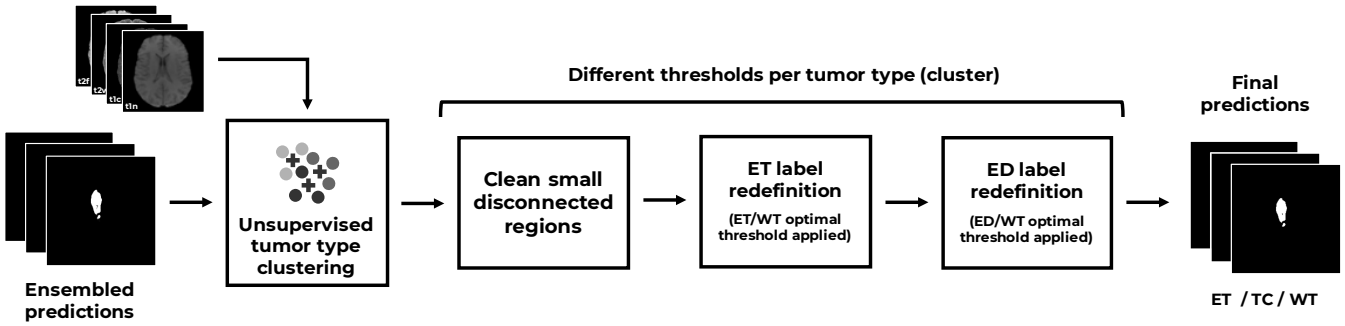

Fig. 2: Adaptive post-processing strategy. The ensemble predictions first underwent a tumor-type clustering using a k-means algorithm (k=5). Then, cleaned of small disconnected regions. Finally, ET and ED labels were redefined based on ET/WT and ED/WT thresholds, respectively.

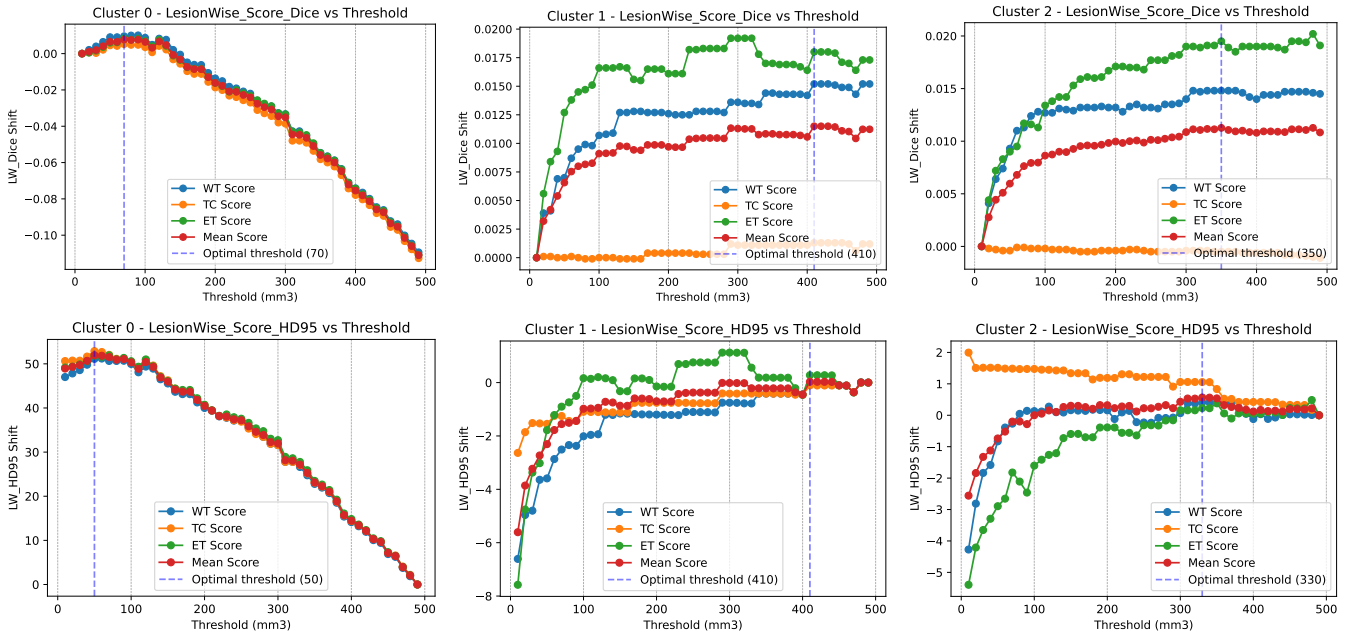

Fig. 3: Threshold search using cross-validation is conducted to find the optimal threshold for removing small disconnected regions within each cluster. LW refers to lesion-wise metrics. The change in metrics is computed and shown according to different thresholds.

learning techniques. In this study, we introduced an ensemble strategy followed by an adaptive post-processing approach leveraging k-means clustering that helped improve the performance of the model ensemble and the generalizability of deep learning models. This method aims to enhance the efficacy of model ensembles and bolster the general applicability of deep learning models in tumor segmentation tasks.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data available in open access as part of the Brain Tumor Segmentation (BraTS) Challenge project through Synapse ID (syn51156910).

6. ACKNOWLEDGMENTS

Partial support for this work was provided by the National Cancer Institute (UG3 CA236536) and by the Spanish Ministerio de Ciencia e Innovación, the Agencia Estatal de Investigación and NextGenerationEU funds, under grants PDC2022-133865-I00 and PID2022-141493OB-I00. The authors gratefully acknowledge the Universidad Politécnica de Madrid (www.upm.es) for providing computing resources on the Magerit Supercomputer.

7. REFERENCES

[1] B. H. Menze, A. Jakab, S. Bauer, and et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[2] Maruf Adewole, Jeffrey D. Rudie, Anu Gbadamosi, and et al., "The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa)," 2023.

[3] Anahita Fathi Kazerooni, Sherjeel Arif, Rachel Madhogarhia, and et al., "Automated tumor segmentation and brain tissue extraction from multiparametric mri of pediatric brain tumors: A

	LW Dice			LW HD95 (mm)		
	ET	TC	WT	ET	TC	WT
nnU-Net (1)	0.76 ± 0.311	0.83 ± 0.242	0.804 ± 0.249	61.87 ± 118.42	34.46 ± 82.29	51.96 ± 94.55
Swin UNETR (2)	0.758 ± 0.3	0.808 ± 0.263	0.78 ± 0.262	61.01 ± 110.28	41.96 ± 89.29	61.93 ± 100.01
SegResNet (3)	0.429 ± 0.317	0.525 ± 0.292	0.826 ± 0.231	89.24 ± 114.93	75.62 ± 99.96	43.07 ± 86.59
Ensemble2 (1+2)	0.772 ± 0.304	0.844 ± 0.232	0.811 ± 0.241	57.19 ± 113.90	29.77 ± 76.29	49.21 ± 91.15
Ensemble3 (1+2+3)	$\mathbf{0.772 \pm 0.304}$	$\mathbf{0.844 \pm 0.232}$	$\mathbf{0.841 \pm 0.216}$	$\mathbf{57.19 \pm 113.90}$	$\mathbf{29.77 \pm 76.29}$	$\mathbf{37.72 \pm 81.66}$
Ensemble2+PP	0.842 ± 0.227	0.854 ± 0.217	0.872 ± 0.187	29.46 ± 83.98	24.68 ± 71.99	25.14 ± 67.59
Ensemble3+PP	0.842 ± 0.227	0.854 ± 0.217	0.872 ± 0.182	29.46 ± 83.98	24.67 ± 71.99	25.22 ± 67.58

Table 1: Quantitative results on 360 validation cases: lesion-wise (LW) Dice coefficients and 95% Hausdorff distances (HD95) were computed for enhancing tumor (ET), tumor core (TC), and whole tumor (WT), respectively. Each value represents the mean $\pm$ standard deviation. PP refers to post-processing by removing small disconnected components.

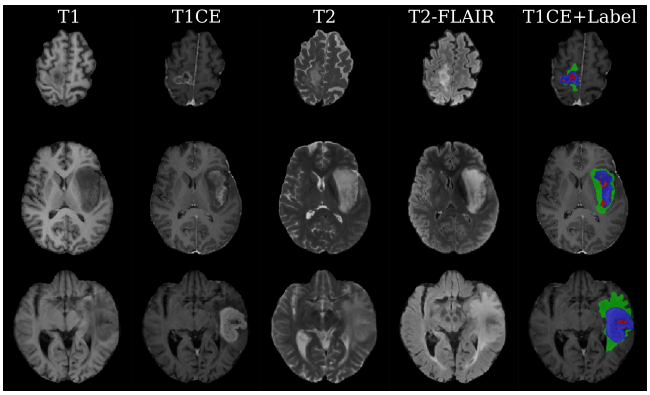

Fig. 4: Qualitative results on the validation dataset. The selected cases show the median of the averaged LW Dice over three tumor regions within each predicted cluster (NCR-red, ED-green, ET-blue). From top to bottom: cluster=0, ET=0.931, TC=0.95, WT=0.88; cluster=1, ET=0.887, TC=0.928, WT=0.98; cluster=2, ET=0.936, TC=0.924, WT=0.908.

multi-institutional study," *Neuro-Oncology Advances*, vol. 5, no. 1, pp. vdad027, 2023.

[4] A. Karargyris, R. Umeton, M.J. Sheller, and et al., "Federated benchmarking of medical artificial intelligence with medperf.," *Nat Mach Intell*, vol. 5, pp. 799–810, 2023.

[5] Florian Kofler, Felix Meissen, Felix Steinbauer, et al., "The Brain Tumor Segmentation (BraTS) Challenge 2023: Local Synthesis of Healthy Brain Tissue via Inpainting," 2023.

[6] Hongwei Bran Li, Gian Marco Conte, Syed Muhammad Anwar, and et al., "The Brain Tumor Segmentation (BraTS) Challenge 2023: Brain MR Image Synthesis for Tumor Segmentation (BraSyn)," 2023.

[7] Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum, "The SRI24 multichannel atlas of normal adult human brain structure," *Human brain mapping*, vol. 31, no. 5, pp. 798–819, 2010.

[8] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, and et al., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[9] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, and et al., "Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20730–20740.

[10] A. Hatamizadeh, V. Nath, Y. Tang, and et al., "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds. 2022, pp. 272–284, Springer, Cham.

[11] Andriy Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*. Springer, 2019, pp. 311–320.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[13] A. Hatamizadeh, Y. Tang, V. Nath, and et al., "UNETR: Transformers for 3D Medical Image Segmentation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 574–584.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] J.J.M. van Griethuysen et al., "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[16] Z. Jiang, A. Parida, S.M. Anwar, Y. Tang, H.R. Roth, M.J. Fisher, R.J. Packer, R.A. Avery, and M.G. Linguraru, "Automatic Visual Acuity Loss Prediction in Children with Optic Pathway Gliomas using Magnetic Resonance Imaging," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. July 2023, pp. 1–5, IEEE.