DOMINO: Domain-aware Model Calibration in Medical Image Segmentation

Skylar E. Stolte[1], Kyle Volle[2,3], Aprinda Indahlastari[4,6,7], Alejandro Albizu[4,5], Adam J. Woods[4,5,6,7], Kevin Brink[8], Matthew Hale[2], and Ruogu Fang[1,4,6]

[1] J. Crayton Pruitt Family Department of Biomedical Engineering, Herbert Wertheim College of Engineering, University of Florida (UF), USA
[2] Department of Mechanical and Aerospace Engineering, Herbert Wertheim College of Engineering, UF, USA
[3] UF Innovation Station at the Research & Engineering Education Facility (REEF), Herbert Wertheim College of Engineering, UF, USA
[4] Center for Cognitive Aging and Memory, McKnight Brain Institute, UF, USA
[5] Department of Neuroscience, College of Medicine, UF, USA
[6] Department of Electrical and Computer Engineering, Herbert Wertheim College of Engineering, UF, USA
[7] Department of Clinical and Health Psychology, College of Public Health and Health Professions, UF, USA
[8] United States Air Force Research Laboratory, Fort Walton Beach, Florida, USA

Abstract. Model calibration measures the agreement between the predicted probability estimates and the true correctness likelihood. Proper model calibration is vital for high-risk applications. Unfortunately, modern deep neural networks are poorly calibrated, compromising trustworthiness and interpretability. Medical image segmentation particularly suffers from this due to the natural uncertainty of tissue boundaries. This is exasperated by their loss functions, which favor overconfidence in majority classes. We address these challenges with DOMINO, a domain-aware model calibration method which leverages the semantic confusability and hierarchical similarity between class labels. Our experiments demonstrate that our DOMINO-calibrated deep neural networks outperform non-calibrated models and state-of-the-art morphometric methods in head image segmentation. Our results show that that our method can consistently achieve better calibration, higher accuracy, and faster inference times than these methods, especially on rarer classes. This performance is attributed to our domain-aware regularization to inform semantic model calibration. These findings show the importance of semantic ties between class labels in building confidence into deep learning models. The framework has the potential to improve the trustworthiness and reliability of generic medical image segmentation models.

Keywords: Image Segmentation · Machine Learning Uncertainty · Model Calibration · Model Generalizability · Whole Head MRI

1 Introduction

Machine learning calibration measures the agreement between the predicted probability estimates and the true correctness likelihood [9]. Proper calibration is vital for high-risk applications. Modern deep neural networks (DNNs) achieve impressive accuracy at poor calibration [9]. Incorrectly calibrated models are unreliable on out-of-distribution data and don't know when they are likely to be incorrect. This discrepancy leaves them vulnerable in critical decision making scenarios such as self-driving cars, surgical robots, disease subtyping, and personalized intervention. On the other hand, well-calibrated models are less certain when decisions are incorrect and comparably certain when correct. Their reliable confidence provides valuable information to establish trustworthiness.

We hypothesize that domain-aware model calibration that leverages the *semantic confusability* and *hierarchical similarity* among class labels can yield well-calibrated and higher performing models. Our approach harnesses the inherent similiarities between labels. To test this hypothesis, we have chosen medical image segmentation because of its fundamental role in medical image analysis. Prior works have found that overly-confident decisions on tissue boundaries can introduce significant errors in brain volume estimations [4]. Head image segmentation is prone to errors due to delicate tissue boundaries, tissue imbalance, and low contrast. These challenges can make widely-adopted open-source software fall short on high-fidelity tasks in patient sub-populations [13,18,3]. Errors in head segmentation can lead to downstream errors in clinical pipelines, like in parameter estimation for non-invasive brain stimulation [2,12]. Hence, we seek to answer the question: can we leverage class relationships to build an automated, high-performing, and well-calibrated head image segmentation model?

We answer this question with DOMINO, a framework that leverages domain information among class labels to calibrate DNNs. Unlike prior works that push class means to be orthogonal [16], we assume some class labels have natural similarity. The choice of loss function is very important to calibration because loss drives how a model learns and performs [17]. Medical image segmentation still largely relies on standard loss functions [1]. We extend these approaches with domain-aware loss regularization to improve model calibration. We study two regularization schemes that are based on confusion matrices (CM) and hierarchical classes (HC), respectively. The former imposes a penalty based on class confusability when using a standard network on a held-out data subset. The latter groups labels into hierarchical classes based on common tissue properties.

Our experiments in Section 3 demonstrate that our DOMINO-calibrated DNNs outperform non-calibrated models and state-of-the-art morphometric methods (e.g., Headreco) in head segmentation. Our method can consistently achieve better calibration (Figure 6) and higher accuracy (Tables 2, 3), especially on rarer classes. This performance is attributed to our domain-aware regularization to inform semantic model calibration. These findings show the importance of semantic ties between class labels in building confidence into DNNs. DOMINO-calibrated models improve model trustworthiness and generalizability. This system could potentiate efficient cross-talk between human and machine [8].

2 Domain-aware Model Calibration

2.1 U-Net Transformers (UNETR) Model

We employ UNETR [10] as our base model due to its superior performance on medical image segmentation. UNETR utilizes a U-Net architecture with a transformer encoder. This approach combats the relative locality of convolutional layers in fully convolutional networks (FCNs). Transformers have revolutionized Natural Language Processing due to superior learning of long-range sequences [19]. Transformers differ from FCNs in that the former encodes images as a sequence of one-dimensional patch embeddings. Self-attention modules learn weighted sums of values that are calculated from hidden layers. Hence, UNETR reformulates 3D image segmentations as sequence-to-sequence predictions. The network passes the transformer's learned global contextual information to a traditional FCN decoder via skip connections at different resolutions. The decoder concatenates localized information with the global multi-scale information from the encoder. Non-regularized UNETR is referred to as UNETR-Base.

2.2 Domain-aware Loss Regularization

Concept Our penalty term addresses a deficit with cross entropy (CE) loss in uncertainty. CE loss maximizes the output associated with the ground truth label class. Due to this, the network tries to increase the true label logit more than the incorrect label logits. This results in networks that are overly confident in the predicted class. On the other hand, the non-selected classes' softmax outputs do not represent the true likelihood distribution.

Our work introduces more meaningful uncertainty into our deep learning model by penalizing incorrect classes. Specifically, we assume that some classes are naturally more similar to others. Prior work shows that network presentation often pushes class means to all be orthogonal to each other [16]. This means that the network assumes that all classes are equally separable. This assumption fights the natural similarities between certain classes. Thus, we hypothesize that a network can learn better class representation by taking advantage of *class similarities*, rather than fighting them. The methods described in this section are applicable to classification and segmentation. In this definition, segmentation is considered analogous to pixel-wise classification in uncertainty problems [14].

Derivation Our regularization term adds to any loss function as follows:

$$\mathcal{L}(y, \hat{y}) + \beta(y')(W)(\hat{y}) \tag{1}$$

where $\mathcal{L}$ is a suitable loss function (here, we use $DiceCE$ which is a combination of Dice score and cross entropy), y is the one-hot encoded true label, and $\hat{y}$ is the softmax output. β can take on any value between zero and one. W represents a generic regularization term of size $N \times N$, where N is the number of classes. The diagonals are zero, whereas the off-diagonals represent the penalties for confusing classes. We propose two domain-aware approaches to design W as below.

Confusion Matrix-based (UNETR-CM) Confusion matrix-based calibration utilizes the natural confusability among class labels using a non-calibrated DNN. First, we train UNETR-base without regularization on the training set.

Then, we evaluate the trained model on a held-out validation set to generated a confusion matrix on all classes. The loss regularization is computed as below:

$$W_{ij} = S \cdot \frac{I_i - C_{ij}}{Ii} \tag{2}$$

Here, i and j represent the row and column indices, respectively. C is the confusion matrix generated when UNETR-Base is applied on a held-out validation set and normalized by class prevalence. W_{ij} represents any given matrix entity. I_i is i^{th} row of the identity matrix. Thus $W_{ii} = 0$ so there is no penalty for the correct class. Finally, S is a scaling factor to make the regularization weights more significant. We set $S = 3$ based on empirical experiments; however, jointly varying β and S can change the balance of the loss function. Low values for both result in no regularization; too high and it begins to effect model accuracy. The correct values for these hyperparameters will depend on the model and dataset.

Hierarchical Class-based (UNETR-HC) Here, We propose a regularization that leverages the hierarchical relationship between semantic labels. Hierarchical groups are more likely to have similar properties than inter-group classes. Hence, confusing within hierarchical groups can allow researchers and clinicians to make more informed and safer mistakes when wrong. Table 1 shows the hierarchy for our head segmentation. We define the matrix penalty shown in Fig 1b by considering which classes are subsets of the same super-class. In Fig 1b, each row represents the penalties for confusing the given class with any other class. The maximum penalty is 3, and penalties are manually lowered within the groups of table 1. The eye class is considered close to two groupings. This method of generating the matrix penalty is more subjective than UNETR-CM, but it incorporates domain knowledge.

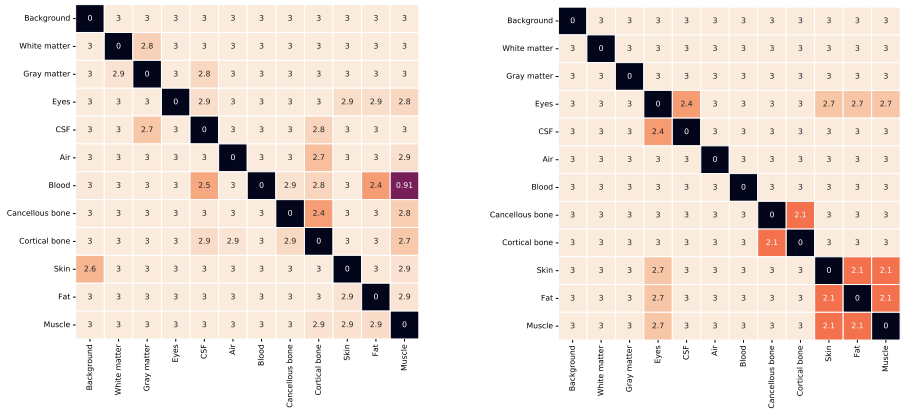
Table 1: **Hierarchical class groupings.** *Eyes are considered to fall within CSF and soft tissue due to have aqueous and fibrous components.

Hierarchical groupings	***Tissues***
Background (BG)	BG
White matter (WM)	WM
Grey Matter (GM)	GM
Cerebrospinal fluid (CSF)	CSF, Eyes*
Bone	Cancellous bone, Cortical bone
Soft tissue	Skin, Fat, Muscle, Eyes*
Air	Air
Major artery (Blood)	Blood

3 Experiments and Results

3.1 Dataset

This study harnesses data from a Phase III clinical trial on cognitive training and non-invasive brain stimulation for cognitive improvements. The study recruited a large participant group within the age range of 65 to 89 years and with signs of age-related cognitive decline. The trial was approved by the Institutional Review Boards at all involved institutions. Structural T1-weighted MRI scans were obtained using a 32-channel, receive-only head coil from a 3-T Siemens MAGNETOM Prisma MRI scanner. MPRAGE sequence parameters: repetition time (TR) = 1800 ms; echo time (TE) = 2.26 ms; flip angle = 8°; field of view (FOV) = 256 × 256 × 256 mm; voxel size = 1 mm^3.



(a) Penalty matrix for UNETR-CM (b) Penalty matrix for UNETR-HC

Fig. 1: Computed matrix penalties (W terms) for both experiments

Ground Truth Trained staff segmented the T1 MRIs into 11 tissues using semi-automated segmentation. These 11 tissues included muscle, fat, skin, cortical bone, cancellous bone, majory artery (blood), air, cerebrospinal fluid (CSF), eyes, grey matter (GM), and white matter (WM). Semi-automated segmentation consists of automated segmentation followed by manual correction. First, base segmentations for WM, GM, and bone were obtained using Headreco, while air was generated in the Statistical Parametric Mapping toolbox (SPM12). Next, these automatic outputs were manually corrected using ScanIP Simpleware™ (version 2018.12, Synopsys, Inc., Mountain View, USA). Bone was separated into cancellous and cortical tissue using thresholding and morphology. Blood, skin, fat, muscle, and eyes (sclera and lens) were manually segmented in Simpleware. CSF was generated by subtracting the other ten tissues from the entire head. The resulting 11 tissue masks served as the ground truths for learned segmentation.

Implementation details We implement UNETR using the Medical Open Network for Artificial Intelligence (MONAI-0.8) in Pytorch 1.10.0 [6]. We split our 113 MRIs into 93 training / 10 validation / 10 testing. Each DNN required 1 GPU, 4 CPU, and 30 GB of memory. Each model was trained for 25,000 iterations with evaluation at 500 intervals. The models were trained on 256 x 256 x 256 images with batch sizes of 2 images. We trained our models with Adam optimization using stochatic gradient descent. UNETR segmentation results took 3 seconds per head. Headreco takes roughly 20 minutes per head.

3.2 Evaluation Metrics

We employ the following metrics on the 11-class and 6-class segmentations. We perform the 6-class comparison because the current field standard in head segmentation (e.g., Headreco) provide different tissues than our method. For example, Headreco [15] uses 8 tissues and SPM uses 6 tissues. Thus, we had to combine tissues into groups for a fair comparison.

Dice represent the overlap of two binary masks [5]: $Dice = \frac{2|Y \cap \hat{Y}|}{|Y|+|\hat{Y}|}$ where Y and $\hat{Y}$ represent the ground truth mask and generated mask for a given tissue, respectively. A perfect overlap between these two generates a Dice score of 1, whereas a 0 represents no mask overlap.

Hausdorff distance (Hausdorff) calculates the average distances between the closest points in two data subsets [11,7]. Hausdorff distances are generally more robust than Dice in respect to the precise boundaries.

$$H(Y, \hat{Y}) = max(h(Y, \hat{Y}), h(\hat{Y}, Y)) \tag{3}$$

$$h(Y, \hat{Y}) = \max_{y \in Y}(\min_{\hat{y} \in \hat{Y}}(d(y, \hat{y}))), \quad h(\hat{Y}, Y) = \max_{\hat{y} \in \hat{Y}}(\min_{y \in Y}(d(\hat{y}, y))) \tag{4}$$

where y represents a point in Y and $\hat{y}$ represents a point in $\hat{Y}$. $H(Y, \hat{Y})$ is the overall modified Hausdorff distance, whereas $h(Y, \hat{Y})$ and $h(\hat{Y}, Y)$ are directed Hausdorff distances. $d(y, \hat{y})$ and $d(\hat{y}, y)$ are Euclidean distances. Smaller the Hausdorff distance indicates better segmentation.

Top-N accuracy Top-N accuracy measures how often your true class falls within your top N highest softmax outputs. This metrics reflect meaning in the outputs that were not the selected class. For instance, higher Top-2 and Top-3 predictions can show that a well-calibrated makes reasonable mistakes that are supported by the data, rather than random misclassifications.

Calibration Curves show the relationship between the predicted probability estimates and the true correctness likelihood. These plots are meant for binary classification, so for segmentation one class "positive" is compared to the rest "negative". The prevalence of positive classes is compared to predicted certainty for that class. Perfect calibration is a straight line from origin to (1,1).

3.3 Calibrated models outperform UNETR-Base on 11-classes

Qualitative analysis Figure 2 shows that UNETR-HC best captures the fine detail of the boundary between GM and CSF. This observation is noticeable in the upper left and upper right "grooves" in the light blue (CSF) color. UNETR-HC attempts to tract out these regions and label them as CSF, whereas the UNETR-Base and UNETR-CM assign more of these pixels as GM. This boundary is a major challenge in automatic segmentation due to partial volume effects.

Quantitative comparison Figure 3 and Table 2 show the Dice, Hausdorff, and Top-N. UNETR-CM performs best in Dice and Top-N accuracy, whereas both UNETR-CM and UNETR-HC outperform UNETR-Base in Hausdorff. This insinuates that UNETR-CM classifies the most pixels correctly, whereas both models capture tissue boundaries.

Table 2: Top-N Accuracy on 11 classes

Method	***Top-1***	***Top-2***	***Top-3***
UNETR-Base	0.876	0.979	0.990
UNETR-HC	0.891	0.984	0.993
UNETR-CM	**0.895**	**0.986**	**0.996**

3.4 UNETR with calibration outperforms or performs comparably to Headreco in 6-class segmentation

Qualitative analysis We consolidate classes to compare fairly with state-of-the-art head segmentation such as Headreco [15]. In order to compare the two methods we combine DOMINO classes that are specific subsets of Headreco classes; for example, cancelous and cortical bone are both labeled as just bone. In Fig 4, the three learned models are contrasted with the ground truth labels and the Headreco output. Differences are highlighted with white rectangles. Our methods show comparable or superior performance to Headreco across all tissue types.

Table 3: Top-N Accuracy on 6 classes

Method	***Top-1***	***Top-2***	***Top-3***
Headreco	0.905	0.977	0.983
UNETR-Base	0.913	0.993	0.998
UNETR-HC	0.924	0.995	0.998
UNETR-CM	**0.928**	**0.996**	**0.999**

Quantitative comparison Figure 5 and Table 3 show the Dice score, Hausdorff distance, and top-1/2/3 accuracy on 6-classes. Calibrated UNETR is comparable to Headreco in WM, GM, and CSF; all of our models outperform Headreco in Air, Bone, and Soft tissue. UNETR-HC's hausdorff distances shows that the regularization can improve 6-class segmentation without retraining. UNETR-CM performs the best in Top-1/2/3 accuracy. Figure 6 shows that DOMINO regularization achieves better calibration than UNETR-Base. All algorithms are about evenly calibrated on GM and air, whereas our methods are better calibrated than Headreco on WM, CSF, Bone, and soft tissue.

4 Conclusions

There is often a trade-off between performance and calibration. This work proposes a novel domain-aware calibration method that improves both model calibration and performance. Our results show that regularization leads to better calibration, increased top-N accuracy, and improved segmentation metrics. The calibrated models perform well on full class and reduced class tasks without retraining. This highly-flexible approach can be applied to widespread medical segmentation. Further, model calibration can help improve cross-talk between automated algorithms and manual labelers. Finally, our calibration can be applied to classification tasks in medical image diagnosis. We will release DOMINO to the community to support open science research.

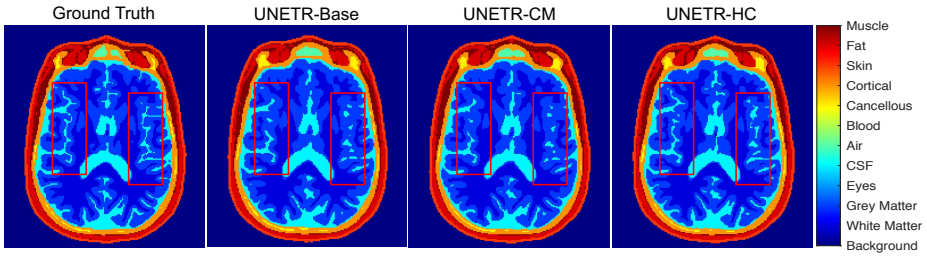


Fig. 2: Sample image slice for 11-tissue segmentation.The red squares show that UNETR-HC captures the GM - CSF boundary better than other methods

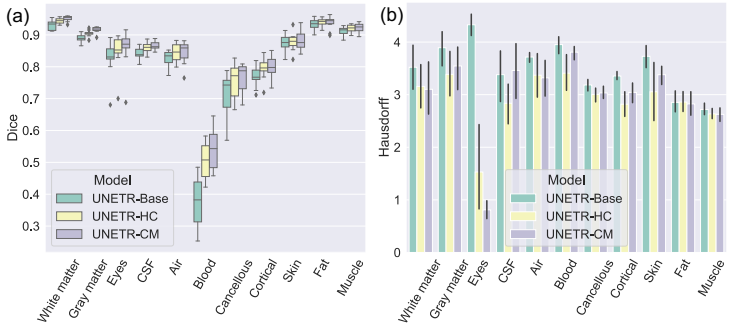


Fig. 3: (a) Dice scores and (b) Hausdorff distances in 11-class segmentation.

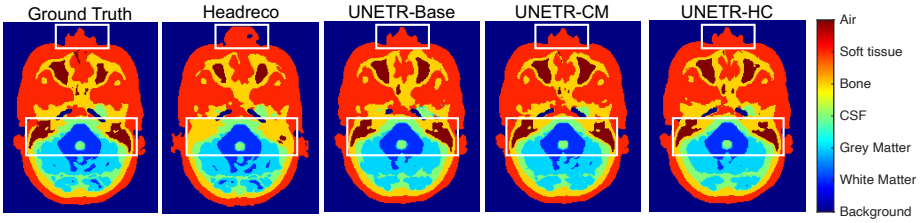


Fig. 4: Sample image slice for 6-tissue segmentation. The white squares highlight important regions where our methods outperformed Headreco

Acknowledgements This work was supported by the National Institutes of Health/National Institute on Aging (NIA RF1AG071469, NIA R01AG054077), the National Science Foundation (1908299), and the NSF-AFRL INTERN Supplement (2130885). We acknowledge NVIDIA AI Technology Center (NVAITC) for their suggestions to this work. We would also like to thank Jiaqing Zhang for assistance in paper formatting.

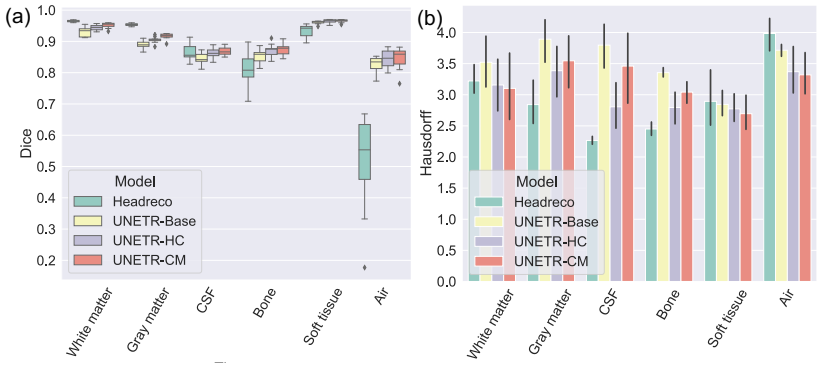


Fig. 5: (a) Dice scores and (b) Hausdorff distances in 6-class segmentation.

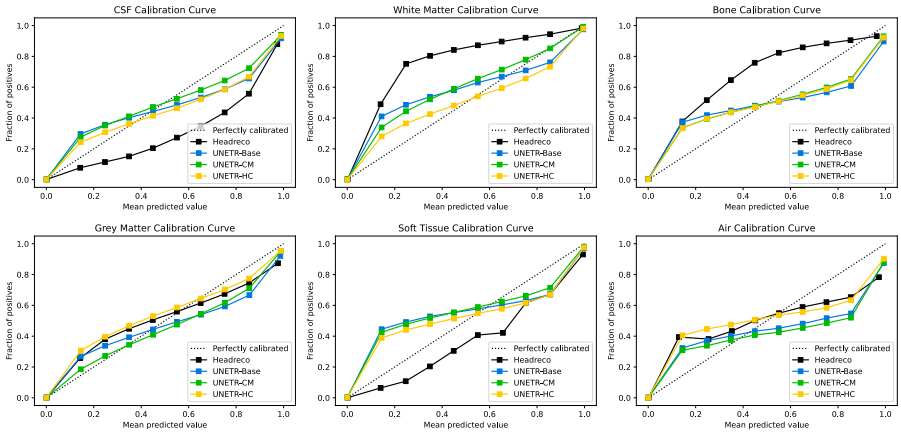


Fig. 6: Calibration curves for 6-class problem.

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76**, 243–297 (2021). https://doi.org/https://doi.org/10.1016/j.inffus.2021.05.008, https://www.sciencedirect.com/science/article/pii/S1566253521001081
2. Albizu, A., Fang, R., Indahlastari, A., O'Shea, A., Stolte, S.E., See, K.B., Boutzoukas, E.M., Kraft, J.N., Nissim, N.R., Woods, A.J.: Machine learning and individual variability in electric field characteristics predict tdcs treatment response. Brain stimulation **13**(6), 1753–1764 (2020)
3. Antonenko, D., Grittner, U., Saturnino, G., Nierhaus, T., Thielscher, A., Flöel, A.: Inter-individual and age-dependent variability in simulated electric fields induced by conventional transcranial electrical stimulation. NeuroImage **224**, 117413 (2021). https://doi.org/https://doi.org/10.1016/j.neuroimage.2020.117413, https://www.sciencedirect.com/science/article/pii/S1053811920308983
4. Ballester, M.A.G., Zisserman, A.P., Brady, M.: Estimation of the partial volume effect in mri. Medical image analysis **6**(4), 389–405 (2002)
5. Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M.B.: Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 92–100. Springer (2019)
6. Consortium, M.: Monai: Medical open network for ai (Mar 2020). https://doi.org/10.5281/zenodo.6114127, https://doi.org/10.5281/zenodo.6114127, If you use this software, please cite it using these metadata.
7. Dubuisson, M.P., Jain, A.K.: A modified Hausdorff distance for object matching. In: Proceedings of 12th international conference on pattern recognition. vol. 1, pp. 566–568. IEEE (1994)

8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/gal16.html
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/guo17a.html
10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
11. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the Hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence **15**(9), 850–863 (1993)
12. Indahlastari, A., Albizu, A., Kraft, J.N., O'Shea, A., Nissim, N.R., Dunn, A.L., Carballo, D., Gordon, M.P., Taank, S., Kahn, A.T., et al.: Individualized tDCS modeling predicts functional connectivity changes within the working memory network in older adults. Brain Stimulation **14**(5), 1205–1215 (2021)
13. Indahlastari, A., Albizu, A., O'Shea, A., Forbes, M.A., Nissim, N.R., Kraft, J.N., Evangelista, N.D., Hausman, H.K., Woods, A.J., Initiative, A.D.N., et al.: Modeling transcranial electrical stimulation in the aging brain. Brain stimulation **13**(3), 664–674 (2020)
14. Jadon, S.: A survey of loss functions for semantic segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). pp. 1–7. IEEE (2020)
15. Nielsen, J.D., Madsen, K.H., Puonti, O., Siebner, H.R., Bauer, C., Madsen, C.G., Saturnino, G.B., Thielscher, A.: Automatic skull segmentation from mr images for realistic volume conductor models of the head: Assessment of the state-of-the-art. Neuroimage **174**, 587–598 (2018)
16. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences **117**(40), 24652–24663 (2020)
17. Taghanaki, S.A., Zheng, Y., Zhou, S.K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., Hamarneh, G.: Combo loss: Handling input and output imbalance in multi-organ segmentation. Computerized Medical Imaging and Graphics **75**, 24–33 (2019)
18. Wilke, M., Schmithorst, V., Holland, S.: Normative pediatric brain data for spatial normalization and segmentation differs from standard adult data. Magnetic Resonance in Medicine **50**(4), 749–757 (2003). https://doi.org/https://doi.org/10.1002/mrm.10606, https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.10606
19. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Huggingface's transformers: State-of-the-art natural language processing (2020)