Multimodal HIE Lesion Segmentation in Neonates: A Study of Loss Functions

Abdul Haseeb and Annayah Usman
Institute of Business Administration, Karachi, Pakistan
{a.haseeb.17132,a.usman_18673}@khi.iba.edu.pk

January 8, 2025

Abstract

Segmentation of Hypoxic-Ischemic Encephalopathy (HIE) lesion in neonatal MRI is crucial but challenging task due to diffused multifocal lesions and limited datasets. Using the BONBID-HIE dataset, we implemented a 3D U-Net with optimized preprocessing, augmentation, and training strategies to overcome data constraints. We compared various loss functions, including Dice, Dice-Focal, Tversky, and Hausdorff Distance Loss, alongside two novel hybrid losses combining Dice-Focal and Tversky with Hausdorff Distance Loss. Results highlight Tversky-HausdorffDT Loss as the best performer, achieving superior Dice and NSD metrics. This study emphasizes the importance of loss function optimization for accurate segmentation of HIE lesions of varying volume.

1 Introduction

Hypoxic-Ischemic Encephalopathy (HIE) refers to a brain injury caused by reduced oxygen or blood flow during prenatal, intrapartum or postnatal period [9]. It occurs in 1.5 to 2.5 per 1000 live births in developed countries and has severe health implications: by the age of 2 years, up to 60% of the patients either die or suffer from neurocognitive deficits like mental retardation, epilepsy, and cerebral palsy [1]. It not only has a high prevalence, affecting around 200,000 newborns worldwide every year, but also a significant economic cost, $2 billion per year in the US alone [3].

Accurate segmentation of HIE lesions in neonatal MRIs is crucial for correct prognosis of the disease. However, that in itself is very challenging since these lesions are typically multifocal, diffused, and sometimes occupy a very small proportion of brain volume. This makes it remarkably more challenging than other segmentation tasks like brain tumors [3].

Reviewing recent literature reveal work around ensemble approaches, heavy data augmentations, and advanced architectures such as Swin-UNetR, nnU-Net, SageResNet, and fusion models such as Swin-UNetR with random forest classifier [2]. While much of the literature has been focused on modifying model architectures and data enhancements, we add to the existing literature on loss function optimization by conducting a comparative analysis of different loss functions and also explore option of hybrid loss functions, with the intention of opening avenues for customized loss function for this domain.

2 Data and Methodology

The dataset used here is Part I of BONBID-HIE MICCAI 2023 Challenge [4]. It contains 3d skull stripped Apparent Diffusion Coefficient (ADC) maps, Z-score normalized ADC maps (ZADC), and binary label masks for 133 HIE patients (85 training cases, 4 validation cases, and 44 testing cases - hidden). ADC maps measure water diffusion in brain tissues, with restricted movement suggesting lesions. ZADC maps are a normalized version of ADC maps showing deviation from normal range. Normal values were constructed from ADC maps of 13 healthy individuals acquired $0 - 14$ days after birth [3].

The dataset here is particularly challenging because while it is the first MRI dataset specifically created for HIE segmentation tasks, it is extremely small to utilize advance architectures properly, it has lesions volumes that vary greatly from $< 1\%$ to $50 - 100\%$, and its maps are susceptible to scanner variances.

Note: We've used Pytorch [10] library mainly for the implementation of this study.

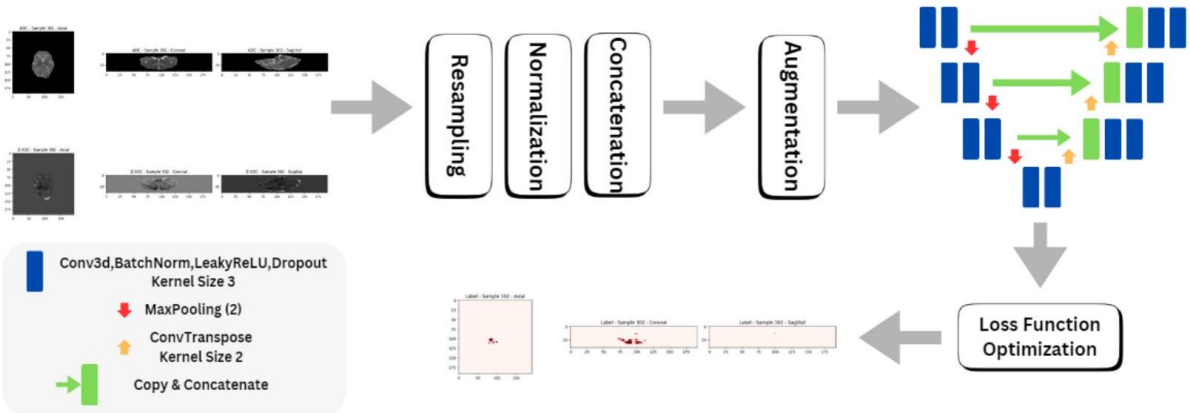

Figure 1: Methodology

2.1 Data Pre-Processing

3D medical images are memory intensive and so to reduce computational overhead during training, the data is preprocessed offline as doing it on-the-fly can lead to duplicate memory allocation. The following three preprocessing steps are undertaken:

- ADC and ZADC maps and label masks are resampled to a fixed size to make Pytorch batching possible and help model learn efficiently without being biased by size variations. Maps and labels are resampled to $(192, 192, 32)$ which is maximum of average dimension across all directories, rounded to a multiple of $2\hat{}4 = 16$ (total down-sampling factor of traditional UNet [11]) to ensure clean down-sampling and up-sampling (without fractional sizes that require inconsistent padding). For maps resizing is done using Trilinear interpolation to provide smooth transition in intensity values and for mask Nearest-Neighbor interpolation is used because it preserves discrete class boundaries. We didn't resample to maximum dimension across all directories purely due to memory constraints as otherwise we believe it to be the superior choice.
- The resampled (resized) maps' intensities are normalized individually using mean and standard deviation to make the model more robust by mitigating differences in scanners and patient anatomy.
- Lastly, both ADC and ZADC maps are concatenated (stacked on top of each other) to make a 2-channel input $(2, 192, 192, 32)$ for UNet to allow the model to leverage information from both modalities.

2.2 Model Specification

We've specified a 3D UNet architecture, due to its strong inductive bias, instead of employing more recent architectures like ViTs, which generally require larger datasets to realize their advantages. Our design includes three encoder and decoder blocks instead of the conventional four, reducing depth and parameter count to prevent overfitting given the limited dataset size. Each encoder block consists of double 3D convolutional (conv) layers followed by batch normalization, LeakyReLU activation, and dropout, concluding with a maxpooling layer for down-sampling. The bottleneck further enhances feature representations through additional double conv layers. The decoder progressively up-samples using transpose convolutions, integrates encoder features via skip connections, and then applies double conv layers. The output layer employs a $(1, 1, 1)$ convolution to generate the segmentation map.

We've used complete 2-channel input, combining ADC and ZADC maps, rather than single channel patches to leverage global context and multimodal information. Batch normalization ensures faster and more stable learning, while dropout minimizes overfitting, with progressively higher rates in deeper layers to encourage independent feature learning across activation maps at higher levels. LeakyReLU maintains a small gradient for negative inputs, ensuring gradient flow through non-lesion areas and enabling the model to learn from both lesion and background regions, which is crucial for sparse lesion segmentation.

2.3 Data Augmentation and Training

During training, the maps are augmented using the TorchIO library [13] with a probability of 0.5 for each transformation: (i) Random Noise (mean = 0.0, std=0.01) simulates realistic scanner artifacts, (ii) Random Anisotropy (downsampling=(1.2, 2.0)) reflects mild-to-moderate real-world variations in image resolution, (iii) Random Blur ((std=(0, 0.5)) addresses scenarios where lesion edges appear blurred due to low resolution or motion artifacts, (iv) Random Gamma (log _gamma $= (-0.1, 0.1)$) introduces subtle realistic changes to brightness and contrast, and (v) Random Elastic Transformation mimics natural variations in soft tissues caused by patient movement or anatomy. Augmentations are crucial in our case to artificially increase the training dataset size, simulate real world imaging variations for improved generalizability, and ensure scanner invariance in the model.

The network is trained on Nvidia Tesla P100 GPU on Kaggle with a batch size of 4, the maximum feasible size given memory constraints. Training is capped at 100 epochs, with early stopping enabled to halt training after 10 epochs with no improvement. To improve generalization and prevent overfitting due to the limited dataset size, the AdamW optimizer is used. Unlike the original Adam where weight decay is implicitly tied to the learning rate, AdamW explicitly applies weight decay by subtracting weight penalty during parameter update, which encourages smaller weights and smoother decision boundaries. A weight decay rate of $1e-3$ complements the small batch size by countering the risk of overfitting due to noisy gradient updates. The learning rate is set to $1e-3$, with exponential decay (factor of 0.9) for faster convergence and reduced oscillations in later epochs. L1 regularization (1e-4) promotes sparsity in the model, reducing overfitting. Gradient clipping (max norm = 1) is applied to prevent gradient explosion, which is crucial for deep architectures like U-Net. These strategies-dropout, weight decay, L1 regularization, and gradient clipping-help stabilize training, reduce overfitting, and enhance generalizability.

2.4 Loss Functions

In this study, we have compared the segmentation performance over various loss functions defined below:

- Dice Loss [6]: It's used to measure the similarity between two sets.

$$\text{DiceLoss } = 1 - \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

Where, A and B are the predicted and ground truth binary masks, respectively.

- Dice Focal Loss (DFL) [7]: Dice Focal Loss combines both the Dice Loss and Focal Loss, focusing on hard-to-classify examples.

$$DFL = (1-\alpha)(1- \text{ DiceLoss }) + \alpha(1- \text{ FocalLoss }); \text{ FocalLoss } = -\lambda_t \left(1-p_t\right)^{\gamma} \log\left(p_t\right) \tag{2}$$

Where α is a balancing factor between Dice loss and Focal loss, p_t is predicted probability mask of the true class, λ_t is a weighting factor for the class, and γ is the focusing parameter that adjusts the rate at which easy examples are down-weighted. By default, γ is kept as 2, equal balance is given to both Dice and Focal loss, and no weight is applied.

- Tversky Loss [12]: Tversky Loss is a generalization of Dice loss and is useful in dealing with class imbalances, particularly when false positives and false negatives need to be weighted differently.

$$\text{TverskyLoss } = 1 - \frac{|A \cap B|}{|A \cap B| + \alpha| \sim B| + \beta| \sim A|} = 1 - \frac{TP}{TP + \alpha FP + \beta FN} \tag{3}$$

Where A and B are same as Dice loss and α and β are hyperparameters that control the weighting of false positives and false negatives. We've kept α as 0.3 and β as 0.7 ($\alpha < \beta$) to emphasize of false negatives more as in medical diagnosis missing a lesion could have severe consequences than wrongly identifying one.

- Hausdorff Distance Loss (HDTL) [5]: Hausdorff Distance is a measure of the maximum distance between two sets. In segmentation, it's used to measure how far apart the predicted and true boundaries are. Hausdorff Distance Loss is the reciprocal of this value to minimize the distance:

$$HDTL = \frac{1}{1 + \text{ HausdorffDistance}} \tag{4}$$

We also defined two hybrid loss functions. One is a linear combination of Dice-Focal and Hausdorff Distance Loss while the other is a linear combination of Tversky and Hausdorff Distance Loss. Through first combination the aim is to strengthen Dice-Focal Loss's ability to segment lesions of varying sizes by incorporating Hausdorff Distance Loss for improved boundary precision. While, for the second combination the aim is to merge Tversky Loss's strength in capturing large lesions with Hausdorff Distance Loss's ability to capture precise boundaries. They're defined as below:

$$DF - HDTLoss = \alpha(DFL) + \beta(\log(HDTL)) \tag{5}$$

$$Tversky - HDTLoss = \alpha(TverskyLoss) + \beta(\log(HDTL)) \tag{6}$$

Taking the log of the Hausdorff distance loss in both combinations reduces the impact of outlier distances, stabilizing training and emphasizing overall boundary alignment rather than rare extreme errors. We've kept α as 0.9 and β as 0.1 after empirically testing combinations from $\{(\alpha, \beta) : \alpha \in [0.5, 0.9], \beta \in [0.1, 0.5]\}$. All of the above are implemented through MONAI library [8].

3 Results and Discussion

3.1 Evaluation Metrics

We evaluate the segmentation quality using the following three metrics defined below:

- Mean Surface Distance (MSD):

$$\text{MSD}(p, q) = \frac{1}{2}\left(\frac{d(\delta(q), \delta(p))}{|\delta(q)|} + \frac{d(\delta(p), \delta(q))}{|\delta(p)|}\right) \tag{7}$$

Where $\delta(x)$ represents the surface of x and $d(x, y)$ represents the shortest surface distance from surface x to surface y. It computes the average distance between the surfaces of two binary masks, measuring how well the predicted and ground truth surfaces align. It considers both directions: from the prediction surface to the ground truth and vice versa.

- Normalized Surface Dice (NSD):

$$NSD(p, q)_\tau = \frac{|\delta(q) \cap \gamma_\tau(p)| + |\gamma_\tau(q) \cap \delta(p)|}{|\delta(q) + \delta(p)|} \tag{8}$$

NSD is a metric used to evaluate the similarity between the boundary surfaces of predicted and ground truth binary masks, allowing for a specified tolerance distance τ. It considers how many boundary points from the two masks fall within this distance, rather than requiring exact overlap. NSD provides a more flexible, surface-focused measure of accuracy, especially useful in scenarios where small boundary mismatches are acceptable.

- Dice coefficient:

$$\text{Dice}(p, q) = \frac{2 \times |p \cap q|}{|p| + |q|} \tag{9}$$

It measures the volumetric overlap between the prediction and ground truth masks.

3.2 Loss Function Comparison

Loss Functions	Dice ↑	MSD ↓	NSD ↑	Epochs
Dice Loss (Baseline)	0.3800	15.0650	0.3850	32
Dice Focal Loss	0.4900	1.7925	0.5275	49
Tversky Loss	0.3525	15.3650	0.3375	38
HausdorffDT Loss	0.3300	Inf	0.2800	29
DiceFocal-HausdorffDT Loss	0.4925	**1.4225**	0.5300	72
Tversky-HausdorffDT Loss	**0.5000**	1.6250	**0.5325**	59

Table 1: Metric-Based Comparison of Loss Functions

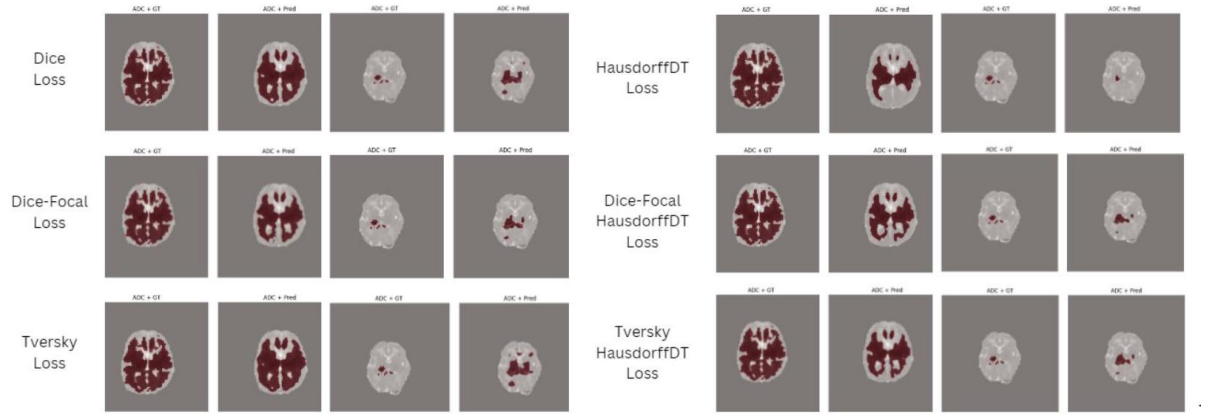

Figure 2: Visualizing Segmentation Masks Across Loss Functions

The results indicate that combining losses yield superior performance than standalone loss functions. Dice-Focal loss reports great improvement over the baseline by a huge margin and understandably so as instead of just dice loss focusing on the overlap between predicted and ground truth regions that help in addressing lesions of varying sizes, focal loss specifically prioritizes hard-to-segment regions, such as smaller or less prominent lesions, over large, well-defined ones. This results in overall better performance.

Hausdorff Distance Loss aims to provide more precise boundaries, but in doing so, it fails to capture complete lesion areas and sometimes misses small lesions altogether. In contrast, Tversky Loss performs well with large lesions and does not miss lesion areas but tends to over-segment, leading to poor performance with very small lesion sizes. Our defined hybrid losses, which aim to leverage the advantages of each individual loss function while mitigating their disadvantages, offer the best performance. Tversky-HausdorffDT Loss has the highest Dice and NSD metrics with second lowest MSD whereas DiceFocal-HausdorffDT Loss has the lowest MSD with second highest Dice and NSD metrics. Overall, Tversky-HausdorffDT Loss performs the best. These results are supported by both evaluation metrics and visualized masks on axial ADC maps. An interesting observation is that all combination losses require more epochs to converge compared to the standalone losses.

4 Conclusion

In this study, we aimed to improve HIE lesion segmentation through loss function optimization and our proposed hybrid loss functions achieved the best Dice, MSD, and NSD metrics. However, despite achieving improvements with hybrid losses as hypothesized, this study has several limitations.

4.1 Limitations

- None of the defined losses did well on extremely small lesions ($< 1\%$ volume) due to high heterogeneity in the dataset and limited capability of the model itself.

- Only the validation set itself is used for evaluation since the test set is held out by the organizers as the dataset is part of BONBID-HIE Challenge 2024 and evaluation on test set require submitting a Docker container to the online platform by the organizers. The performance on the validation set may not reflect the true capability of the methodology defined here.
- Resampled label mask could introduce distortion in segmentation regions in comparison to the actual mask. A more robust approach would be to only resample maps to a fixed size and reverse resample the predicted binary masks to original size, spacing, direction, and origin for true comparison with the actual label mask.

4.2 Future Work

Loss function optimization improves the segmentation performance and while this will direct research towards defining a custom loss function for HIE lesions, the improvement is expected to be marginal only. It would be better to define a custom loss function in relation to more advance architectures.

The main challenge in this problem lies in the data itself, which is highly limited for training any model with strong capabilities. A better approach would be to avoid training altogether and instead use models like SAM, which offer zero-shot generalizability. Specifically, MedSAM-2, which is already fine-tuned on medical imaging, would be ideal. From [3] we know that negative intensities in ZADC maps are correlated with lesion regions. Therefore, we can prompt MedSAM-2 with bounding boxes around negative intensities in ZADC maps and provide ADC maps as input for segmentation.

References

[1] K. A. Allen and D. H. Brandon. "Hypoxic Ischemic encephalopathy: Pathophysiology and experimental treatments". In: *Newborn and Infant Nursing Reviews* 11.3 (2011), pp. 125–133. DOI: 10.1053/j.nainr.2011.07.004.

[2] R. Bao. *AI for Brain Lesion Detection and Trauma Video Action Recognition: First BONBID-HIE Lesion Segmentation Challenge and First Trauma Thompson Challenge, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 16 and 12, 2023, Proceedings*. Springer Nature, 2024.

[3] R. Bao et al. "BOston Neonatal Brain Injury Dataset for Hypoxic Ischemic Encephalopathy (BONBID-HIE): part I. MRI and Manual Lesion annotation". In: *bioRxiv (Cold Spring Harbor Laboratory)* (2023). URL: https://doi.org/10.1101/2023.06.30.546841.

[4] Hypoxic Ischemic Encephalopathy Lesion Segmentation Challenge - Grand Challenge. *Hypoxic Ischemic Encephalopathy Lesion Segmentation Challenge - Grand Challenge*. n.d. URL: https://bonbid-hie2023.grand-challenge.org/bonbid-hie2023/.

[5] D. Karimi and et al. "Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks". In: *IEEE Transactions on Medical Imaging* 39.2 (2019), pp. 499–513.

[6] F. Milletari and et al. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *3DV*. 2016.

[7] MONAI. *Loss functions - MONAI 1.4.0 Documentation*. n.d. URL: https://docs.monai.io/en/stable/losses.html#dicefocalloss.

[8] MONAI. *MONAI - home*. n.d. URL: https://monai.io/.

[9] National Institute of Neurological Disorders and Stroke. *Hypoxic ischemic encephalopathy*. n.d. URL: https://www.ninds.nih.gov/health-information/disorders/hypoxic-ischemic-encephalopathy.

[10] PyTorch. *PyTorch*. n.d. URL: https://pytorch.org/.

[11] O. Ronneberger, P. Fischer, and T. Brox. "U-NET: Convolutional Networks for Biomedical Image Segmentation". In: *arXiv (Cornell University)* (2015). URL: https://doi.org/10.48550/arxiv.1505.04597.

[12] A. Sadegh and et al. *Tversky loss function for image segmentation using 3D fully convolutional deep networks*. 2017. URL: https://arxiv.org/abs/1706.05721.

[13] TorchIO. *TorchIO*. n.d. URL: https://torchio.readthedocs.io/.