ISSN (Print) 1226-1750
ISSN (Online) 2233-6656
Journal of Magnetics 30(4), 849-858 (2025) https://doi.org/10.4283/JMAG.2025.30.4.849

Effects of Loss Functions on MRI Segmentation Accuracy: A Large-Scale Comparative Study

Tae-Gyu Kim[1†], Seung min Hwang[1†], Seojae Jeon[1,2*], and Jang Woo Park[3,4*]

[1]Dept. of Biomedical Engineering, Chonnam National University, Republic of Korea
[2]Korea Institute of Integrated Medical Research, Republic of Korea
[3]Korea Radioisotope Center for Pharmaceuticals, Korea Institute of Radiological & Medical Sciences, Republic of Korea
[4]School of Biomedical Engineering, Chonnam National University, Republic of Korea

(Received 7 November 2025, Received in final form 11 November 2025, Accepted 15 December 2025)

Accurate brain tissue segmentation is essential for constructing individualized head models in noninvasive brain stimulation. This study investigated how loss function design influences segmentation performance using a large dataset of over 500 T1-weighted MRIs. A 3D U-Net was trained to segment the skin, skull, cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM) using either Dice or combined Dice + Cross-Entropy (Dice+CE) losses. Quantitative results showed that Dice+CE achieved higher overall accuracy (0.8577 vs. 0.8550) and mean IoU (0.2863 vs. 0.2716), with notable improvements in low-contrast tissues such as skin, skull, CSF, and GM. WM slightly favored Dice due to its large and homogeneous structure. Qualitative analysis indicated clearer tissue boundaries and fewer misclassifications with Dice+CE. These findings demonstrate that combining Dice and Cross-Entropy losses enhances segmentation accuracy and provide evidence for optimizing brain tissue segmentation model performance.

Keywords : brain tissue segmentation, personalized neuromodulation, electro-magnetic stimulation, magnetic resonance imaging, 3D U-Net

1. Introduction

Recent advances in non-invasive brain stimulation (e.g., transcranial electric stimulation) emphasize personalized approaches to improve clinical reproducibility, reduce inter-subject variability and tailor stimulation to individual neuroanatomy. Personalized stimulation requires that the head model accurately reflects an individual's anatomical structures, which in turn depends on precise tissue segmentation of three-dimensional images such as MRI and CT. Segmentation is the first step in constructing computational head models for electric-field simulations. it extracts key tissues including the scalp, skull, cerebrospinal fluid (CSF), grey matter and white matter from structural MRI or CT images [1]. Accurate segmentation is assessed using metrics such as the Dice similarity coefficient and the modified Hausdorff distance [1], and errors propagate through subsequent modelling stages. Single-modality images may struggle to differentiate certain boundaries; for example, a T1-weighted MRI separates soft tissues well but makes it difficult to distinguish the Skull, CSF interface. T2-weighted MRI highlights CSF, while CT excels at delineating bone. Therefore, combining T1 with T2 or CT images improves skull and CSF segmentation but requires careful co-registration [2]. In tumour segmentation tasks, multimodal MRI sequences (T1, T2, contrast-enhanced T1 and FLAIR) provide complementary information that improves segmentation accuracy compared with single-modality images [3].

Accurate delineation of anatomical structures directly influences diagnostic and therapeutic precision and enables time-efficient workflows. Early clinical practice relied on manual segmentation, which requires clinicians to outline structures on every slice (often hundreds of images) and is therefore labour-intensive, time-consuming and subject to inter and intra-observer variability [4]. Classical automated methods such as thresholding,

*Corresponding author: Tel: +82-2-970-8927
Fax: +82-2-970-8989, e-mail: fr1771@chonnam.ac.kr
e-mail: jangwoo@kirams.re.kr
†These authors contributed equally to this work

edge-based and region-growing algorithms or atlas registration improved efficiency but remained sensitive to noise and intensity inhomogeneity and often produced unstable results [5]. Moreover, traditional methods struggle with small tissues and ambiguous boundaries. Small organs occupy only a tiny fraction of the image volume and have unclear boundaries, leading to poor segmentation performance [5].

Deep learning based segmentation, particularly convolutional neural networks (CNNs), has become the dominant approach in recent years. Three dimension(3D) U-Net and its variants use encoder–decoder architectures with skip connections to capture both contextual and fine-grained features and have become the de-facto standard for volumetric CT and MRI segmentation across a wide range of tasks [5]. Compared with manual or classical algorithms, CNN-based segmentation offers superior accuracy and efficiency, substantially reducing processing time and allowing real-time or near-real-time applications. In the context of non-invasive brain stimulation, precise segmentation of tissues such as bone, fat and air spaces is essential because the electrical conductivity of each tissue influences current flow; detailed segmentation also enables modelling of small structures (e.g., blood vessels) that were previously ignored or required contrast agents [1]. Nevertheless, challenges remain to accurately segmenting tiny structures and low-contrast boundary regions, accommodating anatomical variability across individuals, choosing between single or multi-modal imaging inputs, improving model performance while minimizing computation time and memory consumption, and coping with scanner differences or patient motion [5]. Addressing these issues remains an active area of research.

In current practice, many models segment white matter, grey matter, CSF, bone, fat/muscle and ocular structures using single images (T1, T2 or CT) or combining MRI with CT. Among them, 3D U-Net based models are widely used owing to their strong performance [5]. A critical factor in these models is the loss function, which guides how the network learns from segmentation errors. Popular choices include the Dice loss, the cross-entropy loss and its combinations. The Dice loss maximizes overlap between predicted and reference regions but may be unstable or insensitive to boundary errors. The cross-entropy loss treats segmentation as voxel-wise classification but can be dominated by majority classes. Combining the two losses leverages complementary strengths and mitigates class imbalance. For example, researchers have defined a composite loss as the sum of Dice and cross-entropy losses and reported robust performance across diverse class distributions [6]. Another study noted that cross-entropy alone was influenced by mainstream pixels and that adding the Dice term stabilized training and improved segmentation accuracy [7]. In a recent multimodal brain-tumour segmentation model (ResSAXU-Net), a fusion loss combining Dice and cross-entropy losses helped address network convergence and data imbalance, yielding Dice similarity coefficients above 0.95 across tumor subregions [8]. These studies suggest that combining multiple loss functions can improve segmentation performance, however most prior works have focused on specific lesions or limited datasets. There is still a lack of systematic, large-scale analyses investigating how the choice of loss function affects overall segmentation accuracy and boundary awareness in general MRI based brain tissue segmentation. In addition, existing studies have primarily concentrated on internal brain structures such as GM, WM and CSF [9], while cases that perform simultaneous segmentation of the entire head including skin and skull using only MRI data remain very limited [10].

Therefore, in this study, we applied a large-scale MRI dataset and a 3D U-Net architecture to compare the effects of the Dice loss function and the combined Dice + Cross-Entropy loss function on the segmentation of major brain tissues, including skin, skull, CSF, GM, and WM. Through this comparison, we aimed to analyze differences in segmentation accuracy and boundary recognition across tissues depending on the loss function and to provide insights for optimizing brain tissue segmentation model performance.

2. Methods

Model training used the publicly available IXI dataset [11]. The cohort comprises approximately 582 healthy volunteers and includes multiple MRI sequences (T1, T2,

Table 1. Information of publicly available IXI dataset.

Parameter	Description
Number of subjects	582 healthy volunteers
Imaging sequences	T1, T2, Proton Density, MRA, DTI
Scanner details	Hammersmith Hospital: Philips 3T Guy’s Hospital: Philips 1.5T Institute of Psychiatry: GE 1.5T
Image Data format	NIfTI-1
Original Image size	256 × 256 × 150
Spatial resolution (Voxel size)	0.96 × 0.96 × 1.2 mm^3
Data used in this study	552 T1-weighted volumes

Proton Density weighted, MRA, and DTI). Scans were acquired on three scanners across London sites: Hammersmith Hospital (Philips 3T), Guy's Hospital (Philips 1.5T), and the Institute of Psychiatry (GE 1.5T) and are distributed in NIfTI-1 format. Native volumes measure 256 × 256 × 150 voxels with a voxel size of 0.96 × 0.96 × 1.2 mm^3. In this study, only T1-weighted images were considered, and a total of 552 T1 volumes were used for model development. Information about the detailed MRI image data set is shown in Table 1.

2.1. Data Labeling

For personalized neuromodulation, five tissue classes were segmented: skin, skull, cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM). Training labels were generated with an automatic pipeline built from open-source tools. Extracranial tissues and CSF were obtained using the atlas-based charm workflow in SimNIBS v4.5 [12], which applies deformable atlas registration to produce subject specific label maps. Parenchymal tissues (GM, WM) were derived with FastSurfer [13], a CNN-based neuroanatomical segmentation framework, from which GM and WM masks were extracted. The cerebellum was not treated as a separate class and was merged into GM or WM as appropriate.

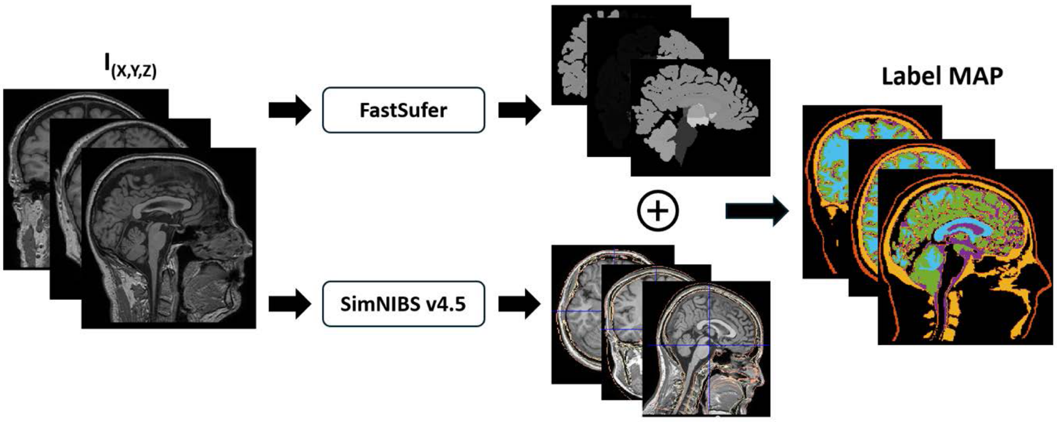

Fig. 1. (Color online) Labeling workflow with Fastsurfer & SImNIBS.

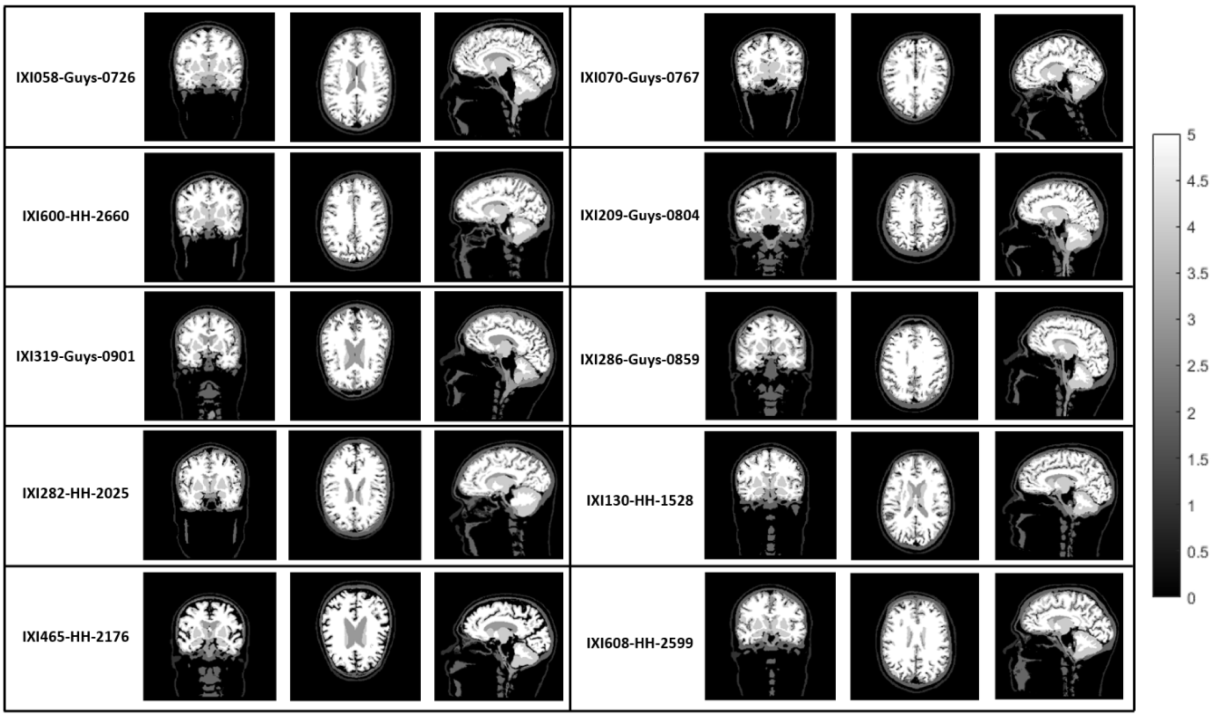

Fig. 2. Labeled images which is randomly sampled from 552 MRI dataset.

Table 2. Label information: label indices for background, skin, skull, cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM).

Tissue	Labeling Number
Background	0
Skin	1
Skull	2
CSF	3
Gray Matter	4
White Matter	5

Figure 1 illustrates the workflow: FastSurfer provides GM/WM masks, SimNIBS-*charm* provides skin, skull, CSF masks, and the outputs are fused to obtain the final label map. For quality control, ten labeled volumes were randomly sampled for visual inspection. Representative examples are shown in Fig. 2. Class indices for each tissue are summarized in the accompanying Table 2.

2.1.1. Automated Post-processing.

It is important to note that no manual corrections or human editing were applied to the labels. However, to refine the quality of the automatically fused labels, a targeted automated post-processing step was implemented using MATLAB. This procedure was specifically applied to the skin and skull tissue classes to mitigate artifacts from the initial automated segmentation tools. For the skin mask, a connected-component analysis was performed to identify and retain only the largest connected structure, effectively removing smaller, isolated noise components. Similarly, for the skull mask, noise reduction was achieved by applying a volume threshold, where all connected components consisting of fewer than 70 voxels were programmatically removed. This automated cleaning ensured cleaner segmentation boundaries for these complex tissues prior to model training.

2.2. Data preprocessing

Model development was conducted in MATLAB R2024a. Prior to training, all input data underwent preprocessing, performed entirely within the MATLAB environment. No data augmentation was applied during either the training or testing phases.

2.2.1. Intensity normalization

All native T1-weighted MRI volumes were min–max normalized to the [0,1] range prior to training to reduce inter-scan intensity variability, improve numerical stability, and avoid on-the-fly scaling (thereby lowering GPU memory load). Because raw MR intensities vary with acquisition settings and hardware, this preprocessing step was applied uniformly across the dataset to mitigate distributional shifts and support better generalization. Min–max normalization was computed as:

$$I_{norm}(X,Y,Z) = \frac{I_{(X,Y,Z)} - I_{min}}{I_{max} - I_{min}} \tag{1}$$

where $I(X, Y, Z)$ denotes the original voxel intensity at location (X, Y, Z), and I_{min} and I_{max} are the minimum and maximum intensities within the volume, respectively.

2.2.2. Spatial preprocessing (Image Resizing and Padding)

All T1-weighted MRI volumes were resampled to an isotropic voxel size of 1×1×1 mm³ to harmonize spatial resolution across datasets. Each volume was then standardized to a uniform input shape by zero-padding to 256×256×256 voxels, yielding consistent dimensions across the axial, coronal, and sagittal planes. This procedure minimizes inter-scan spatial mismatch and enforces a consistent input format for both model training and downstream deployment.

2.3. Network architecture: 3D U-Net

A three-dimensional(3D) U-Net was implemented to achieve high-fidelity tissue segmentation in volumetric MRI [14]. The model extends the canonical 2D U-Net to 3D, thereby preserving through-slice spatial context that is essential for analyzing voxelwise volumes. The network follows an encoder–decoder topology: the encoder stacks 3D convolutions with batch normalization and ReLU activations, interleaved with max-pooling for hierarchical feature abstraction. The decoder employs transposed up-convolutions and skip connections to progressively restore spatial resolution while retaining fine-grained localization. Skip connections concatenate shallow, spatially precise features with deeper, semantically rich representations, improving delineation of small structures and ambiguous boundaries.

Input volumes were partitioned into 64×64×64 voxel patches sampled from the preprocessed 256×256×256 T1-weighted images [15]. The segmentation target comprised six classes: background, skin, skull, cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM). A voxelwise softmax at the output layer produced class probabilities. Two training objectives were examined: Dice loss and a composite Dice–cross-entropy loss, enabling direct comparison of their effects on segmentation performance.

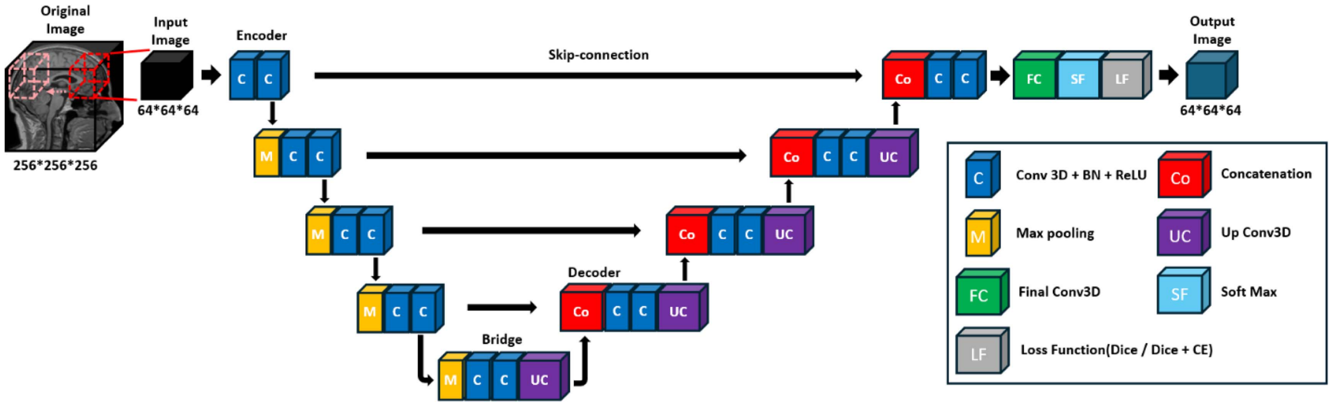

Fig. 3. (Color online) Patch-based 3D U-Net for MRI tissue segmentation. From each preprocessed T1-weighted volume (256×256×256), cubic patches (64×64×64) are sampled as inputs.

The 3D U-Net backbone affords greater through-slice continuity and volumetric precision than 2D architectures, which is advantageous for delineating ambiguous boundaries and small structures in medical images. Accordingly, a 3D U-Net was trained for MRI tissue segmentation, and performance differences attributable to the loss formulation were systematically evaluated. The network configuration and input strategy are illustrated in Fig. 3.

2.4. Training configuration

A patch-based 3D U-Net was trained on 552 T1-weighted MRI volumes. Subjects were partitioned into training/validation/test sets in a 7:2:1 ratio (386/110/56), ensuring no subject overlap across splits. Each volume was preprocessed to 256×256×256 voxels, from which cubic input patches of 64×64×64 voxels were uniformly sampled. To mitigate class imbalance, a center-voxel label constraint was imposed so that sampled patches were evenly distributed across background, skin, skull, CSF, GM, and WM [15]. Training was executed on a work-station equipped with two Intel Xeon Gold 6526Y CPUs, four NVIDIA RTX 4000 Ada GPUs (20 GB each total GPU memory 80 GB), and 2 TB system RAM (32×64 GB RDIMMs). Optimization used Adam with an initial learning rate of 1×10^{-5} and a mini-batch size of 24. For each epoch, 24 patches were drawn per volume from the training set; validation was performed every 50 iterations to monitor convergence and select checkpoints. Two loss formulations were evaluated: (i) pure Dice loss and (ii) a composite Dice+cross-entropy objective. For each loss, models were trained for 30 and 50 epoch to assess the effect of training duration on performance. The learning options referenced above are summarized and presented in Table 3.

Table 3. Comparative Training Setup: Dice-only vs. Hybrid (Dice+CE) Losses.

Training Option	Baseline (Dice)	Hybrid (Dice+CE)
Model type	3D U-Net	
Data Set	552 T1 MRI Image(3D voxel image) 7(Training/386) : 2(Validation/110) : 1(Test/56)	
Optimizer	Adam	
learning rate	1e-5	
Validation Frequency	50	
Mini BatchSize	24	
Input Size	64 × 64 × 64	
Epoch	30	50
Loss function	Dice Loss	Dice + Cross-Entropy Loss

2.5. Evaluation metrics

The effect of the loss function on MRI tissue-segmentation performance was evaluated using multiple metrics: global accuracy, per-class accuracy, mean accuracy, the intersection over union (IoU), and per-class IoU. Definitions of these metrics are provided below.

2.5.1. Global Accuracy/Accuracy/Mean Accuracy

Accuracy quantifies the proportion of correctly classified voxels among all voxels. Global accuracy evaluates this quantity over the entire segmentation irrespective of class, whereas per-class accuracy is computed separately for each class using the same definition. Let TPc denote the number of voxels correctly assigned to class C, FPc the number of voxels incorrectly predicted as class C, and FNc the number of voxels that truly belong to class C but were predicted otherwise. Then

$$GlobalAccuracy = \frac{\sum_C TP_C}{\sum_C TP_C + \sum_C FN_C} \quad (2)$$

Per-class accuracy denotes, for a given class C, the proportion of voxels correctly identified as C:

$$Accuracy_C = \frac{TP_C}{TP_C + FN_C} \quad (3)$$

This metric indicates how accurately the model recognizes voxels of class C. The mean accuracy is defined as the unweighted average of per-class accuracies across all C classes:

$$Mean\ Accuracy = \frac{1}{C}\sum_{C=1}^{C} Accuracy_C \quad (4)$$

2.5.2. IOU/Mean IOU

The intersection over union (IoU), also known as the Jaccard index, quantifies the degree of overlap between the predicted and reference regions. At the voxel level for class cc, let TP_C, FP_C, and FN_C denote the numbers of true-positive, false-positive, and false-negative voxels, respectively. The classwise IoU is defined as:

$$IoU_C = \frac{FN_C}{TP_C + FP_C + FN_C} \quad (5)$$

The mean IoU is the unweighted average across all CC classes:

$$Mean\ IoU = \frac{1}{C}\sum_{C=1}^{C} IoU_C \quad (6)$$

2.6. Statistical Analysis

To rigorously determine whether the observed differences in segmentation performance between the single Dice loss and the composite Dice + Cross-Entropy (CE) loss were statistically meaningful, a formal statistical analysis was conducted on the per-subject metrics obtained from the 56 test volumes. Since both loss functions were applied to the exact same set of subjects, the performance comparison constitutes a paired data scenario. Consequently, a paired two-sample t-test was chosen to assess the difference in mean performance for all key metrics, including Global Accuracy, Mean Dice Score, and Per-class IoU. This test is appropriate for determining if the mean difference between the paired observations is significantly non-zero, provided the differences are approximately normally distributed. The analysis aimed to test the following hypotheses: the Null Hypothesis H_0 stated that the mean performance difference between the two loss functions is zero (i.e., there is no significant difference); conversely, the Alternative Hypothesis (H_1) stated that the mean performance difference is not zero (i.e., there is a statistically significant difference). All statistical calculations were performed using the MATLAB R2024a environment. Consistent with conventional statistical practice in medical imaging research, a result was deemed statistically significant if the P-value was less than 0.05 ($p < 0.05$).

3. Results

Performance was compared between a Dice-only model and a hybrid Dice+cross-entropy (Dice+CE) model on a held-out test set of 56 T1-weighted MRI volumes. Global accuracy exceeded 0.85 for both models, with a marginal gain for Dice+CE (0.8577) over Dice-only (0.855; Δ = +0.0027).

Per-class accuracies showed broadly consistent improvements for Dice+CE in non-background tissues. Background accuracy was essentially unchanged across models. In contrast, Dice+CE yielded higher accuracies for skin (0.1479 vs. 0.0812; Δ = +0.0667), skull (0.2430 vs. 0.1925; Δ = +0.0505), CSF, and GM. White matter presented an exception, where the Dice-only model performed better (0.5084 vs. 0.4779; Δ = -0.0305). Overall, the hybrid loss offered small gains in global accuracy and more consistent improvements across most tissue classes, with a trade-off observed in WM.

3.1. Intersection-over-Union (IoU)

Class-wise IoU generally favored the hybrid Dice+CE model over the Dice-only baseline. Notable gains were observed for skin (0.090 vs. 0.0609; Δ =+0.0291$), skull (0.141 vs. 0.1104; Δ = +0.0306), CSF (0.064 vs. 0.0562; Δ = +0.0078), and GM (0.2082 vs. 0.1915; Δ = +0.0167), with a modest improvement also seen for background. White matter was the only exception, where performance was essentially unchanged, with a slight advantage for Dice-only (0.3129 vs. 0.311; Δ = -0.0019). Overall, IoU results corroborate the accuracy findings, indicating more consistent tissue-wise gains from the composite loss.

Classwise IoU likewise favored the hybrid Dice+CE loss over the Dice-only baseline. Improvements were observed for skin (0.090 vs. 0.0609 Δ=+0.0291), skull (0.141 vs. 0.1104 Δ=+0.0306), CSF (0.064 vs. 0.0562 Δ=+0.0078), and GM (0.2082 vs. 0.1915 Δ=+0.0167), along with a modest gain for background. White matter

Table 4. Comparison of Segmentation Performance Metrics Between Dice Loss and Dice–Cross-Entropy Loss.

Tissue / Metric	Accuracy (Dice)	Accuracy (Dice+CE)	IOU (Dice)	IOU (Dice+CE)
Back Ground	0.9557	0.9545	0.8974	0.9035
Skin	0.0812	0.1479	0.0609	0.09
Skull	0.1925	0.243	0.1104	0.141
CSF	0.094	0.1045	0.0562	0.064
GM	0.3405	0.3557	0.1915	0.2082
WM	0.5084	0.4779	0.3129	0.311
Global Accuracy	0.855 ± 0.0178	0.8577 ± 0.0213	-	-
Mean Accuracy	0.362 ± 0.0581	0.3806 ± 0.0745	-	-
Mean IOU	-	-	0.2716 ± 0.0468	0.2863 ± 0.0639

was the sole exception: the two models performed similarly, with a slight advantage for Dice-only (0.3129 vs. 0.311 Δ=−0.0019). Overall, IoU results are consistent with the accuracy analysis, indicating more reliable tissue-wise segmentation from the composite loss.

Taken together, both the mean accuracy (0.3806 vs. 0.3620) and the mean IoU (0.2863 vs. 0.2716) were higher for the hybrid Dice+CE loss than for Dice-only, indicating an overall improvement in segmentation performance. Gains were particularly pronounced for tissues with ambiguous boundaries (e.g., skin, skull, CSF, and GM), where the composite loss yielded more reliable delineation. Summary values for each evaluation metric are reported in Table 4.

Qualitative comparison was performed on four randomly sampled test volumes in Fig. 4. With Dice-only training, skin–skull interfaces frequently exhibited boundary ambiguity and cross-class leakage. Applying the hybrid

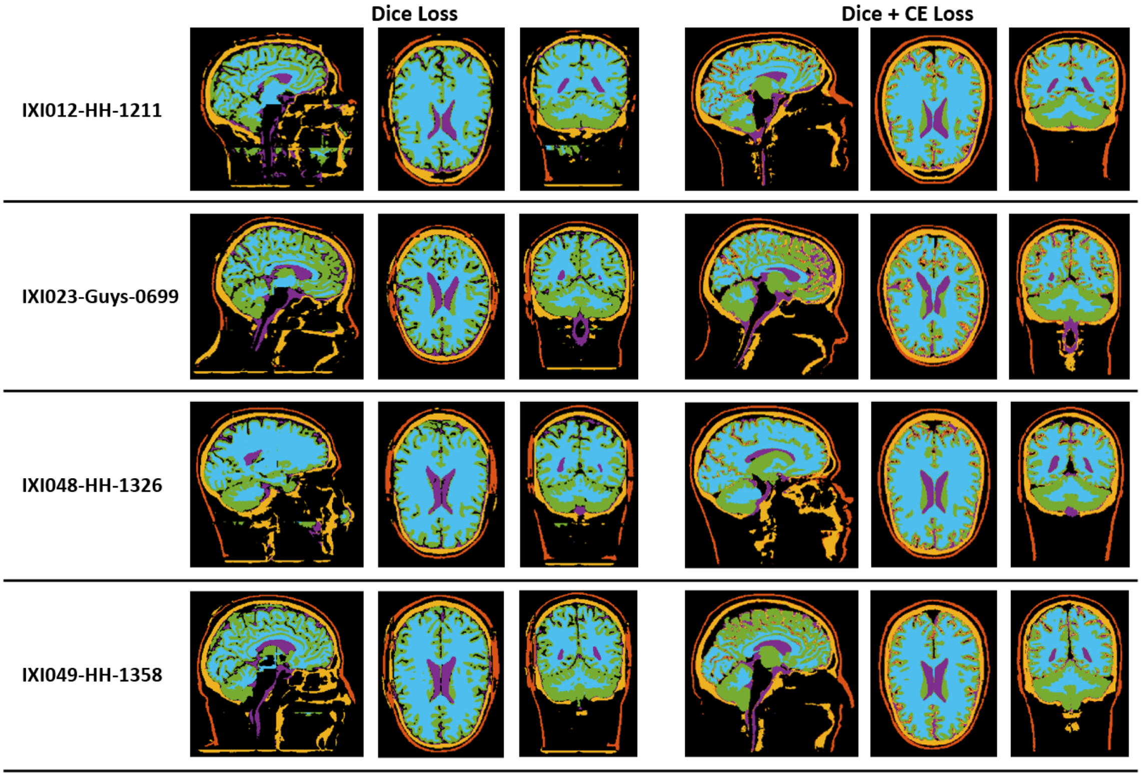

Fig. 4. (Color online) Qualitative comparison of tissue segmentation on four randomly selected test subjects (representative axial, coronal, and sagittal slices). Left: model trained with Dice loss only. Right: model trained with the hybrid Dice+cross-entropy (Dice+CE) loss.

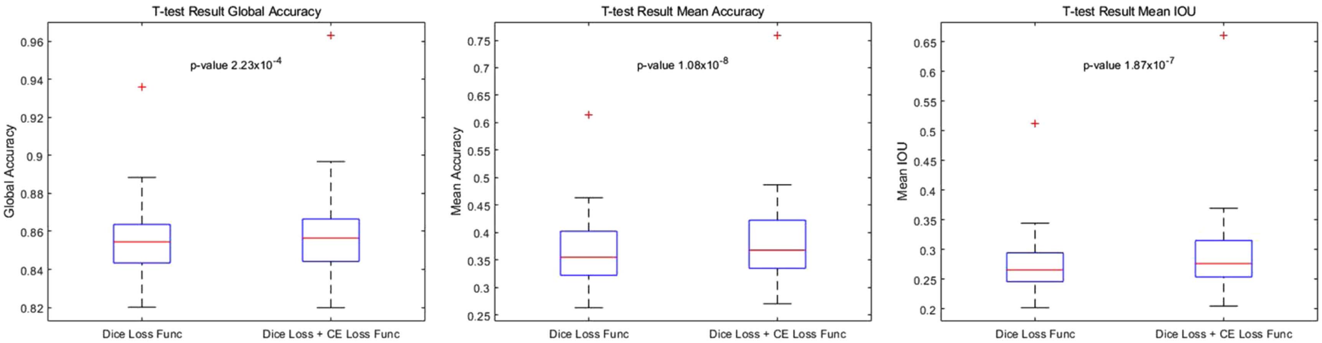

Fig. 5. (Color online) Comparison of Segmentation Accuracy Metrics between Dice Loss and Dice + Cross-Entropy Loss on 56 Test dataset.

Dice+CE loss reduced these errors; although some edge regions remained imperfect, the overall delineation was more stable and accurate than with Dice alone. For CSF, the Dice-only model occasionally produced comparatively clean contours in extracerebral spaces, but in gray and white matter the same model showed prominent mis-segmentation—extending labels into anatomically unrelated regions and creating spurious islands. These errors were substantially attenuated under Dice+CE, yielding more faithful GM/WM parcellation. In sum, the hybrid loss produced superior qualitative segmentations, particularly in boundary recognition and region-specific precision, while retaining comparable performance in easier compartments.

3.2. Statistical Significance of Performance Gains.

To confirm the robustness of the observed performance improvements, a paired two-sample t-test was conducted comparing the per-subject results from the Dice loss and the Dice + CE loss across 56 test volumes. The analysis demonstrated that the composite Dice + CE loss resulted in a statistically highly significant improvement across all primary evaluation metrics. Specifically, the gains in Global Accuracy ($p < 0.001$), Mean Accuracy ($p < 0.001$), and Mean IoU ($p < 0.001$) were all found to be highly significant, strongly supporting the claim that combining the cross-entropy term with the Dice term provides a robust enhancement to the network's segmentation capability. Full statistical results, including T-statistics and precise p-values, are detailed in Table 5.

Table 5. Paired T-test results comparing segmentation performance between Dice Loss and Dice + CE Loss (N=56).

Metric	p-value	Significance
Global Accuracy	2.23×10^{-4}	$p < 0.001$***
Mean Accuracy	1.08×10^{-8}	$p < 0.001$***
Mean IOU	1.87×10^{-7}	$p < 0.001$***

4. Discussion

This study explored how different loss functions influence model performance and learning characteristics in brain tissue segmentation using a large-scale MRI dataset. By comparing pure Dice loss with the combined Dice + Cross-Entropy (Dice+CE) loss, we confirmed that loss function design is a key factor balancing segmentation accuracy and boundary recognition. Overall, Dice+CE achieved higher accuracy and IoU, showing clear improvements in tissues with ambiguous boundaries (e.g., skin, skull, cerebrospinal fluid (CSF), and gray matter (GM)). In contrast, for large and continuous structures like white matter (WM), Dice loss performed slightly better, suggesting a close link between loss function characteristics and tissue morphological complexity.

Dice+CE outperformed Dice alone in both IoU and Accuracy, mainly because the two losses complement each other [16]. Pointed out that Dice loss suffers from unstable gradients due to small denominators in gradient computation [17]. Noted that Cross-Entropy (CE) loss can be dominated by the background class in medical images, making it less suitable for class imbalance [16]. Demonstrated that combining Dice and CE produced higher segmentation scores (DSC 0.672) than either Dice (0.638) or CE (0.601) alone. Our findings align with these results, confirming that Dice+CE achieves superior segmentation performance.

The Dice+CE loss showed marked improvement in tissues with fuzzy or low-contrast boundaries, such as skin, skull, CSF, and GM [18]. Reported that standard Dice loss, focusing mainly on overlap, can cause vanishing gradients near fine boundaries, leading to local inaccuracies [19, 20]. Found that CE provides stronger classification signals for difficult boundary voxels ignored by Dice, improving precision. This complementary effect likely allowed pixel-level recovery in thin, irregular structures like skin, enhancing both IoU and accuracy [21]. Argued that CE mitigates output imbalance defined by false positives/negatives, which may explain the reduced leakage errors at skull–skin interfaces. Moreover, the combination of Dice's global shape consistency and CE's local precision likely reduced noisy islands in CSF and GM, improving segmentation reliability. Thus, the study verified that Dice+CE is advantageous for fine-grained brain segmentation.

White matter (WM) performed slightly better with pure Dice loss [22]. Analyzed 20 loss functions across six public datasets and explained that Dice, as a region-based loss, directly maximizes overlap between predictions and ground truth, aligning well with the risk-minimization principle [23]. Because Dice was originally proposed for white-matter lesion segmentation, it effectively captures size and localization agreement rather than pixel accuracy [23, 24]. Described CE as measuring voxel-wise probability differences, adding a second optimization target. Therefore, while Dice+CE improves precision for small or sparse objects, pure Dice optimized solely for regional overlap retains a slight advantage for large, continuous regions like WM.

This study has several limitations. Using automatically generated label maps (SimNIBS charm and FastSurfer) may reduce boundary accuracy and anatomical alignment, requiring expert correction in some cases. No data augmentation was applied, limiting generalization across posture, head size, noise, and scanner variation. Moreover, the experiments focused on a single 3D U-Net architecture, making it unclear whether the observed effects are architecture-specific or general. Performance evaluation relied on quantitative metrics and did not include expert or clinical assessments.

In conclusion, using over 500 MRI samples, this study confirmed that the choice of loss function significantly affects model performance in brain tissue segmentation. The combined Dice+CE loss yielded higher accuracy and IoU, particularly improving segmentation of tissues with unclear boundaries, while pure Dice slightly outperformed in large, continuous structures. These results indicate that combining Dice's global shape optimization with CE's voxel-level precision offers a balanced approach, suggesting that adaptive loss function strategies can enhance the reliability of MRI segmentation models.

Based on this study, three main directions for future research are suggested to further enhance model generalizability, accuracy, and utility.

Firstly, while our study utilizes a large dataset (552 volumes) to achieve strong initial generalizability, real-world clinical deployment demands robustness across unseen domains (e.g., new scanners or acquisition protocols), and therefore, future work must focus on applying advanced data augmentation techniques to drastically improve Domain Generalization capability. This involves implementing sophisticated methods like intensity-based augmentation (simulating noise and contrast shifts) and realistic elastic deformation, which introduce simulated spatial and intensity variability into the training data. By exposing the model to a wider range of synthetic variations than those present in the original dataset, we can explicitly train the model to be robust across different hardware and acquisition settings, thereby maximizing its stability in diverse clinical environments.

Secondly, enhancing segmentation requires optimizing the training process beyond the current state. This includes exploring Boundary-Sensitive Loss Functions (such as Boundary Loss or Distance Map-based Loss) to explicitly address and minimize minor inaccuracies at thin structures and tissue interfaces. Concurrently, architectural refinement, such as integrating Attention Mechanisms or deeper residual connections into the 3D U-Net, is necessary to better capture both global context and fine-grained local details, thus boosting overall accuracy.

Finally, the anatomical segmentation scope must be expanded. Beyond the current five tissues, future models should include electrically relevant structures, such as segmenting the Cerebellum distinctly, and explicitly modeling ocular structures and air cavities. Most importantly for high-fidelity head modeling, the single skull class must be refined into a multi-layered structure by delineating the compact bone and the inner spongy bone (Diploe), providing the detailed anatomical prerequisites for the most advanced computational simulations.

Acknowledgments

This study was supported by a grant of the Korea Institute of Radiological and Medical Sciences (KIRAMS), funded by Ministry of Science and ICT (MSIT), Republic of Korea (No. 50539-2025). In addition, This research was supported by the Bio&Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00356694 & No. RS-2024-00410364), and by the "Regional Innovation System & Education (RISE)" through the Gwangju RISE Center, funded by the Ministry of Education (MOE) and the Gwangju Metropolitan Government, Republic of Korea (2025-RISE-05-011).

References

[1] M. Wang, K. Zheng, Y. Xin, X. Chen, Y. Liu, H. Luo, J. Tang, T. Yuan, H. Wen, P. Wei, and Q. Liu, arXiv preprint arXiv:2509.01192 (2025).

[2] J. Gomez-Tames and M. Fernández-Corazza, J. Clin. Med. **13**, 3084 (2024).

[3] H. Xue, Y. Yao, and Y. Teng, Quant. Imaging Med. Surg. **14**, 1122 (2024).

[4] I. Despotović, B. Goossens, W. Philips, Comput. Math. Methods Med. **1**, 450341 (2015).

[5] X. Liu, L. Qu, Z. Xie, J. Zhao, Y. H. Shi, and Z. J. Song, Diagnostics **13**, 239 (2023).

[6] Z. Huang, J. Ye, H. Wang, Z. Deng, Z. Yang, Y. Su, J. Liu, T. Li, Y. Gu, S. Zhang, Y. Qiao, L. Gu, and J. He, Sci. Rep. **15**, 29795 (2025).

[7] Y. Chen, L. Quan, C. Long, Y. Chen, L. Zu, and C. Huang, Technol. Health Care **32**, 183 (2024).

[8] Z. Xiong, Sci. Rep. **15**, 24179 (2025).

[9] A. M. Mendrik, K. L. Vincken, H. J. Kuijf, M. Breeuwer, W. H. Bouvy, J. de Bresser, A. Alansary, M. de Bruijne, A. Carass, A. El-Baz, A. Jog, R. Katyal, A. R. Khan, F. van der Lijn, Q. Mahmood, R. Mukherjee, A. van Opbroek, S. Paneri, S. Pereira, M. Persson, M. Rajchl, D. Sarikaya, O. Smedby, C. A. Silva, H. A. Vrooman, S. Vyas, C. Wang, L. Zhao, G. J. Biessels, and M. A. Viergever, Comput Intell Neurosci. **1**, 813696 (2015).

[10] J. D. Nielsen, K. H. Madsen, O. Puonti, H. R. Siebner, C. Bauer, C. G. Madsen, G. B. Saturnino, and A. Thielscher, Neuroimage. **174**, 587 (2018).

[11] IXI Dataset, available at https://brain-development.org/ixi-dataset/ (accessed Sep. 22, 2025).

[12] O. Puonti, K. Van Leemput, G. B. Saturnino, H. R. Siebner, K. H. Madsen, and A. Thielscher, NeuroImage. **219**, 117044 (2020).

[13] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, NeuroImage. **219**, 117012 (2020).

[14] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, in Med. Image Comput. Comput.-Assist. Interv. (MICCAI 2016), Lect. Notes Comput. Sci. **9901**, 424 (2016).

[15] S. E. Stolte, A. Indahlastari, J. Chen, A. Albizu, A. Dunn, S. Pedersen, K. B. See, A. J. Woods, and R. Fang, Imaging Neurosci. **2**, imag-2-00090 (2024).

[16] M. Yeung, L. Rundo, Y. Nan, E. Sala, C. B. Schönlieb, and G. Yang, J. Digit. Imaging. **36**, 739 (2023).

[17] R. Yousef, S. Khan, G. Gupta, T. Siddiqui, B. M. Albahlal, S. A. Alajlan, and M. A. Haq, Diagnostics **13**, 1624 (2023).

[18] R. E. Jurdi, C. Petitjean, P. Honeine, V. Cheplygina, and F. Abdallah, Proc. Mach. Learn. Res. **143**, 158 (2021).

[19] S. M. Hosseini, Proc. Brit. Mach. Vis. Conf. **35**, 1 (2024).

[20] C. Acebes, A. H. Moustafa, O. Camara, and A. Galdran, in Med. Image Comput. Comput.-Assist. Interv. (MICCAI 2024), Lect. Notes Comput. Sci. **15008**, 710 (2024).

[21] S. A. Taghanaki, Y. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, Comput. Med. Imaging Graph. **75**, 24 (2019).

[22] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A.L. Martel, Med. Image Anal. **71**, 102035 (2021).

[23] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, in Med. Image Comput. Comput.-Assist. Interv. (MICCAI 2019), Lect. Notes Comput. Sci. **11765**, 92 (2019).

[24] R. Azad, M. Heidary, K. Yilmaz, M. Hüttemann, S. Karimijafarbigloo, Y. Wu, A. Schmeink, and D. Merhof, arXiv preprint arXiv:2312.05391 (2023).