Maastricht University

HNT-AI

Citation for published version (APA):

Salahuddin, Z., Chen, Y., Zhong, X., Rad, N. M., Woodruff, H. C., & Lambin, P. (2023). HNT-AI: An Automatic Segmentation Framework for Head and Neck Primary Tumors and Lymph Nodes in FDG-PET/CT Images. In V. Andrearczyk, V. Oreiller, A. Depeursinge, & M. Hatt (Eds.), *Head and Neck Tumor Segmentation and Outcome Prediction: 3rd Challenge, HECKTOR 2022, Held in Conjunction with MICCAI 2022, Proceedings* (3 ed., Vol. 13626 LNCS, pp. 212-220). Springer Verlag. https://doi.org/10.1007/978-3-031-27420-6_21

Document status and date:
Published: 01/01/2023

DOI:
10.1007/978-3-031-27420-6_21

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 15 May. 2024

HNT-AI: An Automatic Segmentation Framework for Head and Neck Primary Tumors and Lymph Nodes in FDG-PET/CT Images

Zohaib Salahuddin[1(✉)], Yi Chen[1,2], Xian Zhong[1,3], Nastaran Mohammadian Rad[1], Henry C. Woodruff[1,4], and Philippe Lambin[1,4]

[1] The D-Lab, Department of Precision Medicine, GROW-School for Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands
z.salahuddin@maastrichtuniversity.nl
[2] Key Laboratory of Intelligent Medical Image Analysis and Precise Diagnosis, College of Computer Science and Technology, Guizhou University, Guiyang, China
[3] Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[4] Department of Radiology and Nuclear Medicine, GROW-School for Oncology, Maastricht University Medical Center, Maastricht, The Netherlands

Abstract. Head and neck cancer is one of the most prevalent cancers in the world. Automatic delineation of primary tumors and lymph nodes is important for cancer diagnosis and treatment. In this paper, we develop a deep learning-based model for automatic tumor segmentation, HNT-AI, using PET/CT images provided by the MICCAI 2022 Head and Neck Tumor (HECKTOR) segmentation Challenge. We investigate the effect of residual blocks, squeeze-and-excitation normalization, and grid-attention gates on the performance of 3D-UNET. We project the predicted masks on the z-axis and apply k-means clustering to reduce the number of false positive predictions. Our proposed HNT-AI segmentation framework achieves an aggregated dice score of 0.774 and 0.759 for primary tumors and lymph nodes, respectively, on the unseen external test set. Qualitative analysis of the predicted segmentation masks shows that the predicted segmentation mask tends to follow the high standardized uptake value (SUV) area on the PET scans more closely than the ground truth masks. The largest tumor volume, the larget lymph node volume, and the total number of lymph nodes derived from the segmentation proved to be potential biomarkers for recurrence-free survival with a C-index of 0.627 on the test set.

Team Name: The_DLab

Keywords: 3D UNet · Grid-attention · Residual networks · Squeeze-and-excitation · Segmentation biomarkers

Y. Chen and X. Zhong—These authors contributed equally as second authors.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
V. Andrearczyk et al. (Eds.): HECKTOR 2022, LNCS 13626, pp. 212–220, 2023.
https://doi.org/10.1007/978-3-031-27420-6_21

1 Introduction

Head and Neck (H&N) cancers are among the most common cancers worldwide, including a group of tumors arising in the lip, oral cavity, pharynx, larynx, and paranasal sinuses [1]. The prognosis of H&N cancers varies greatly with 5-year survival ranging from 85.1% in patients with the localized disease to 40.1% in those with distant disease [2]. This variance in outcomes emphasizes the importance of accurate and timely diagnosis and staging, where imaging plays a crucial role. ^{18}F-FluoroDeoxyGlucose (FDG)-Positron Emission Tomography (PET) and Computed Tomography (CT) are two important imaging modalities for the initial staging and follow-up of H&N cancers. Nowadays, there is an increasing need for the development of automatic segmentation algorithms for H&N tumor diagnosis and staging, gross tumor volume (GTV) delineations in radiotherapy planning, as well as establishment and validation of radiomics models. PET and CT focused on metabolic and morphological characteristics, respectively, which may provide complementary and synergistic information for cancerous lesion segmentation as well as tumor characteristics predictive for patient outcome. The manual delineation of tumors in head and neck cancer suffers from inter-observer variability [3]. Hence, it is essential to develop automatic segmentation tools for H&N cancers that are accurate, fast, robust, and reproducible.

Deep neural networks are becoming popular due to the increasing amount of available data and computational resources. Convolutional Neural Networks (CNN) constitute a class of deep neural networks that have demonstrated state-of-the-art performance on a variety of medical image segmentation challenges, e.g., multi-center, multi-vendor, and multi-disease cardiac segmentation (M&Ms) challenge [4], multi-modality whole heart segmentation [5], and auto segmentation for thoracic radiation treatment planning [6]. No-new-UNet (nnUNet) has emerged as a self-configuring method that makes automatic design choices related to pre-processing, network architecture, and hyper-parameter tuning. It has demonstrated state-of-the-art performance in many medical image segmentation challenges [7].

The first and second editions of HEad and neCK TumOR (Hecktor) segmentation and outcome prediction challenge were held in 2020 and 2021 [8] [11]. These challenges aim to provide multi-centric data and use a standardized evaluation criteria to develop and validate automatic segmentation tools for H&N primary tumors in FDG-PET and CT images. The new edition of Hecktor 2022 extends Hecktor 2021 challenge by adding the additional task of lymph nodes segmentation (GTVn) along with the primary tumor (GTVp) segmentation, and it also provides data from three additional centers. In Hecktor 2021, a well-tuned 3D nn-UNet with squeeze-and-excitation normalization [13] demonstrated the best segmentation performance on the test set [12].

In this paper, we propose an HNT-AI segmentation framework based on nnUNet that incorporates residual blocks, squeeze-and-excitation channel-wise attention, and grid-attention gates. We also investigate if the primary tumor

and lymph node segmentations produced by the proposed algorithm can serve as valuable biomarkers for recurrence-free survival prediction.

2 Material and Methods

2.1 Data

The training dataset for GTVp and GTVn segmentation comes from 7 different centers and consists of 524 training cases. A CT image and a corresponding registered FDG-PET image are provided for each case. The clinical information about each patient, including center, gender, age, tobacco and alcohol consumption, performance status (Zubrod), and HPV status is provided for 489 training cases. Additionally, the test dataset for evaluating the segmentation performance consists of 359 cases and it comes from 3 centers. Training cases from two of these centers are present in the training dataset.

2.2 Data Preprocessing

The CT images and the corresponding PET images are first resampled to the maximal bounding box covered by the field of view of both modalities. The CT images have a higher image resolution than the corresponding PET images. The median resolution of CT images in the training set was $0.98 \times 0.98 \times 3.27$ mm^3, and the mean resolution of PET images in the training set was $4.26 \times 4.26 \times 3.27$ mm^3. We resample both the CT and the PET images to the resolution of $1 \times 1 \times 3$ mm^3. The images are resampled with spline interpolation, and the corresponding segmentations are resampled using nearest-neighbor interpolation. We clip the intensity values of CT images at 0.5 and 99.5 percentiles. We apply z-score normalization on the CT images and min-max normalization on the PET images.

2.3 Network Architecture

The proposed architecture is shown in Fig. 1. This architecture is a modified adaption of the nnUNet architecture [7]. Each encoder block comprises of residual skip connections [14]. Each convolutional block consists of $3 \times 3 \times 3$ convolutional layer following by drop-out layer with $p = 0.5$, instance normalization and LeakyReLU activation function. Squeeze-and-excitation channel-wise attention mechanism is employed at each layer to learn important features by fusing spatial and channel-wise features [15]. Grid-attention gates enable the network to identify spatially important areas in the network and consequently aid in false positive reduction [16]. Grid attention gates are incorporated at each skip connection.

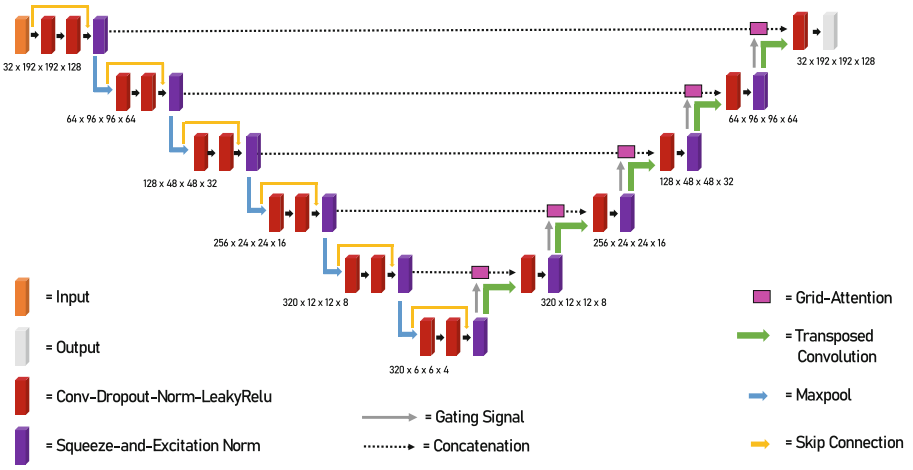

Fig. 1. The proposed architecture for 3D-UNet with skip connections, squeeze-and-excitation channel-wise attention mechanism and grid-attention gates.

2.4 Experimental Settings

We used a single RTX 3090 Graphics Card for training. The batch size was set at 2. Adam optimizer was used with an initial learning rate of 10^{-3}. Cosine annealing scheduler was used to reduce the learning rate from 10^{-3} to 10^{-6} every 30 epochs. The model was trained for 250 epochs. The training process took 9 h to complete. The patch size was set to be $192 \times 192 \times 128$. Data augmentation comprising of random rotation ($-15°$ to $+15°$), random scaling (0.85 to 1.15), elastic deformation, addition of Gaussian noise (0 to 0.1), mirroring, and gamma correction (0.75 to 1.25) was used to avoid overfitting. Five-fold cross-validation was used to find the best network architecture and for hyperparameter tuning.

2.5 Loss Function

The loss function comprised of multi-class dice loss [17] and cross entropy loss. The network is trained with deep supervision; thus, the computation of loss function occurs at each decoder block. This allows the training to occur at each layer. The loss computed at each decoder block is assigned a decreasing weight from higher resolution to lower resolution.

2.6 Post Processing

We observe false positive predictions for GTVp and GTVn outside the area of interest in the brain and the lung region. Furthermore, we observed small false positive predictions that occur at a distance from the correct ground truth prediction. To reduce the false positives, we projected the predicted binary mask

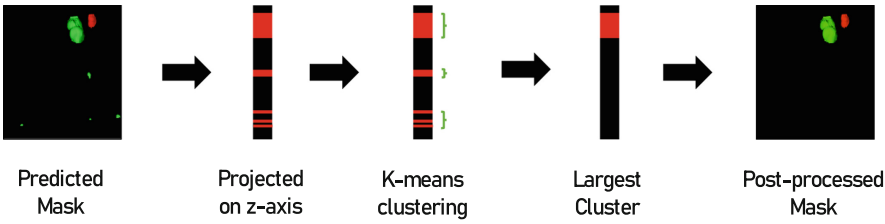

Fig. 2. Post-processing based on k-means clustering.

onto the z-axis, and all the indices that contain a prediction are set to one. K-means clustering is then applied to find the cluster with the largest cumulative volume [9]. The indices on the z-axis that correspond to other clusters are set to zero. The predicted binary mask slices that correspond to the largest cluster are left remaining. The workflow for the post-processing is shown in Fig. 2.

2.7 Evaluation Metric

Aggregate dice score (DSC_{agg}) is used as an evaluation metric. DSC_{agg} is defined as:

$$DSC_{agg} = \frac{\sum_{i}^{N}\sum_{k} \hat{y_{i,k}} y_{i,k}}{\sum_{i}^{N}\sum_{k}(\hat{y_{i,k}} + y_{i,k})}$$

where $y_{i,k}$ is the ground truth for GTV_p or GTV_n and $\hat{y_{i,k}}$ is the corresponding prediction. This metric is used because some images do not contain any tumor or lymph node. The intersections and unions for all the images are accumulated and then DSC_{agg} is calculated at the end.

3 Results

3.1 Five-Fold Cross-Validation

Table 1 shows the performance of modified 3D-UNet architecture for five-fold cross-validation. We started with nnUNet as a baseline. The proposed model that contains residual skip connections, squeeze-and-excitation channel-wise attention mechanism and grid attention gates achieves the best performances. Figure 3 shows the qualitative performance of the proposed algorithm on an image (CHUV-021) from the validation set. As shown, the predicted segmentation mask closely follows the boundary of the area with high SUV intake in the PET images compared to the ground truth label.

Table 1. Aggregated Dice DSC_{agg} for GTV_n and GT_n for five-fold cross validation.

Model	DSC_{agg} GTV_p	DSC_{agg} GTV_n	$Mean$ DSC_{agg}
nnUNet	0.752 ± 0.471	0.731 ± 0.621	0.742 ± 0.546
3D UNet + Resnet	0.751 ± 0.508	0.744 ± 0.419	0.748 ± 0.463
3D UNet + SE	0.768 ± 0.398	0.759 ± 0.507	0.764 ± 0.452
Proposed	$\mathbf{0.777 \pm 0.351}$	$\mathbf{0.768 \pm 0.522}$	$\mathbf{0.773 \pm 0.437}$

Table 2. Aggregated Dice DSC_{agg} for GTV_n and GT_n of the three attempts on the test set

Model	DSC_{agg} GTV_p	DSC_{agg} GTV_n	$Mean$ DSC_{agg}
nnUNet	0.756	0.729	0.743
3D UNET (Resnet + SE + Grid-Attention)	0.768	0.758	0.763
Ensemble	0.774	0.758	0.767

3.2 Test Set

Table 2 shows the result of our three attempts on the unseen test set. In the first attempt, we benchmarked the performance of nnUNet. For the second submission, 90% of the dataset was used for training and 10% of the dataset was used for validation. The proposed 3D UNet model with residual skip connections, squeeze-and-excitation channel-wise attention and grid attention gates was trained for 1000 epochs. For the final attempt, we combined the model used in attempt 2 with the five other models from five-fold cross-validation. The ensemble model was obtained by taking an average of the soft-max probabilities of the six models.

3.3 Tumor and Lymph Node Volume as a Biomarker for Prognosis

The prognostic value of tumor and lymph node volume derived from the segmentation are summarized in Table 3. Five-fold cross-validation shows that the model combining the largest tumor volume, largest lymph node volume and number of lymph nodes leads to the highest C-index of 0.597 ± 0.056 for recurrence-free survival (RFS) prediction. The model showed a C-index of 0.627 in the test set.

4 Discussion

The proposed 3D-UNet architecture based on residual skip connections, squeeze-and-excitation channel-wise attention, and grid-attention gates outperformed other network configurations during five-fold cross-validation. The proposed algorithm achieved DSC_{agg} GTV_p score of 0.777 ± 0.351 and DSC_{agg} GTV_n score of 0.768 ± 0.522 during five-fold cross-validation. On the unseen test set, the algorithm obtained DSC_{agg} GTV_p score of 0.768 and DSC_{agg} GTV_p score of

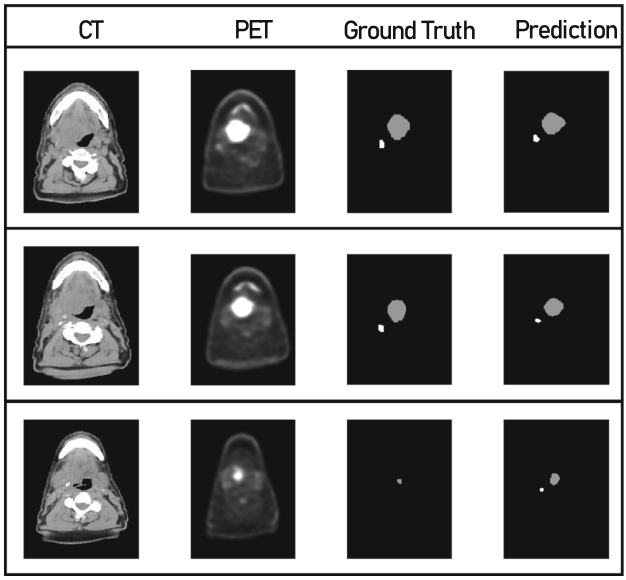

Fig. 3. Qualitative analysis of the predicted GTV_p and GTV_n predictions. The predicted mask tends to follow the areas of high SUV uptake more closely as compared to the ground truth labels.

Table 3. Prognostic value of tumor and lymph node volume for prediction of recurrence-free survival

Variables in the model	C-index ± SD
Five-fold Cross-Validation Results	
Largest Tumor Volume	0.565 ± 0.069
Largest Lymph Node Volume	0.534 ± 0.083
Number of Lymph Nodes	0.564 ± 0.033
Largest Tumor Volume + Largest Lymph Node Volume	0.570 ± 0.059
Largest Tumor Volume + Largest Lymph Node Volume + Number of Lymph Nodes	$\mathbf{0.597 \pm 0.056}$
Test Set Results	
Largest Tumor Volume + Largest Lymph Node Volume + Number of Lymph Nodes	0.627

0.758. Furthermore, the largest segmented tumor volume, the largest segmented lymph node volume and the total number of segmented lymph nodes appeared to be promising biomarkers for recurrence-free survival. These three features obtained a C-index of 0.627 on the external test set for RFS prediction. The proposed model demonstrates superior performance on the external test set as compared to five-fold cross-validation. This could not be investigated due to the absence of labels for the external test set.

This study has some limitations. It is important to determine that the developed algorithms are fair with respect to age, gender, and other biases. Therefore, metrics should also be calculated for these subgroups to report any biases. We observed qualitatively that in some of the cases, the proposed segmentation algorithm performs a better delineation. Therefore, an *in silico* trial that can obtain a preference score to determine whether the radiologists prefer automatic segmentation or manual delineations needs to be conducted [18]. We also observed that false positives occur in areas where there is no probability of the tumor occurring. The segmentation algorithm can be made more interpretable by including an anatomical prior that calculates the probability of tumor occurrence at each pixel value [19]. Furthermore, we also need to estimate uncertainty to avoid silent failures on out-of-distribution cases [20].

5 Conclusion

In this paper, we proposed a segmentation framework demonstrating promising performance for segmenting primary tumors and lymph nodes in head and neck cancer. We also validated the quality of the segmentation by using the largest segmented tumor and lymph node volumes as biomarkers for recurrence-free survival prediction. In the future, we need to increase interpretability by incorporating anatomical priors, estimating uncertainty, and calculating the fairness of the algorithm by evaluating the algorithm with respect to biases.

Acknowledgements. We acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), ERC-2020-PoC: 957565-AUTO. DISTINCT. Authors also acknowledge financial support from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement: ImmunoSABR n° 733008, MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103, JTI-IMI2-2020-23-two-stage IMI-OPTIMA n° 101034347, and TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295). This work was supported by the Dutch Cancer Society (KWF Kankerbestrijding), Project number: 14449/2021-PoC. This study has received funding from the National Natural Science Foundation of China (No. 92059201). This study has received funding from the China Scholarship Council (No. 202006675008).

References

1. Pfister, D.G., et al.: Head and neck cancers, version 2.2020, NCCN clinical practice guidelines in oncology. J. Nat. Compr. Cancer Netw. **18**(7), 873–898 (2020)
2. Marcus, C., et al.: PET imaging for head and neck cancers. Radiol. Clin. **59**(5), 773–788 (2021)
3. van der Veen, J., Gulyban, A., Nuyts, S.: Interobserver variability in delineation of target volumes in head and neck cancer. Radiother. Oncol. **137**, 9–15 (2019)
4. Campello, V.M., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans. Med. Imaging **40**(12), 3543–3554 (2021)

5. Zhuang, X., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med. Image Anal. **58**, 101537 (2019)
6. Yang, J., et al.: Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. Med. Phys. **45**(10), 4568–4581 (2018)
7. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
8. Oreiller, V., et al.: Head and neck tumor segmentation in PET/CT: the HECKTOR challenge. Med. Image Anal. **77**, 102336 (2022)
9. Kodinariya, T.M., Makwana, P.R.: Review on determining number of Cluster in K-means clustering. Int. J. **1**(6), 90–95 (2013)
10. Andrearczyk, V., et al.: Overview of the HECKTOR challenge at MICCAI 2022: automatic head and neck tumor segmentation and outcome prediction in PET/CT. In: Head and Neck Tumor Segmentation and Outcome Prediction (2023)
11. Andrearczyk, V., et al.: Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images. 3D Head and Neck Tumor Segmentation in PET/CT Challenge. Springer, Cham (2021)
12. Xie, J., Peng, Y.: The head and neck tumor segmentation based on 3D U-Net. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) HECKTOR 2021. LNCS, vol. 13209, pp. 92–98. Springer, Cham (2022)
13. Iantsen, A., Visvikis, D., Hatt, M.: Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images. In: Andrearczyk, V., Oreiller, V., Depeursinge, A. (eds.) HECKTOR 2020. LNCS, vol. 12603, pp. 37–43. Springer, Cham (2021)
14. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
16. Schlemper, J., et al.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53**, 197–207 (2019)
17. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
18. Primakov, S.P., et al.: Automated detection and segmentation of non-small cell lung cancer computed tomography images. Nat. Commun. **13**(1), 1–12 (2022)
19. Salahuddin, Z., et al.: Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput. Biol. Med. **140**, 105111 (2022)
20. Mehrtash, A., et al.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans. Med. Imaging **39**(12), 3868–3878 (2020)