Pattern Recognition Letters 203 (2026) 111–118

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Preprocessing strategies and their influence on deep learning-driven MRI segmentation

Ariadna Jiménez-Partinen [a,b], Ezequiel López-Rubio [a,b], Fátima Nagib-Raya [c], Esteban J. Palomo [a,b,*], Rafael M. Luque-Baena [a,b]

[a] ITIS Software, University of Málaga, Málaga, 29071, Spain
[b] Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma BIONAND, Málaga TechPark, 29590, Spain
[c] Department of Radiology, Hospital Regional Universitario de Málaga, Avenida de Carlos Haya, 84, Málaga, 29010, Spain

ARTICLE INFO

Editor: Maria De Marsico

Keywords:
MRI
Medical imaging
Segmentation
Deep learning

ABSTRACT

In this work, a comprehensive analysis of the impact of intensity value regularization methods on 3D MRI segmentation for three neurological disorders: glioblastoma, multiple sclerosis, and epilepsy, is presented. The experiments were conducted through three architectures: nnU-Net (convolutional neural network), WNet (hybrid combining convolutional and transformer elements), and Primus (transformer-based), considering both FLAIR and T1-weighted images, as well as FLAIR-only scenarios.

The statistical analysis conducted underscores the crucial role of intensity regularization in the performance. The results indicate that among the intensity regularization methods tested in this study, KDE, White-stripe, and Z-score standardizations proved to be particularly effective. Furthermore, nnU-Net is the most robust architecture against intensity variability, with small improvements of around 3%. Meanwhile, methods incorporating TF elements are more sensitive to these variations. WNet demonstrates slightly greater gains, around 6%. While Primus can be less stable and underperform compared to nnU-Net and WNet in most cases; nonetheless, it remains a promising and competitive option. Additionally, it has been demonstrated that adding an extra channel does not necessarily guarantee improved performance, while also increasing computational cost.

1. Introduction

Object segmentation is a key research topic in computer vision, especially in medical imaging, where lesion segmentation can be a time-consuming factor and it can be affected by subjectivity due to the intra- and inter-observer variability. Magnetic Resonance Imaging (MRI) is a fundamental tool used to analyze brain anatomy for diagnosing, monitoring, and assessing treatment responses in neurological disorders, including brain tumors, Alzheimer's disease, multiple sclerosis (MS), and epilepsy [1]. Accurate voxel-level outlining of abnormalities and regions of interest is crucial for diagnosis, prognosis, and treatment guidance [2]. Nevertheless, MRI scans may be affected by various factors, such as noise, artifacts, and intensity variations coming from acquisition protocols, patient position and characteristics, resonance machines, magnetic field strength, or scans from the same patient at different time points [3]. Therefore, in MRI analysis, preprocessing is an essential step to correct these imperfections to obtain more homogeneous and comparable images intra- and inter-patients [4].

Deep Learning-based segmentation methods require a previously accurate preprocessing step [5], but its impact on the performance and robustness is under-explored. In the state-of-the-art, most works on MRI lesion segmentation apply registration and skull-stripping steps [6,7], with only a few incorporating intensity normalization [8–11] or bias field correction [12,13] because performance improvement relies on architecture modifications or hyperparameter optimization [14]. Nonetheless, some studies have studied the effect of intensity regularization. For instance, Reinhold et al. [15] concludes that while normalization methods enhance the results, they have similar effects on MRI synthesis. Ghazvanchahi et al. [16] states that White-stripe and Z-score standardizations improve white matter lesion (WML) segmentation using Fluid-Attenuated Inversion Recovery (FLAIR) images. Jacobsen et al. [17] employs T1-weighted (T1-w) images for cerebellum segmentation, showing that intensity normalization strategies influence the CNN performance. Meanwhile, Kondrateva et al. [18] notes a negligible effect of different inter-subject alignment approaches and some image enhancement with 3D segmentation considering a multimodal

* Corresponding author.
E-mail addresses: ariadna@uma.es (A. Jiménez-Partinen), ezeqlr@lcc.uma.es (E. López-Rubio), fatimanagib@gmail.com (F. Nagib-Raya), ejpalomo@uma.es (E.J. Palomo), rmluque@uma.es (R.M. Luque-Baena).

https://doi.org/10.1016/j.patrec.2026.02.030
Received 22 June 2025; Received in revised form 30 October 2025; Accepted 27 February 2026
Available online 8 March 2026

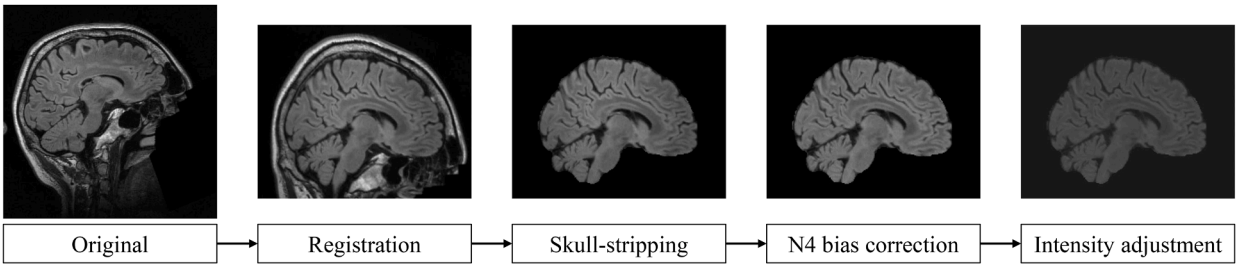


Fig. 1. Preprocessing pipeline. Intensity values are normalized for visualization purposes.

input – 4 MRI sequences –. There is a lack of studies on the effects of preprocessing and input influence on 3D segmentation DL-based architectures across different architectures and considering various pathologies. This results in non-uniform and non-reproducible protocols, which hinder the understanding of DL solutions and their application in clinical settings [19].

Our proposal examines the impact of intensity regularization in the preprocessing step for 3D MRI segmentation tasks. The main contribution is the comprehensive analysis conducted of five different intensity regularization strategies, unimodal and multimodal approaches, and three different Depp Learning (DL) architectures – convolutional neural network (CNN), Transformer (TF), and hybrid (CNN + TF) – are explored across three pathologies –glioblastoma (GBM), multiple sclerosis (MS), and epilepsy –. This study aims to identify shortcomings, requirements, and potential improvements in 3D segmentation solutions.

2. Methodology

2.1. Datasets

Three open-access annotated datasets, composed of different MRI sequences, from different pathologies have been considered to reliably study the impact of intensity regularization on 3D segmentation performance.

Glioblastoma (GBM) is the most common and lethal brain tumor [20]. The UPenn-GBM dataset [21] includes MRI scans – T1-weighted (T1-w), T2-weighted (T2-w), Fluid-Attenuated Inversion Recovery (FLAIR) images, and T1-weighted with gadolinium (T1-Gd) – from 611 patients diagnosed with *de novo* glioblastoma, featuring registered and skull-stripped images. The annotated regions correspond to enhancing tumor (ET), the necrotic tumor core (NCR), and edema (ED). ET and NCR are visible in T1-Gd scans, while ED is revealed as a hyperintense signal in FLAIR scans. This study employs a multimodal approach, utilizing both FLAIR and T1-Gd sequences, to address the multiclass segmentation problem, as the FLAIR approach alone is insufficient, as only ED is distinguishable.

Multiple Sclerosis (MS) is a chronic neurological disease that damages myelin sheaths, which protect nerve fibers, disrupting normal neurological function [22]. The annotated MSLesSeg dataset [23,24] comprises 53 patients with different time points, with already preprocessing steps: registering and skull-stripping.

Epilepsy is a neurological disorder that can be considered focal if the seizures originate in a specific cerebral area [25]. Focal cortical dysplasias (FCD) are circumscribed malformations of cortical development leading to an increased risk of seizures with functional disruptions [26]. While FCD type I is very subtle and presents a normal brain MRI, FCD II cases are characterized by abnormalities on MRI (mainly on FLAIR), such as a non-uniformly localized signal increase at the gray-white matter junction, resulting in subtle blurring, abnormal gyration patterns, and slight variation in cortical thickness [27]. Schuch et al. [28] created an open-access annotated dataset of 85 epilepsy patients – 78 suspected FCDII, 5 MRI-negative, and 2 with other abnormalities – and 85 healthy controls. The MRI scans are intra-subject registered and defaced. The details of the open-access datasets, before the application of additional preprocessing as described below, are summarized in Table 1.

Table 1
Datasets summary. "Reg" = registration; "Skull-str." = skull-stripping.

Dataset	Classes	N. Patients / Control	Base preprocessing
UPenn-GBM [21]	Multiclass	611	SRI24 reg. + skull-str.
MSLesSeg [23,24]	Binary	53	MNI152 reg. + skull-str.
Epilepsy [28]	Binary	78 / 85	Defacing

2.2. Preprocessing

Raw MRI data undergo preprocessing to ensure intra- and inter-subject homogeneity before learning tasks. This involves four main steps: registration, information reduction, bias correction of inhomogeneous intensities, and intensity regularization. The preprocessing pipeline is illustrated in Fig. 1 and described below.

Registration: MRI sequences were registered from an original to a common space to ensure correct alignment across subjects. This iterative transformation was already performed in the UPenn-GBM dataset to the SRI24 space and the MSLesSeg dataset to the MNI152 space, while the epilepsy dataset was affine and deformable registered to the $1mm^3$ MNI152 space.

Skull-stripping: non-brain tissues were removed to reduce and focus the information. This phase has already been performed in the UPenn-GBM and MSLesSeg datasets. For the epilepsy dataset, the brain extraction function from the ANTs library was used [29].

N4 bias field correction: the bias field is observed as a slight non-uniformity in the intensity values corresponding to the same tissue in different locations, which hinders the consistency of medical images [30]. The N4 algorithm [31] was applied to homogenize intensities from the same tissue in the UPenn-GBM, MSLesSeg, and epilepsy datasets.

Intensity regularization strategies: since MRI sequences come from different resonance machines, intensity value regularization is essential to get intensity values within the same dynamic range and aligned histograms. To analyze the influence of intensity modification on DL solutions, five regularization strategies were explored [15]:

0. *None*: intensities are unmodified.
1. *Min-Max normalization* (MinMax):

$$I_{minmax} = \frac{I - I_{min}}{I_{max} - I_{min}}, \tag{1}$$

where I is the intensity value of a voxel, and I_{max} and I_{min} are the maximum and minimum intensity values of the image, respectively.
2. *Kernel Density Estimate-based (KDE) standardization*:

$$I_{KDE} = \frac{I}{\rho}, \tag{2}$$

where ρ is the white-matter (WM) intensity peak.
3. *White-stripe standardization* (Ws):

$$I_{ws} = \frac{I - \mu_{ws}}{\sigma_{ws}}, \tag{3}$$

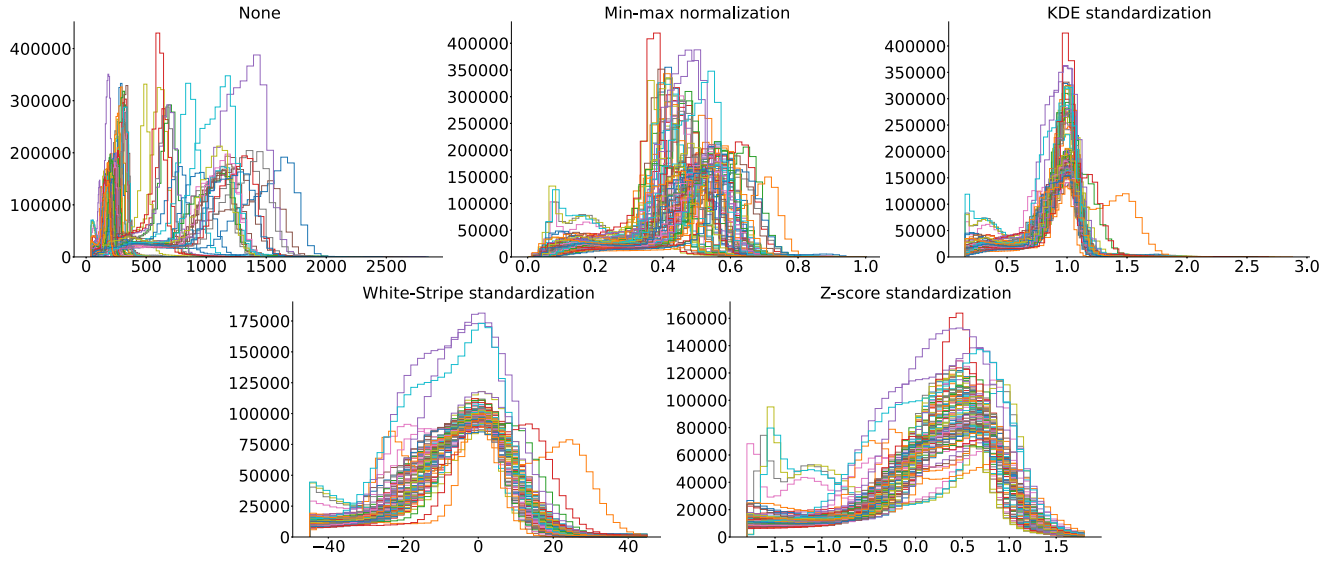
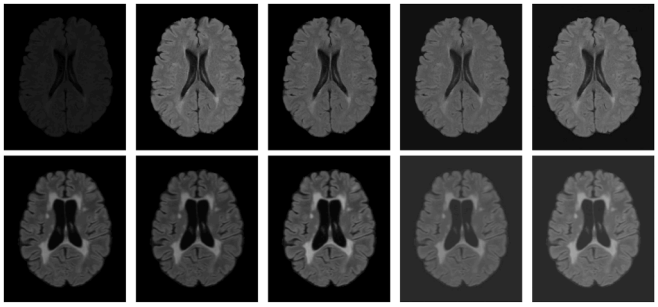


Fig. 2. Histograms of FLAIR scans from the MSLesSeg dataset for each proposed intensity regularization method, with background intensities thresholded for visualization purposes. The x axis represents the available range of possible values a voxel could have, and the y axis represents the total number of voxels that have each specific value, i.e., the absolute frequency of each intensity value.



(a) None (b) MinMax (c) KDE (d) Ws (e) Zs

Fig. 3. FLAIR axial samples from subjects 5 (first row) and 39 (second row) of the MSLesSeg dataset with each proposed intensity regularization approach.

is based on parameters obtained from a sample of normal-appearing white-matter (NAWM) through the segmentation of the WM. Let μ_{ws} be the WM intensity peak, and σ_{ws} the standard deviation within 10% of it.

4. *Z-score standardization* (Zs):

$$I_{zs} = \frac{I - \mu}{\sigma} \tag{4}$$

where μ and σ are the mean and the standard deviation of the intensity values, respectively.

Selected intensity regularization strategies were applied independently after the previous steps, as is described in Fig. 1. Fig. 2 depicts the histograms of FLAIR scans from the MSLesSeg dataset; and Fig. 3 shows the differences between intensity regularization strategies comparing two subjects.

3. Experiments

3.1. Setup

Experiments are coded as *XYZ*, where *X* indicates the pathology, *Y* corresponds to unimodal or multimodal approach – for GBM only a multimodal strategy is addressed due to in FLAIR scans only edema (ED) region can be visualized –, and *Z* denotes the intensity regularization method applied to the previously preprocessed images, as aforementioned in Section 2.2. Details are reported in Table 2.

Table 2
Experiment code details.

X	Y	Z
6 - UPenn-GBM	0 - FLAIR	0 - None
7 - MSLesSeg	1 - FLAIR + T1-w	1 - MinMax
8 - Epilepsy + control		2 - KDE
9 - Epilepsy		3 - Ws
		4 - Zs

To analyze the impact of preprocessing approaches on a broad range of DL-based methods, we selected three 3D segmentation models: nnU-Net[32,33], which is a CNN-based model considered the benchmark in 3D MRI segmentation [34]; WNet [35], a novel hybrid model that integrates convolutional and transformer elements; and Primus [36], which is an innovative entirely Transformer-based architecture.

The three models are embedded in the "nnunet" library, and the custom configuration implemented inherits from "*3d_fullres*", with a batch size of 8 and "*NoNormalization*" scheme in order to provide the input data already preprocessed with the complete pipeline detailed in Section 2.2. For nnU-Net, WNet, and Primus, the "nnUNetTrainer", "WNet3D_S", and "Primus_M" trainers, along with their default training parameters, were used. Each dataset was randomly split into training (80%) and testing (20%) subsets, ensuring that sequences from the same patient across different time points were included in the same subset. The training subset was used to conduct a 5-fold cross-validation scheme with 250 epochs. The best model identified during each training was then independently evaluated on the same held-out test subset.

3.2. Results

For evaluating segmentation performance, the commonly used Dice Similarity Coefficient (DSC) was selected. The DSC is a commonly used overlap measure that quantifies the similarity between the predicted segmentation and the ground truth, ranging from 0 to 1. A DSC of 1 indicates perfect overlap, while a value of 0 indicates no overlap. Table 3 reports DCS values as the $mean \pm standard\ deviation$ from predicting the held-out test subset with the best model identified during each training

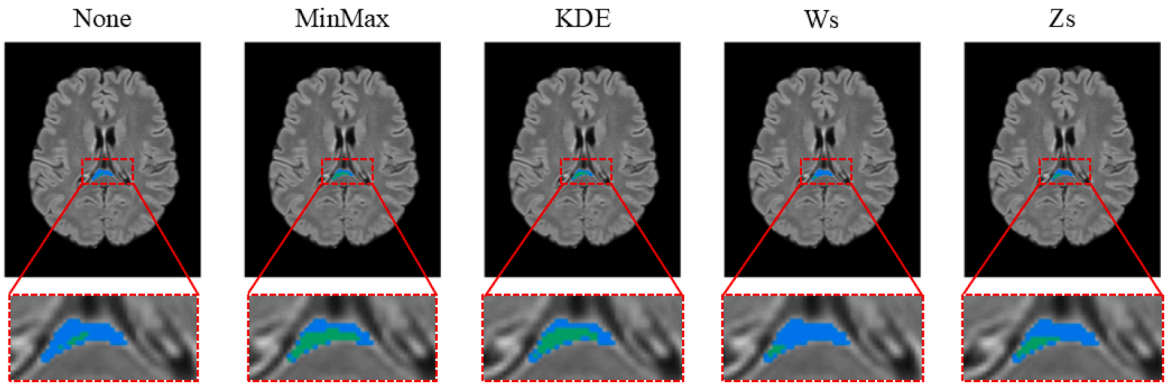


Fig. 4. Pimus prediction example in a FLAIR sequence of MSLesSeg with each intensity regularization method (71Z). Blue = ground truth mask; Green = intersection between ground truth and predicted masks. Intensity values are normalized for visualization purposes.

Table 3
DSC obtained on the same held-out external test subset by each model trained through 5-fold cross-validation in the format *mean ± standard deviation*. CNN = Convolutional Neural Network; TF = Transformer. The highest model values are in **bold**, and the highest by intensity regularization are *underlined*.

	Architecture			CNN	CNN + TF	TF
Dataset	Sequences	Intensity regularization	Code	nnU-Net	WNet	Primus
UPenn-GBM	FLAIR + T1-Gd	None	610	**0.862 ± 0.002**	0.835 ± 0.025	0.505 ± 0.402
	FLAIR + T1-Gd	MinMax	611	**0.866 ± 0.003**	0.864 ± 0.001	0.850 ± 0.001
	FLAIR + T1-Gd	KDE	612	**0.868 ± 0.002**	0.867 ± 0.002	0.853 ± 0.002
	FLAIR + T1-Gd	Ws	613	**0.866 ± 0.002**	0.759 ± 0.098	0.849 ± 0.001
	FLAIR + T1-Gd	Zs	614	**0.773 ± 0.052**	0.174 ± 0.347	0.409 ± 0.229
MSLesSeg	FLAIR	None	700	**0.685 ± 0.008**	0.676 ± 0.009	0.636 ± 0.009
	FLAIR	MinMax	701	0.685 ± 0.002	**0.686 ± 0.007**	0.635 ± 0.008
	FLAIR	KDE	702	**0.690 ± 0.006**	0.682 ± 0.007	0.640 ± 0.009
	FLAIR	Ws	703	0.688 ± 0.010	**0.691 ± 0.006**	0.675 ± 0.008
	FLAIR	Zs	704	0.681 ± 0.010	**0.686 ± 0.005**	0.663 ± 0.012
	FLAIR + T1-w	None	710	**0.683 ± 0.014**	0.590 ± 0.174	0.003 ± 0.002
	FLAIR + T1-w	MinMax	711	**0.680 ± 0.008**	0.678 ± 0.014	0.643 ± 0.007
	FLAIR + T1-w	KDE	712	0.681 ± 0.008	**0.686 ± 0.007**	0.653 ± 0.008
	FLAIR + T1-w	Ws	713	0.677 ± 0.011	**0.685 ± 0.004**	0.659 ± 0.008
	FLAIR + T1-w	Zs	714	0.679 ± 0.009	**0.680 ± 0.013**	0.654 ± 0.007
Epilepsy + Control	FLAIR	None	800	**0.245 ± 0.034**	0.199 ± 0.100	0.087 ± 0.016
	FLAIR	MinMax	801	0.224 ± 0.056	**0.233 ± 0.027**	0.124 ± 0.013
	FLAIR	KDE	802	**0.213 ± 0.060**	0.177 ± 0.114	0.141 ± 0.031
	FLAIR	Ws	803	0.211 ± 0.024	**0.221 ± 0.027**	0.091 ± 0.013
	FLAIR	Zs	804	0.231 ± 0.049	**0.272 ± 0.029**	0.142 ± 0.038
	FLAIR + T1-w	None	810	**0.256 ± 0.065**	0.221 ± 0.038	0.004 ± 0.002
	FLAIR + T1-w	MinMax	811	**0.233 ± 0.039**	0.173 ± 0.124	0.144 ± 0.012
	FLAIR + T1-w	KDE	812	**0.282 ± 0.027**	0.209 ± 0.115	0.166 ± 0.010
	FLAIR + T1-w	Ws	813	**0.236 ± 0.026**	0.214 ± 0.029	0.130 ± 0.023
	FLAIR + T1-w	Zs	814	**0.258 ± 0.050**	0.227 ± 0.041	0.166 ± 0.030
Epilepsy	FLAIR	None	900	0.206 ± 0.037	**0.267 ± 0.033**	0.215 ± 0.020
	FLAIR	MinMax	901	0.201 ± 0.032	**0.267 ± 0.046**	0.253 ± 0.024
	FLAIR	KDE	902	0.244 ± 0.035	**0.313 ± 0.026**	0.294 ± 0.055
	FLAIR	Ws	903	0.215 ± 0.029	**0.296 ± 0.014**	0.251 ± 0.020
	FLAIR	Zs	904	0.228 ± 0.037	**0.308 ± 0.026**	0.303 ± 0.051
	FLAIR + T1-w	None	910	0.268 ± 0.038	**0.290 ± 0.023**	0.119 ± 0.034
	FLAIR + T1-w	MinMax	911	0.290 ± 0.046	0.294 ± 0.053	**0.297 ± 0.036**
	FLAIR + T1-w	KDE	912	**0.324 ± 0.039**	0.320 ± 0.017	0.303 ± 0.049
	FLAIR + T1-w	Ws	913	0.274 ± 0.025	**0.318 ± 0.023**	0.293 ± 0.036
	FLAIR + T1-w	Zs	914	0.295 ± 0.048	0.298 ± 0.038	**0.333 ± 0.042**

in the 5-fold cross-validation scheme; i.e., the test subset was predicted five times per experiment. A qualitative MS example is depicted in Fig. 4 where the prediction of each intensity regularization approach is illustrated.

There are some findings worth mentioning. Firstly, epilepsy lesions are clearly the worst-segmented because of their reduced size and challenging discernment. A representative sample is depicted in Fig. 5 where the DSC attained is 0.787. Secondly, Primus underperforms compared to CNN-related architectures, while the highest DSC values are equally distributed between nnU-Net and WNet. In the UPenn-GBM dataset, nnU-Net performs best, while WNet excels in the epilepsy dataset. Results for MSLesSeg are more evenly distributed. Additionally, Primus shows significant outliers at values of 610, 710, and 800, indicating unstable training with raw intensity values (XY0). This instability arises from the

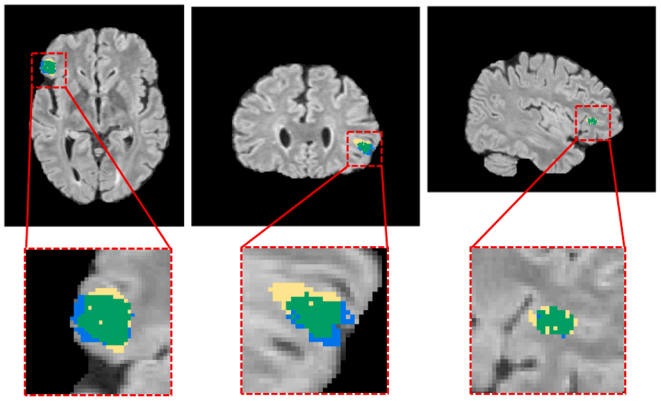


Fig. 5. nnU-Net prediction example in a FLAIR sequence of epilepsy with Z-score standardization (914). DSC of 0.787. Blue = ground truth mask; Yellow = predicted mask; Green = intersection between ground truth and predicted masks. Intensity values are normalized for visualization purposes.

sensitivity of the attention blocks in TF methods to initial embedding scales, which depend directly on the intensities, resulting in a saturated attention layer and unstable gradients. Thirdly, focusing on the intensity regularization method, KDE achieves the highest values, closely followed by Ws and Zs, while None and MinMax strategies retrieve lower outcomes. This fact is discussed in depth in Subsection 3.3.

Fig. 6 illustrates the trade-off between the time-consuming and the DSC achieved. This figure explores the behavior across different models, represented by marker shapes, and the task addressed, indicated by marker color. Additionally, the use of one or two input sequences is represented by the intensity of the color. The isolated instances of the Primus and WNet models correspond to the outliers previously discussed.

Considering each pathology, the DCS values are quite similar, regardless of the number of input channels and the specific model used. However, epilepsy outcomes exhibit more dispersed values. Regarding architectures, it is clear that Primus, followed by WNet, requires more training time to converge, as TF-based models are slower than CNN models. There is a noticeable difference in performance between the datasets due to the complexity of the problem and the limited number of samples available.

Additionally, the multimodal version (using FLAIR and T1-w), indicated in a more intense color, requires more computational time due to the incorporation of this extra channel. However, in most cases, the performance is very similar. To analyze in more detail whether the inclusion of an additional sequence improves performance, we compare the differences between using a unimodal approach (only FLAIR scans) and a multimodal approach (both FLAIR and T1-weighted scans), as shown in Fig. 7. For nnU-Net, adding T1-weighted images actually worsens performance, especially in epilepsy instances (8YZ and 9YZ cases). For WNet and Primus, there are instances where a multimodal approach leads to higher performance due to the aforementioned irregularities in the training process. For Primus, when we exclude outlier values, adding an extra channel does not generally improve the DCS. Meanwhile, for WNet, the performance differences are negligible, making it unclear whether the additional channel is beneficial. Overall, incorporating the extra channel appears to have no substantial effect on performance. Additionally, when considering computational costs and time requirements, 3D segmentation of MS and epilepsy lesions with any DL-based proposed architecture may be unnecessary.

3.3. Statistical analysis

This section provides a statistical analysis conducted to assess the statistical significance of the observed differences in test results of

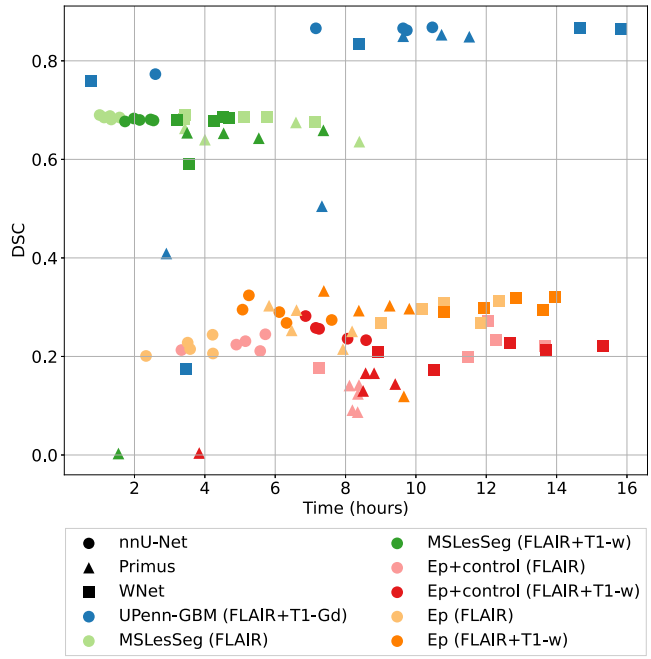


Fig. 6. Trade-off between mean DSC values – Table 3 – and training time. Marker shapes and colors represent the architecture and dataset, respectively.

Table 3. Specifically, we analyze the effect of intensity adjustment on DSC performance across all experiments for each proposed model. Due to the limited sample size, the non-parametric Wilcoxon rank-sum test [37] was conducted. The Wilcoxon test evaluates the null hypothesis (H_0) that DSC values from 5-fold cross-validation with proposed intensity adjustments come from the same distribution, using a significance level of $\alpha = 0.05$. The alternative hypothesis (H_1) "*less*" was tested, where the distribution of the first group (using the left methodology in pairwise combinations) is stochastically less. Besides, the Bonferroni correction was applied to adjust the confidence interval, α, making the criterion for significance stricter. Fig. 8 depicts the $p-values$ obtained from the Wilcoxon test as a heatmap. In this representation, a $p-value \leq 0.05$ (blue tonality) indicates a rejection of H_0. Additionally, the $p-values$ that meet the Bonferroni correction criteria ($p-value \leq 0.005$) are highlighted with yellow edges.

In addition, Cliff's Delta effect size measure was calculated. The obtained value indicates the difference in the probability of one group being larger than the other. In Fig. 9, Cliff's Delta values are represented as a heatmap, ranging from -1 to 1. Positive values suggest that observations from the first group are greater than those from the second group (indicated by blue tones), while negative values indicate the opposite (red tones). The intensity of the color represents the magnitude of the value; the larger the value, the greater the difference between the two groups.

Firstly, it can be observed that both tests exhibit similar patterns, thereby corroborating and complementing the information. In broad terms, the various negligible outcome values in the nnU-Net comparisons, as shown in Fig. 8, indicate a minimal influence of the intensity regularization strategy used across the three pathologies considered. Complementing this information with Fig. 9, it is evident that for GBM, the KDE standardization stands out compared to the others. However, for MS and epilepsy, there are no clear indicators of a superior method, and the magnitude of Cliff's Delta values remains small. This suggests that the nnU-Net outcomes are quite similar, indicating its robustness against input fluctuations. WNet, as illustrated in Figs. 8 and 9, exhibits some statistical significance spread. This demonstrates

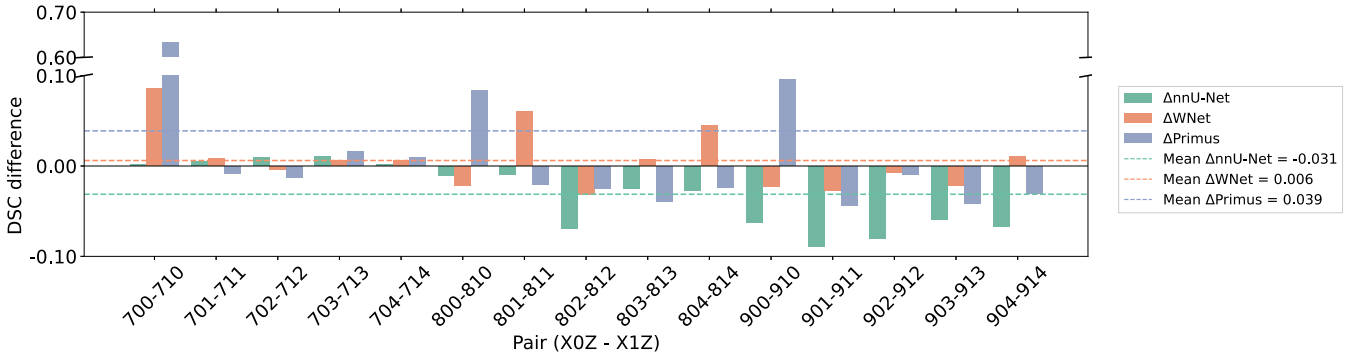
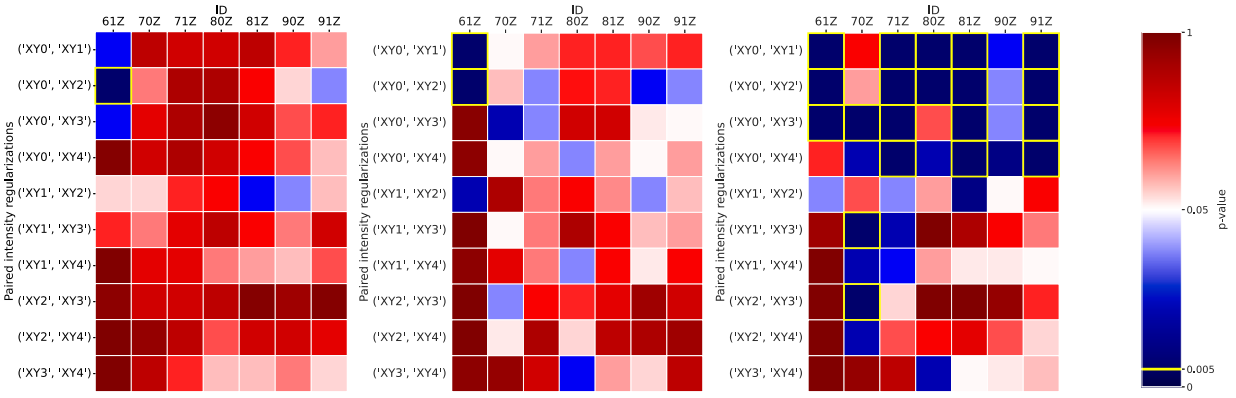
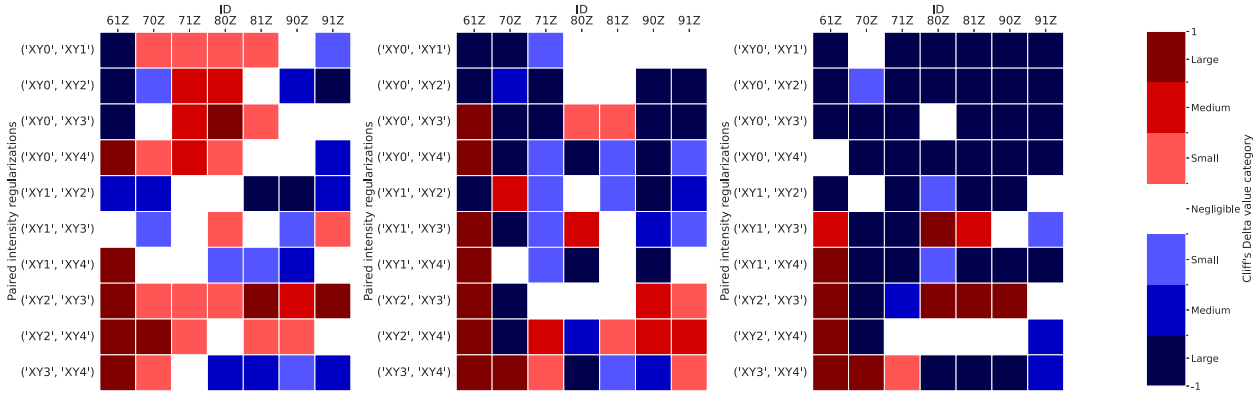


Fig. 7. Absolute difference between performance obtained using a unimodal input, i.e., FLAIR scans; and multimodal, FLAIR and T1-w sequences in the MSLesSeg and Epilepsy datasets. Experiment codes are detailed in Table 2.



(a) nnU-Net (b) WNet (c) Primus

Fig. 8. Heatmap of $p-values$ from the Wilcoxon rank-sum test with H_1 : "*less*". Statistical significance is indicated by blue tones (p-value < 0.05) and yellow edges for Bonferroni correction (p-value < 0.005). Experiment codes are detailed in Table 2.



(a) nnU-Net (b) WNet (c) Primus

Fig. 9. Heatmap corresponding to Cliff's Delta values, indicating effect sizes as "negligible", "small", "medium", or "large". Blue tones represent higher values in the first group, while red tones indicate higher values in the second group. Experiment codes are detailed in Table 2.

that the KDE, Ws, and Zs outperformed both the None and MinMax options.

Finally, the Primus analysis highlights several significant differences, even applying the Bonferroni correction, as shown in Fig. 8. It also obtains high magnitude values in Cliff's Delta measure, as depicted in Fig. 9, which can be attributed to larger spread values and some anomalous training examples. This indicates a general sensitivity to inputs and unstable behavior. Nonetheless, in the GBM results, KDE once again outperformed other methods, and in the cases of MS and epilepsy, both the Ws and Zs also demonstrate strong performances.

4. Conclusions

This work presents a comprehensive study on the effects of intensity value regularizations and the number of input channels – whether unimodal or multimodal– across CNN-based, TF-based, and hybrid models for 3D MRI segmentation of three neurological disorders: glioblastoma, multiple sclerosis, and epilepsy.

The experiments conducted across the nnU-Net, WNet, and Primus architectures indicate that proper intensity processing is essential. Among the intensity regularization methods tested in this study, KDE, White-stripe, and Z-score standardizations proved to be particularly effective. Particularly, KDE and White-stripe are designed to exploit anatomical MRI properties, supporting the need not to isolate images from their context and consider the biological meaning of intensity values.

A statistical analysis of the results showed that, regardless of the pathology, nnU-Net is the most robust architecture against intensity variability, with small improvements of around 3%. Meanwhile, methods incorporating TF elements are more sensitive to these variations. WNet exhibits slightly larger gains by about 6%. Nonetheless, for the epilepsy dataset, WNet outperformed nnU-Net by 30%. On the other hand, Primus can be more unstable and underperform compared to nnU-Net and WNet in most cases. Nonetheless, it remains a promising and competitive option in medical imaging due to its TF elements, which allow for the easy integration of other clinical data formats. Additionally, it has been demonstrated that incorporating an extra channel may decrease performance while increasing computational cost.

The findings can be summarized as follows: intensity regularization is essential, adding more information does not guarantee better performance, and CNN-based methods remain the benchmark in medical imaging segmentation.

Future work should explore additional segmentation architectures, such as mamba-based models, and focus on improving epilepsy performance through hyperparameter optimization algorithms or data augmentation techniques using generative models – diffusion models or GAN-based – methods should be applied to enhance results in epilepsy cases.

CRediT authorship contribution statement

Ariadna Jiménez-Partinen: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization; **Ezequiel López-Rubio:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization; **Fátima Nagib-Raya:** Writing – review & editing, Writing – original draft, Validation, Investigation, Conceptualization; **Esteban J. Palomo:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization; **Rafael M. Luque-Baena:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Data availability

Open-access dataset were used and correctly referenced.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is partially supported by the Ministry of Science and Innovation of Spain, grant number PID2022-136764OA-I00, project name Automated Detection of Non Lesional Focal Epilepsy by Probabilistic Diffusion Deep Neural Models. It includes funds from the European Regional Development Fund (ERDF). It is also partially supported by the Fundación Unicaja under project PUNI-003_2023, project name Intelligent System to Help the Clinical Diagnosis of Non-Obstructive Coronary Artery Disease in Coronary Angiography, and the Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma BIONAND under project ATECH-25-02. The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. They also gratefully acknowledge the support of NVIDIA Corporation with the donation of an RTX A6000 GPU with 48Gb. The authors also thankfully acknowledge the grant of the Universidad de Málaga and the Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma BIONAND. Funding for open access charge: Universidad de Málaga/CBUA.

References

[1] A. Kujur, Z. Raza, A.A. Khan, C. Wechtaisong, Data complexity based evaluation of the model dependence of brain MRI images for classification of brain tumor and alzheimer's disease, IEEE Access 10 (2022) 112117–112133.
[2] M.S. Sheela, G. Amirthayogam, J.J. Hephzipah, et al., Advanced brain tumor classification using DEEPBELEIF-CNN method, Babylonian J. Mach. Learn. 2024 (2024) 89–101.
[3] Q. Liu, M. Liu, Y. Zhu, L. Liu, Z. Zhang, Y. Wang, DAUNet: A deformable aggregation UNet for multi-organ 3D medical image segmentation, Pattern Recognit. Lett. 191 (2025) 58–65.
[4] M.M. Mijwil, R. Doshi, K.K. Hiran, O.J. Unogwu, I. Bala, Mobilenetv1-based deep learning model for accurate brain tumor classification, Mesopotamian J. Comput. Sci. 2023 (2023) 29–38.
[5] S. Marina, Improving diagnostic accuracy of brain tumor MRI classification using generative AI and deep learning techniques, Babylonian J. Artif. Intell. 2025 (2025) 55–63.
[6] M. Zhang, H. Yu, G. Cao, J. Huang, Y. Cheng, W. Zhang, X. Yuan, R. Yang, Q. Li, L. Cai, et al., Three-branch feature enhancement and fusion network for focal cortical dysplasia lesions segmentation using multimodal imaging, Brain Res. Bull. 222 (2025) 111268.
[7] R.A. Zeineldin, M.E. Karar, Z. Elshaer, J. Coburger, C.R. Wirtz, O. Burgert, F. Mathis-Ullrich, Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI, Sci. Rep. 14 (1) (2024) 3713.
[8] X. Zhang, Y. Zhang, C. Wang, L. Li, F. Zhu, Y. Sun, T. Mo, Q. Hu, J. Xu, D. Cao, Focal cortical dysplasia lesion segmentation using multiscale transformer, Insights Imag. 15 (1) (2024) 222.
[9] Y. Amri, A.B. Slama, Z. Mbarki, R. Selmi, H. Trabelsi, Automatic glioma segmentation based on efficient U-net model using MRI images, Intell.-Based Med. 11 (2025) 100216.
[10] N. Mushtaq, A.A. Khan, F.A. Khan, M.J. Ali, M.M.A. Shahid, et al., Brain tumor segmentation using multi-view attention based ensemble network, Comput. Mater. Cont. 72 (3) (2022).
[11] D. Rastogi, P. Johri, M. Donelli, S. Kadry, A.A. Khan, G. Espa, P. Feraco, J. Kim, Deep learning-integrated MRI brain tumor analysis: feature extraction, segmentation, and survival prediction using replicator and volumetric networks, Sci. Rep. 15 (1) (2025) 1437.
[12] P. Belwal, S. Singh, U-Net Approach to MRI-Based segmentation of multiple sclerosis lesions, Procedia Comput. Sci. 259 (2025) 1316–1325.
[13] R. Kadri, B. Bouaziz, M. Tmar, F. Gargouri, Innovative multi-modal approaches to alzheimer's disease detection: transformer hybrid model and adaptive MLP-Mixer, Pattern Recognit. Lett. 190 (2025) 15–21.
[14] M. Soni, M.A. Shnan, Scalable neural network algorithms for high dimensional data, Mesopotamian J. Big Data 2023 (2023) 1–11.
[15] J.C. Reinhold, B.E. Dewey, A. Carass, J.L. Prince, Evaluating the impact of intensity normalization on MR image synthesis, in: Proceedings of SPIE–the International Society for Optical Engineering, 10949, 2019, p. 109493H.
[16] A. Ghazvanchahi, P.J. Maralani, A.R. Moody, A. Khademi, Effect of intensity standardization on deep learning for WML segmentation in multi-centre FLAIR MRI, in: Medical Imaging with Deep Learning, PMLR, 2024, pp. 1923–1940.
[17] N. Jacobsen, A. Deistung, D. Timmann, S.L. Goericke, J.R. Reichenbach, D. Guellmar, Analysis of intensity normalization for optimal segmentation performance of a fully convolutional neural network, Zeitschrift für Medizinische Physik 29 (2) (2019) 128–138.
[18] E. Kondrateva, P. Druzhinina, A. Dalechina, S. Zolotova, A. Golanov, B. Shirokikh, M. Belyaev, A. Kurmukov, Negligible effect of brain MRI data preprocessing for tumor segmentation, Biomed. Signal Process. Control 96 (2024) 106599.

[19] J. Shin, Revolutionizing medical imaging with artificial intelligence real-Time segmentation for enhanced diagnostics, EDRAAK 2024 (2024) 18–25.
[20] J.A. Schwartzbaum, J.L. Fisher, K.D. Aldape, M. Wrensch, Epidemiology and molecular pathology of glioma, Nature Clin. Pract. Neurol. 2 (9) (2006) 494–503.
[21] S. Bakas, C. Sako, H. Akbari, M. Bilello, A. Sotiras, G. Shukla, J.D. Rudie, N.F. Santamaría, A.F. Kazerooni, S. Pati, et al., The university of pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics, Sci. Data 9 (1) (2022) 453.
[22] H. Lassmann, J. Van Horssen, D. Mahad, Progressive multiple sclerosis: pathology and pathogenesis, Nature Rev. Neurol. 8 (11) (2012) 647–656.
[23] A. Rondinella, F. Guarnera, E. Crispino, G. Russo, C. Di Lorenzo, D. Maimone, F. Pappalardo, S. Battiato, Icpr 2024 competition on multiple sclerosis lesion segmentation-methods and results, in: International Conference on Pattern Recognition, Springer, 2024, pp. 1–16.
[24] F. Guarnera, A. Rondinella, E. Crispino, G. Russo, C. Di Lorenzo, D. Maimone, F. Pappalardo, S. Battiato, MSLesSeg: baseline and benchmarking of a new multiple sclerosis lesion segmentation dataset, Sci. Data 12 (1) (2025) 1–10.
[25] T.M. Salmenpera, J.S. Duncan, Imaging in epilepsy, J. Neurol. Neurosur. Psychiatry 76 (suppl 3) (2005) iii2–iii10.
[26] T. Bast, G. Ramantani, A. Seitz, D. Rating, Focal cortical dysplasia: prevalence, clinical presentation and epilepsy in children and adults, Acta Neurol. Scand. 113 (2) (2006) 72–81.
[27] I. Najm, D. Lal, M. Alonso Vanegas, F. Cendes, I. Lopes-Cendes, A. Palmini, E. Paglioli, H.B. Sarnat, C.A. Walsh, S. Wiebe, et al., The ILAE consensus classification of focal cortical dysplasia: an update proposed by an ad hoc task force of the ILAE diagnostic methods commission, Epilepsia 63 (8) (2022) 1899–1919.
[28] F. Schuch, L. Walger, M. Schmitz, B. David, T. Bauer, A. Harms, L. Fischbach, F. Schulte, M. Schidlowski, J. Reiter, et al., An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II, Sci. Data 10 (1) (2023) 475.
[29] N.J. Tustison, P.A. Cook, A.J. Holbrook, H.J. Johnson, J. Muschelli, G.A. Devenyi, J.T. Duda, S.R. Das, N.C. Cullen, D.L. Gillen, et al., The ANTsx ecosystem for quantitative biological and medical imaging, Sci. Rep. 11 (1) (2021) 9068.
[30] S. Song, Y. Zheng, Y. He, A review of methods for bias correction in medical images, Biomed. Eng. Rev. 1 (1) (2017).
[31] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, J.C. Gee, N4ITK: Improved N3 bias correction, IEEE Trans. Med. Imag. 29 (6) (2010) 1310–1320.
[32] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, NnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[33] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, P.F. Jaeger, Nnu-net revisited: a call for rigorous validation in 3d medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 488–498.
[34] L. Huang, A. Miron, K. Hone, Y. Li, Segmenting medical images: from UNet to res-UNet and nnUNet, in: 2024IEEE 37Th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2024, pp. 483–489.
[35] Y. Zhou, L. Li, L. Lu, M. Xu, NnWNet: rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 20852–20862.
[36] T. Wald, S. Roy, F. Isensee, C. Ulrich, S. Ziegler, D. Trofimova, R. Stock, M. Baumgartner, G. Köhler, K. Maier-Hein, Primus: Enforcing attention usage for 3d medical image segmentation, (2025). arXiv preprint arXiv:2503.01835
[37] F. Wilcoxon, Individual comparisons by ranking methods, in: Breakthroughs in Statistics: Methodology and Distribution, Springer, 1992, pp. 196–202.