CALF: A Conditionally Adaptive Loss Function to Mitigate Class-Imbalanced Segmentation

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Imbalanced datasets pose a considerable challenge in training deep learning (DL) models for medical diagnostics, particularly for segmentation tasks. Imbalance may be associated with annotation quality, limited annotated datasets, rare cases, or small-scale regions of interest (ROIs). These conditions adversely affect model training and performance, leading to segmentation boundaries which deviate from the true ROIs. Traditional loss functions, such as Binary Cross Entropy, replicate annotation biases and limit model generalization. We propose a novel, statistically driven, conditionally adaptive loss function (CALF) tailored to accommodate the conditions of imbalanced datasets in DL training. It employs a data-driven methodology by estimating imbalance severity using statistical methods of skewness and kurtosis, then applies an appropriate transformation to balance the training dataset while preserving data heterogeneity. This transformative approach integrates a multifaceted process, encompassing preprocessing, dataset filtering, and dynamic loss selection to achieve optimal outcomes. We benchmark our method against conventional loss functions using qualitative and quantitative evaluations. Experiments using large-scale open-source datasets (i.e., UPENN-GBM, UCSF, LGG, and BraTS) validate our approach, demonstrating substantial segmentation improvements. Code availability: https://anonymous.4open.science/r/MICCAI-Submission-43F9/.

Keywords: adaptive loss function · medical image segmentation

1 Introduction

The application of artificial intelligence (AI) and deep learning (DL) methods has become a revelation in the high-risk and high-stakes world of medicine, particularly in tasks such as image segmentation [1,2,16]. Segmentation algorithms play a crucial role in isolating regions of interest (ROIs) from medical images, enabling disease diagnosis and biomarker discovery [3]. However, the clinical adoption of these methods faces critical challenges, notably the availability of large and high quality data sufficient for training [4,7]. Furthermore, available annotations may be subject to several flaws, resulting in sparse annotations or those that fail to accurately capture true ROIs. These challenges contribute to data imbalance and compromise DL training and generalization [5].

Loss function selection is crucial, as it plays a fundamental role in iteratively guiding the optimization of model parameters [7,8]. Conventional loss functions such as Binary Cross Entropy (BCE), Focal, Tversky, and Dice Losses have been widely adopted, some including characteristics that could mitigate class imbalance [3,6]. However, they often suffer from inherent limitations that hinder their effectiveness, one of which includes their sensitivity to the quality and format of annotation [5]. Since most DL models learn based on the provided annotations, these loss functions tend to reinforce annotation biases. For example, if training labels are polygon-based, roughly drawn, or exceed the boundary limits of the true ROI, the model predictions are likely to mimic these idiosyncrasies, restricting generalization to complex-shaped ROIs (e.g., disease-affected regions with irregular or amorphous boundaries) [7]. Traditional loss functions also exhibit difficulties in handling foreground-background imbalances where the ROI occupies a small fraction of the overall image. BCE overemphasizes the dominant background class and poorly segments small or low-contrast structures [4]. Focal Loss partially addresses this by down-weighting easy-to-classify pixels but requires careful hyperparameter tuning [6]. Dice and Tversky Losses consider pixel-wise agreement between predictions and ground truth, but are susceptible to over-segmentation and fail where under-segmentation is more detrimental [3].These loss functions lack adaptability to varying dataset characteristics, sometimes requiring manual tuning [5]. The rigid nature of these functions limits their applicability to diverse segmentation tasks, necessitating the development of a more dynamic approach which can adjust to data heterogeneity and uncertainty.

In this paper, we introduce the *Conditionally Adaptive Loss Function*(CALF), a novel, statistically driven approach for segmentation tasks on imbalanced datasets. Our key contributions include: (1) developing a dynamic loss function that adapts to the statistical characteristics of the dataset, mitigating annotation biases and enhancing segmentation performance; (2) a hybrid data processing approach integrates preprocessing techniques, flexible dataset filtering mechanisms, and dynamic loss function selection to address data scarcity and imbalance while preserving dataset heterogeneity; (3) a configurable data filtering system that introduces a dataset balancing mechanism, allowing controlled variation in the ratio of ROI-present to ROI-absent images, thereby enabling robust model evaluation under different data conditions; and (4) a comprehensive performance evaluation in which we compare CALF against various loss functions in tumor segmentation tasks, demonstrating superior generalization in rare-class segmentation scenarios.

2 Methodology

Let $\mathbf{x}_i \in \mathbb{R}^{w \times h}$ represent an input grayscale image, where w and h denote the width and height of the image, respectively. Each image is associated with a binary segmentation mask $\mathbf{y}_i \in \{0,1\}^{w \times h}$, where the pixel values indicate the foreground and background regions. Given a dataset of images $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and their corresponding masks $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, the objective is to train a model

$f_{\mathbf{w}}$ with parameters $\mathbf{w}$ that accurately predicts the segmentation mask for any given input image. Loss functions guide model optimization to distinguish between foreground and background regions. However, conventional loss functions do not account for statistical distributions of segmented objects across different images. This hinders model performance, as real-world medical datasets often exhibit uneven distributions of foreground regions. Therefore, the proposed approach dynamically adjusts according to skewness and kurtosis measures to analyze the distribution of foreground regions and counters with appropriate transformations.

2.1 Skewness and Kurtosis

Skewness quantifies the asymmetry of a probability distribution, which means it describes the distribution of foreground object sizes within a dataset. Kurtosis describes the shape of a probability distribution by measuring whether it is peaked or flat, reflected by the variability of the foreground object size. The definitions are described in Table 1 and include the foreground areas $A = \{A_1, A_2, \ldots, A_N\}$, the mean μ and the standard deviation σ. $S < 0$ indicates a distribution with a longer left tail (that is, larger objects in the foreground are more common), $S > 0$ implies the distribution has a longer right tail (i.e., smaller objects dominate), and $S \approx 0$ suggests the sizes of the foreground objects are more symmetrically distributed across the dataset. A high kurtosis ($K > 0$) indicates a distribution with a sharp peak and heavy tails (that is, a mix of very large and small objects). A low kurtosis ($K < 0$) corresponds to a flatter distribution, suggesting that the size of the foreground objects is more uniform across the dataset. Thresholds describing skewness and kurtosis level have been long-established [9].

Table 1: Statistical moments defined as mean, standard deviation, skewness and kurtosis. CALF utilizes skewness and kurtosis to identify imbalances in the raw data distribution that could impact DL training.

Name	**Indicator**	**Formula**
Mean	Central Tendency	$\mu = \dfrac{1}{N}\sum_{i=1}^{N} A_i$
Standard Deviation	Dispersion	$\sigma = \sqrt{\dfrac{1}{N}\sum_{i=1}^{N}(A_i - \mu)^2}$
Skewness	Asymmetry	$S = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(A_i - \mu)^3}{\sigma^3}$
Kurtosis	Tailedness/Peakness	$K = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(A_i - \mu)^4}{\sigma^4} - 3$

2.2 Conditionally Adaptive Loss Function

The proposed loss function, CALF, can be formulated as described in Table 2. The skewness and kurtosis describe the raw data distribution (i.e., whether the distribution of values is extreme or moderate), which are then used as indicators in identifying the most appropriate transformation. These transformations are designed to minimize and stabilize variance and improve normality by handling small and large values as shown below [9,10]:

- $S \leq -1$: Fisher transformation compresses the large objects found.
- $-1 < S \leq -0.5$: shows large objects present but not dominant. Logit slightly expands small and compresses large values.
- $-0.5 < S < 0$: indicates sparse bright regions. Arcsine expands extreme values while compressing mid-range values.
- $S \geq 1$: Log10 improves separation of small foreground objects.
- $0.5 \leq S < 1$: involves small objects but not dominant. Natural log moderately expands small and compresses large objects.
- $0 < S \leq 0.5, K < 0$: show a uniform distribution with Log10 simply spreading out smaller regions for better separation.
- $0 < S \leq 0.5, K \geq 0$: has many small but not extreme objects, with Log10 balancing size and distribution.

The loss function dynamically adapts to the various distributional properties of any given dataset, ensuring that segmentation models remain sensitive to variations in the size of objects and their distribution. The method stabilizes optimization, prevents bias towards dominant object sizes, and improves segmentation performance across diverse datasets.

Table 2: Conditionally adaptive loss function. The loss function automatically detects raw data distribution (i.e., skewness and kurtosis) and applies the appropriate transformation to counteract the imbalance. Unlike other loss functions, such as Focal or Tversky, this method requires no user input.

Transformation	**Condition**	**Formula**
$\mathcal{L}_{\text{Fisher}}$	$S \leq -1$	$-\mathbb{E}\Big[y \cdot \frac{1}{2}\ln\Big(\frac{1+p}{1-p}\Big) + (1-y) \cdot \frac{1}{2}\ln\Big(\frac{1+(1-p)}{1-(1-p)}\Big)\Big]$
$\mathcal{L}_{\text{Logit}}$	$-1 < S \leq -0.5$	$-\mathbb{E}\Big[y \cdot \ln\frac{p}{1-p} + (1-y) \cdot \ln\frac{1-p}{p}\Big]$
$\mathcal{L}_{\text{Arcsine}}$	$-0.5 < S < 0$	$-\mathbb{E}\Big[y \cdot \arcsin(\sqrt{p}) + (1-y) \cdot \arcsin(\sqrt{1-p})\Big]$
$\mathcal{L}_{\text{Log10}}$	$S \geq 1$	$-\mathbb{E}\Big[y \cdot \log_{10}(p) + (1-y) \cdot \log_{10}(1-p)\Big]$
$\mathcal{L}_{\text{Natural Log}}$	$0.5 \leq S < 1$	$-\mathbb{E}\Big[y \cdot \ln p + (1-y) \cdot \ln(1-p)\Big]$
$\mathcal{L}_{\text{Log10}}$	$0 < S \leq 0.5$ and $K < 0$	$-\mathbb{E}\Big[y \cdot \log_{10}(p) + (1-y) \cdot \log_{10}(1-p)\Big]$
$\mathcal{L}_{\text{BCE-Dice}}$	$0 < S \leq 0.5$ and $K \geq 0$	$\mathcal{L}_{\text{Dice}} = -\frac{2\sum(yp)+\epsilon}{\sum y + \sum p + \epsilon}$; $\mathcal{L}_{\text{BCE}} + (1 - \mathcal{L}_{\text{Dice}})$

3 Experiments

CALF performance was evaluated on the basis of the workflow outlined in Figure 1. It involves the extraction of four open source datasets (Sec. 3.1), preprocessing using a custom data loader, benchmarking which involves comparing CALF against selected loss functions and segmentation models, and tracking both qualitative and quantitative performance (Sec. 3.2 and Sec. 4).

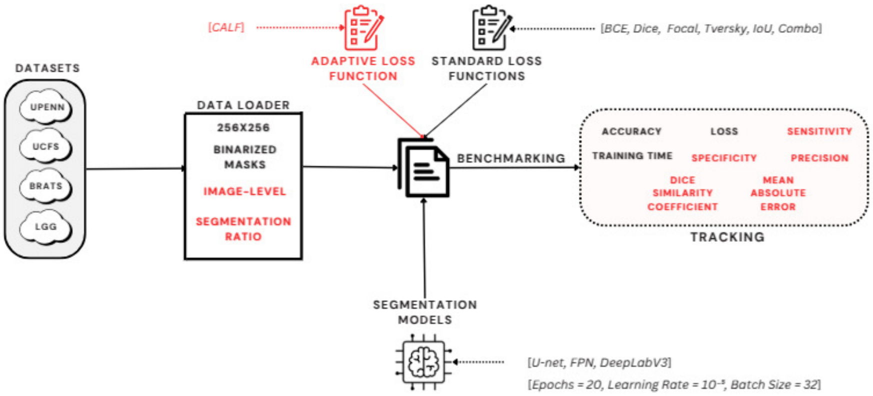


Fig. 1: Workflow of the experiments.

3.1 Datasets

We conducted experiments using four open source high-quality brain cancer datasets from The Cancer Imaging Archive (TCIA) [11] (Table 3). These were chosen to capture various imbalanced conditions for testing and validating CALF. The **UCSF-PDGM** [12], **BraTS** [13], **UPENN-GBM** [14], and **LGG-1p19q Deletion** [15] datasets contain gliomas and glioblastomas MRI scans.

The datasets contain combinations of T1, T2, T1 contrast-enhanced (T1 CE), FLAIR, DWI and SWI. Resolutions ranged from 240×240 to 256×256 pixels. The data processing involved the conversion of three-dimensional images into two-dimensional slices, which were subsequently saved in the Portable Network Graphics (PNG) format. The ground truth labels were converted from grayscale to binary in all datasets. The final dataset consisted of 1,410 patients and 589,838 2D images. An overview of the data (including the training and testing divisions) is given in Table 3.

3.2 Benchmarking

We trained our datasets using U-Net [16], DeepLabV3 [17], and FPN [18] and seven different loss functions (BCE [19], Dice [20], Tversky [21], IoU [22], Focal

Table 3: Summary of dataset used in experiments, including total patient and image numbers, as well as division into training and testing.

Dataset	**Patients**	**Images**	**Training Images**	**Testing Images**
UPENN-GBM	611	284,115	256,620	27,495
UCSF	495	230,175	207,900	22,275
BraTS	145	67,425	60,900	6,525
LGG	159	8,123	7,328	795
Total	1,408	589,838	532,748	57,090

[23], BCE-Dice [24], and CALF). Comparative loss functions were selected after reviewing 25 different loss functions identified by querying the `{loss_function_name} segmentation` in the Dimensions.ai database. These were sorted based on: citation count, use cases (particularly for imbalanced data), and their frequent application in segmentation tasks. A custom data loader was created to specify a tumor-to-non-tumor ratio (from 0 to 1), ensuring models were provided with annotated image–mask pairs, as imbalances also included an inequitable distribution of images with tumors vs. those without. The quantitative metrics used are shown in Figure 1. These were collected along with qualitative analysis that evaluated the precision of the segmentation through visual inspection.

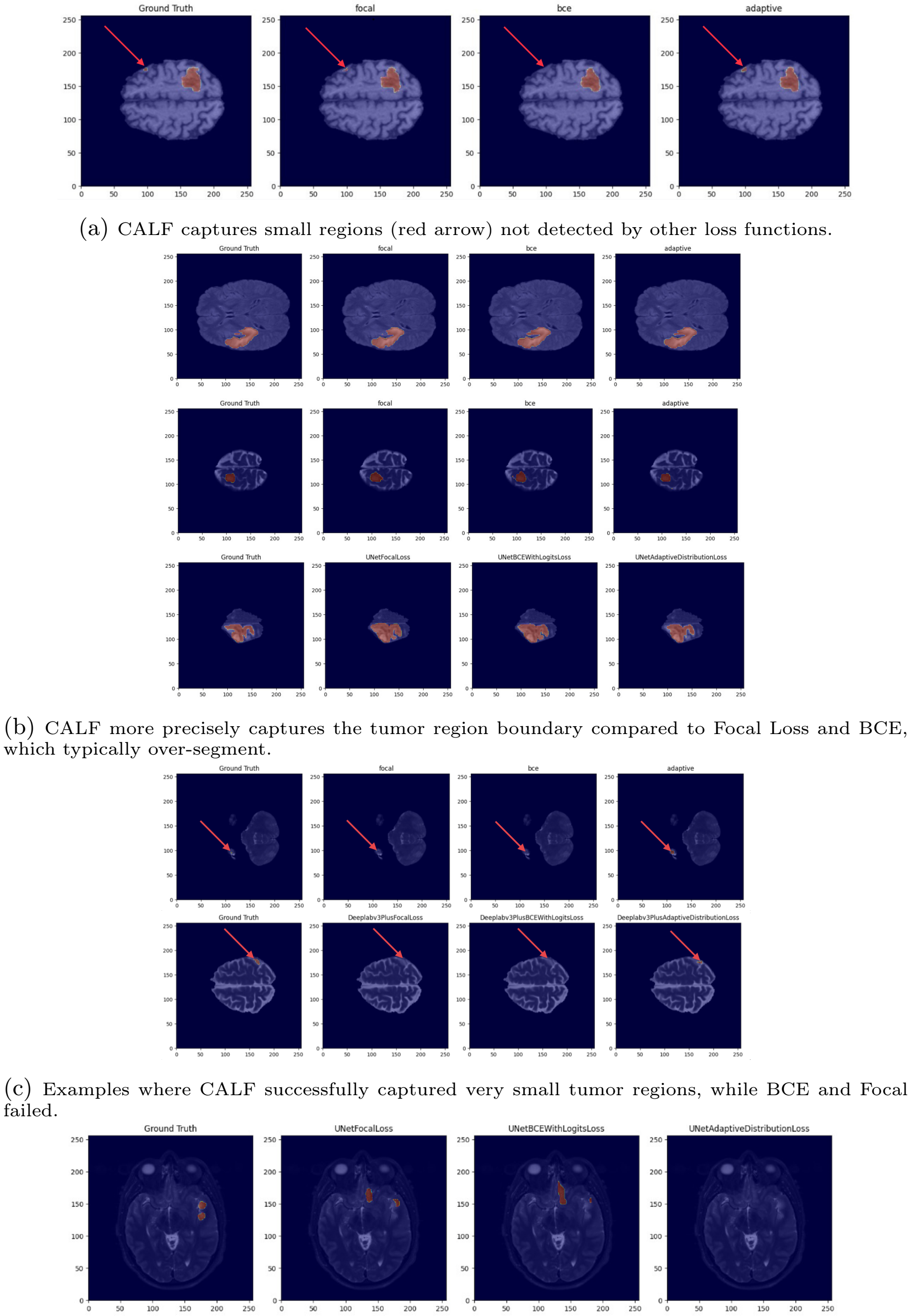
4 Results

Several combinations of model-loss functions with varying ratios were tested to determine how adaptable the loss functions are under various data scarcity conditions. Table 4 presents a comparative analysis of three best performing loss functions (BCE, Focal, and CALF) across the models, with a ratio of **40.9%** (the 'default' ratio, representing the total number of tumor cases available). Our proposed loss function CALF demonstrated consistent performance, quantitatively competing with BCE and outperforming Focal loss in multiple cases. Figure 2 also demonstrates CALF segmentation performance in comparison with BCE and Focal Loss.

Table 4: Model Performance with BCE, Focal, and CALF with ratio of 0.409.

Model	**BCE**			**Focal**			**CALF**		
	Accuracy	DSC	MAE	Accuracy	DSC	MAE	Accuracy	DSC	MAE
U-Net	**0.9991**	**0.9589**	**0.0008**	0.9990	0.9552	0.0009	0.9990	0.9577	0.0009
DeepLab v3	**0.9969**	0.8597	**0.003**	0.9961	0.7998	0.0037	**0.9969**	**0.8598**	**0.003**
FPN	**0.9981**	**0.9128**	**0.0018**	0.9977	0.8892	0.0022	0.9979	0.9072	0.002

As evident in Table 4, BCE showed strong overall performance, achieving the highest DSC and accuracy when trained on U-Net, while the proposed CALF



(a) CALF captures small regions (red arrow) not detected by other loss functions.

(b) CALF more precisely captures the tumor region boundary compared to Focal Loss and BCE, which typically over-segment.

(c) Examples where CALF successfully captured very small tumor regions, while BCE and Focal failed.

(d) An instance where CALF did not predict a mask that was captured by Focal loss and BCE.

Fig. 2: Comparison of different loss functions' performance.

performed equally well and even surpassed BCE in DeepLabV3. The Focal loss function, on the other hand, struggled significantly in DeepLabV3, with a DSC score of only 0.7998, highlighting its weakness in this architecture.

To further investigate the performance of CALF, we trained FPN on an imbalanced dataset with a ratio of **10%**. The outcomes presented in Table 5, demonstrate that the proposed CALF outperformed alternative loss functions in this particular scenario, which is the focal point of our investigation. Table 5 highlights the robustness of CALF, which consistently outperformed other loss functions in the case of imbalanced dataset. This finding underscores the efficacy of our loss function in addressing severe class imbalance. Although the BCE loss function exhibited specificity, it demonstrated challenges in detecting small tumor regions. In contrast, CALF exhibited more consistent performance in this regard.

Table 5: FPN model performance for various loss functions with a ratio of 0.1.

Loss	Accuracy	DSC	Specificity	Sensitivity	Precision	MAE
BCE	0.9964	0.8187	0.9992	0.7430	0.9174	0.0035
Tversky	0.9943	0.7548	0.9964	0.8135	0.7106	0.0056
IoU	0.9945	0.7683	0.9962	0.8473	0.7092	0.0054
Focal	0.9957	0.7701	**0.9996**	0.6488	**0.9573**	0.0042
Dice	0.9948	0.7722	0.9967	0.8218	0.7343	0.0051
BCE-Dice	0.9961	0.8265	0.9976	**0.8534**	0.8046	0.0038
CALF	**0.9965**	**0.8267**	0.9991	0.7598	0.9113	**0.0034**

5 Conclusion

CALF, a conditionally adaptive loss function, was introduced to address the challenges posed by class imbalance in medical image segmentation. By leveraging statistical characteristics like skewness and kurtosis, appropriate transformations are applied to mitigate imbalance and optimize learning. Experiments were conducted on four large-scale, open-source tumor segmentation datasets: UCSF-PDGM [12], BraTS [13], UPENN-GBM [14] and LGG [15]. CALF exhibited a consistent improvement in segmentation accuracy, achieving high-level and persistent qualitative and quantitative results compared to standard loss functions, which exhibited variability in their performance. For example, BCE quantitatively performed well, particularly in its detection of larger tumor regions. However, qualitatively, it over-segmented ROIs, illustrating poor performance in rare-class scenarios. While CALF provides a promising approach to improving medical image segmentation, future research could also explore its adaptability to varying noise levels and integration with semi-supervised learning techniques. Additionally, expanding its application to other imaging modalities and segmentation tasks could further demonstrate its flexibility and validate its robustness.

References

1. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & van der Laak, J. A. W. M.: A survey on deep learning in medical image analysis. Medical image analysis, **42**, 60–88 (2017).
2. Shen, D., Wu, G., Suk, H. I.: Deep learning in medical image analysis. Annual review of biomedical engineering, **19**, 221–248 (2017).
3. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M. J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 240–248 (2017).
4. Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., ... & Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis, **36**, 61–78 (2017).
5. 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 (2020).
6. Abraham, N., Khan, N. M.: A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI), pp. 683–687 (2019).
7. Ma, J., Zhang, Y., Gu, S., Zhu, L., Guo, D., Sun, Z., Yu, W.: Loss functions for medical image segmentation: A survey. IEEE Transactions on Medical Imaging, **40**(11), 3253–3268 (2021).
8. Yeung, M., Sala, E., Schönlieb, C. B., Rundo, L.: Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalances in medical image segmentation. Computers in Biology and Medicine, **134**, 104314 (2021).
9. Bulmer, M. G.: *Principles of Statistics*. Dover Publications, New York, p. 88 (1979).
10. Stevens, J. P.: *Applied Multivariate Statistics for the Social Sciences*. Routledge, New York (2009).
11. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., ... & Prior, F.: The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. Journal of Digital Imaging, **26**(6), 1045–1057 (2013). https://www.cancerimagingarchive.net/.
12. Soltaninejad, M., Yang, Z., Lamb, B., et al.: UCSF-PDGM: A Public MRI Dataset for Glioma Segmentation. The Cancer Imaging Archive (2022). https://www.cancerimagingarchive.net/
13. Menze, B.H., Jakab, A., Bauer, S., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BraTS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2015.2388036
14. Bakas, S., Akbari, H., Sotiras, A., et al.: Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM Collection. The Cancer Imaging Archive (2017). https://www.cancerimagingarchive.net/
15. Clark, K., Vendt, B., Smith, K., et al.: The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013).
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015).
17. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution,

and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, **40**(4), 834–848 (2017).

18. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic Feature Pyramid Networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6399–6408 (2019).
19. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, pp. 164–169 (2016). https://www.deeplearningbook.org/
20. Milletari, F., Navab, N., Ahmadi, S. A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: International Conference on 3D Vision (3DV), pp. 565–571 (2016).
21. Salehi, S. S. M., Erdogmus, D., Gholipour, A.: Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In: International Workshop on Machine Learning in Medical Imaging (MLMI), pp. 379–387 (2017).
22. Rahman, M. A., Wang, Y.: Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. In: International Symposium on Visual Computing (ISVC), pp. 234–244 (2016).
23. Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017).
24. 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. In: Nature Methods, vol. 18, no. 2, pp. 203–211 (2021).
25. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., Dalca, A. V.: VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Transactions on Medical Imaging, **38**(8), 1788–1800 (2019).
26. Dice, L. R.: Measures of the Amount of Ecologic Association Between Species. Ecology, **26**(3), 297–302 (1945).
27. Willmott, C. J., Matsuura, K.: Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. Climate Research, **30**(1), 79–82 (2005).