SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 6, 2026

LOSS FUNCTION ANALYSIS FOR CLASS-IMBALANCED MULTI-ORGAN SEGMENTATION OF THE GASTROINTESTINAL TRACT IN MRI

Moavia Hassan[*1], Muhammad Javed Iqbal[2], Muhammad Ilyas[3], Muhammad Ahsan Rafique[4], Esha Husnain[5]

[*1,2,3,4,5]*Department of Computer Science, University of Engineering and Technology, Taxila, Punjab, Pakistan*

[*1]moaviahassan112@gmail.com

DOI:https://doi.org/10.5281/zenodo.20729853

Keywords
medical image segmentation, gastrointestinal tract, loss function, class imbalance, transformer, magnetic resonance imaging, deep learning

Article History
Received: 20 April 2026
Accepted: 31 May 2026
Published: 17 June 2026

Corresponding Author: *
Moavia Hassan

Abstract
MRI guided radiotherapy for the abdominal cancers should must be marked on every scan slice to stomach, small bowel and large bowel so the radiation can avoid healthy tissues. It is often marked by hand, and different experts often outline the same organ differently. While the Deep learning can perform this task automatically, but the data makes it hard to accurate marking. Such as UW-Madison gastrointestinal (GI) tract dataset almost contains 57% no organ and remaining covers only a portion of image when organ appears that leaves the classes heavily imbalanced. The training loss is the main mechanism that drives a network to attend to such rare foreground, yet it is usually chosen by convention rather than by evidence. We compare five losses under identical conditions on a fixed 2.5D network that pairs a SegFormer MiT-B2 encoder with a U-Net decoder: Dice, soft binary cross-entropy (SoftBCE), their combination, Tversky, and a Focal-Dice combination. Training and evaluation use a patient-grouped split and per-image-averaged Dice, intersection over union (IoU), sensitivity, specificity, and precision. All five reach comparable overall Dice within 0.007 (0.9006 to 0.9072), so overall accuracy is largely insensitive to the loss here. The error profile differs sharply, however: Tversky gives the highest sensitivity (0.9465) at the lowest precision (0.9091), SoftBCE the highest precision (0.9363) at the lowest sensitivity (0.9255), and Focal-Dice the best balanced Dice (0.9072). The small bowel stays hardest under every loss. The loss should therefore be chosen for the clinically preferred balance between missing tissue and over-contouring, not for overall accuracy.

1. INTRODUCTION

Radiotherapy is central to treating abdominal cancers, and its accuracy depends on knowing where the organs at risk sit relative to the tumour. MRI-guided linear accelerators now image the patient immediately before each treatment fraction, which supports daily adaptive planning that follows the large day-to-day motion of the gastrointestinal organs. To use this, the stomach, small bowel, and large bowel must be outlined on every slice so the dose can be shaped around them. Doing this by hand is slow and varies between observers, and it bottlenecks time-limited online workflows, which is why automatic segmentation matters.

Encoder-decoder networks based on U-Net [1] are the standard tool for biomedical segmentation, and self-configuring pipelines built on them remain strong baselines [2]. Transformer encoders such as SegFormer [3], volumetric transformer

models such as UNETR and Swin UNETR [4, 5], and hybrid transformer-convolutional designs [6, 7] have since improved the modelling of long-range context, which helps for organs with extended, variable shapes. The UW Madison GI tract dataset [8] is the public benchmark for this task, and recent encoder-decoder work on it reports Dice near 0.90 [9].

The dataset's defining trait is severe class imbalance. Roughly 57% of slices carry no labelled organ, and where an organ appears it covers a small fraction of pixels. This pushes pixel-wise objectives toward the background, so a network can reach high pixel accuracy while under-segmenting the organs themselves [10]. The structures to which a model attends are governed largely by the loss function. Many losses target this problem: the overlap-based Dice loss [11, 12], the Tversky loss with asymmetric penalties on false negatives and positives [13], the Focal loss that emphasises hard pixels [14], the Focal Tversky [15] and boundary losses [16], and compound or unified objectives that mix region and pixel terms [17-22].

Even so, the loss in GI segmentation studies is usually inherited or fixed by convention rather than settled by a matched comparison. Loss surveys note that the objective strongly affects imbalanced problems and that none is universally best [17, 18], but task-specific evidence for GI MRI is thin. This work fills that gap with a controlled experiment: the architecture, data, optimisation, and every other factor are held fixed, and only the loss changes, so any difference is attributable to the loss. The fixed model is a 2.5D hybrid that stacks neighbouring slices and pairs a SegFormer MiT-B2 encoder with a U-Net decoder, used as a standard backbone rather than as a contribution. We make three contributions: a like-for-like comparison of five widely used losses on UW Madison at accuracy on par with published work; the finding that overall Dice is largely loss-insensitive here while the sensitivity-precision balance is strongly loss-dependent; and a practical rule in which the loss is chosen by clinical priority rather than by overall accuracy.

2. Materials and Methods

2.1. Method Overview

Figure 1 summarizes the pipeline. Each target slice is combined with its two neighbours into a three-channel 2.5D input, passed through a SegFormer MiT-B2 encoder and a U-Net decoder, and turned into one mask per organ. During training the prediction is compared with the ground truth by one of five losses; at validation the same model is scored with five metrics. Only the loss differs between runs, while the dataset, preprocessing, architecture, and optimisation described below stay identical throughout.

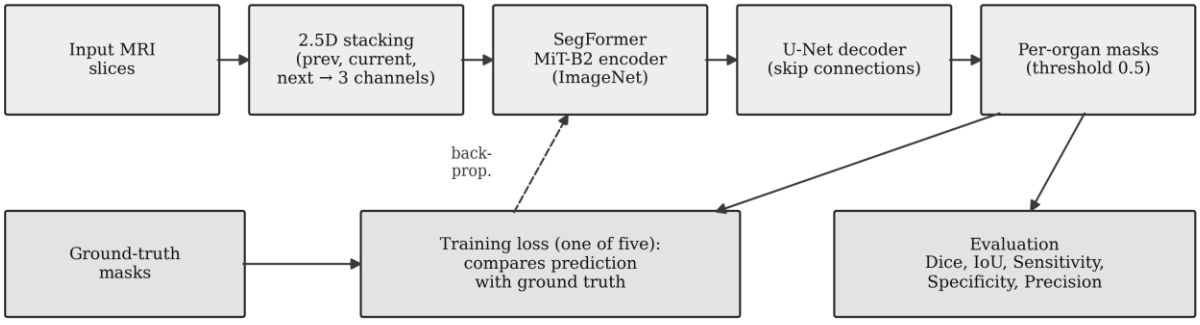

Figure 1. Overview of the proposed segmentation pipeline

2.2. Dataset

UW Madison GI tract publicly available dataset was used in this study [8], which holds 38,496 axial MRI slices from 85 patients imaged over several days. Each slice is stored as a 16-bit grayscale image, and the large bowel, small bowel, and stomach are marked with run-length-encoded masks. These masks are spread very unevenly across the data (Figure 2). The large bowel turns up in 36.6% of the slices and the small bowel in 29.1%, while the stomach appears in only 22.4%. In 56.9% of slices there is no organ at all. The small bowel is the hardest of the three, being thin, convoluted, and variable across patients and days.

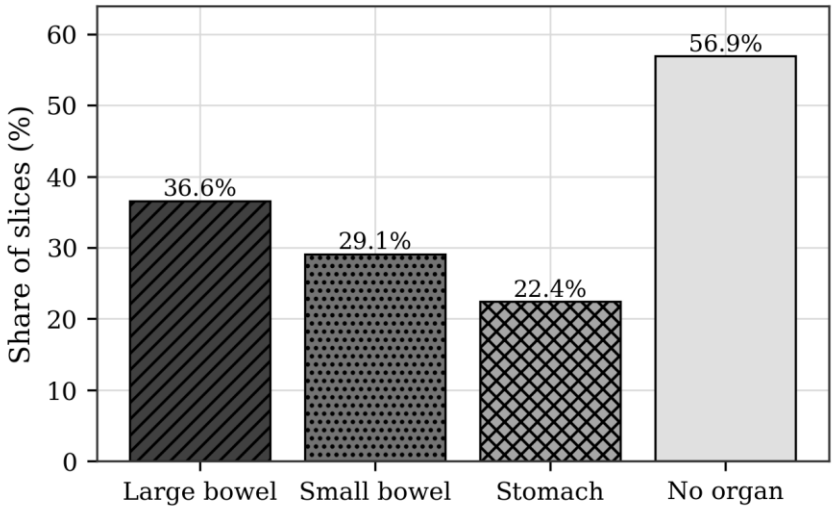

Figure 2. Proportion of slices in the UW Madison GI tract dataset

2.3. 2.5D Input Representation

To add context along the body axis without the cost of full 3D volumes, the model uses a 2.5D input. For each target slice, the preceding and following slices from the same scan are stacked with it to form three channels; at a scan boundary the target slice replaces the missing neighbour. Each slice is min-max normalized to [0, 1] independently, and inputs are resized to 224 by 224 pixels.

2.4. Network Architecture

The network is a hybrid transformer-convolutional model: a SegFormer MiT-B2 encoder [3] with a U-Net decoder [1], built with the segmentation-models-pytorch library [23]. The encoder applies efficient self-attention [24] to yield multi-scale hierarchical features, unlike the single-scale Vision Transformer [25], and the decoder up-samples and fuses them through skip connections to a full-resolution prediction per organ. The encoder is pretrained on ImageNet [26, 27]. We treat this as a fixed, standard backbone so the loss is the only variable. The output gives one logit map per organ as independent binary segmentations, which fits a dataset where organs may co-occur and a slice may hold any subset of them.

2.5. Loss Functions

Five loss configurations were compared as shown in Figure 3. For the loss we have used different metrices which is analyzed below. Let p be the predicted probability of a pixel and g its binary label, with sums taken over all pixels and organ channels and s a small smoothing constant. Dice loss measures overlap and is insensitive to foreground size, which makes it robust to imbalance:

$$L_{Dice} = 1 - \frac{2 \sum p\, g + s}{\sum p + \sum g + s} \quad (1)$$

Equation 2 shows the SofBCE mathematical formulation that obtains pixel wise object and it's smooth to optimize but biased toward the majority background under imbalance:

$$L_{SoftBCE} = -\frac{1}{N}\sum[g \log p + (1-g)\log(1-p)] \quad (2)$$

Equally weighted sum of L = 0.5 L_Dice + 0.5 L_SoftBCE is a common baseline. The Tversky loss [13] generalizes Dice by weighting false negatives and false positives separately:

$$L_{Tversky} = 1 - \frac{\sum pg}{\sum pg + \alpha \sum(1-g)p + \beta \sum g(1-p)} \quad (3)$$

With $\alpha = 0.3$ and $\beta = 0.7$, so missed foreground is penalised more than over-segmentation and recall is favoured. Finally, the Focal-Dice combination pairs Dice with the Focal loss [14], which scales each pixel by $(1 - p_t)$ raised to $\gamma = 2$ to focus on hard boundary pixels; the two terms are equally weighted.

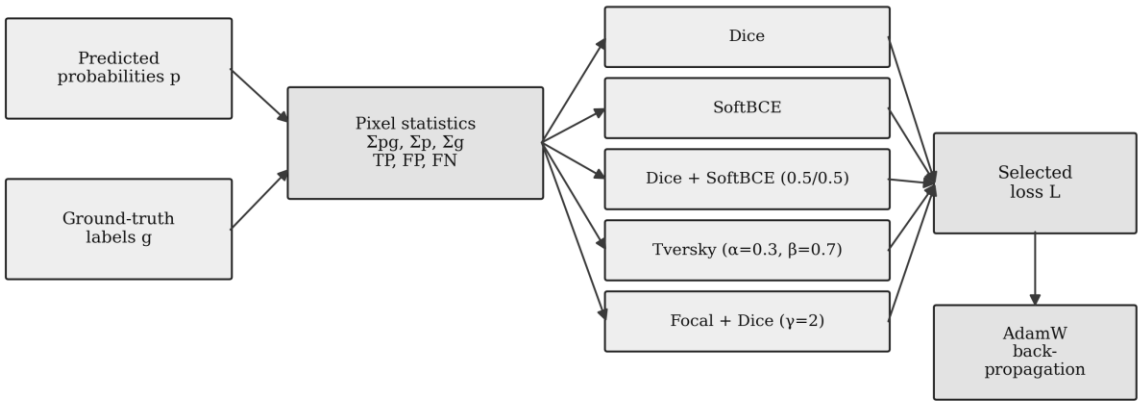

Figure 3. Computation of the five training losses from the predicted probabilities and ground-truth labels

2.6. Experimental Setup

All five models trained under identical conditions, the loss being the only change. A grouped five-fold split kept each patient wholly within one set; a single fixed fold was used, with 30,832 slices from 68 patients for training and 7,664 from 17 patients for validation. The same augmentation was applied in training through the Albumentations library [28] horizontal and vertical flips, shift-scale-rotation, elastic and grid distortion, and brightness, contrast, and gamma jitter. Optimisation used AdamW [29] at a learning rate of 2×10^{-4} and weight decay 1×10^{-2}, with cosine annealing [17] to 1×10^{-6}, automatic mixed precision, a batch size of 16, and 18 epochs. A fixed seed of 42 was used throughout, and all runs used PyTorch [30] on a single NVIDIA T4 GPU.

2.7. Evaluation Metrics

Performance was measured with the Dice coefficient, IoU, sensitivity, specificity, and precision, computed per organ and averaged for the overall values, with predictions thresholded at 0.5. In terms of true and false positives and negatives (TP, FP, TN, FN), the metrics are defined as follows:

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (4)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (5)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Metrics follow the per-image-averaging convention used by prior work on this dataset, under which a slice that is empty and correctly predicted empty scores 1 through a unit smoothing term; this keeps the values comparable to the literature. For each loss, the epoch with the best validation Dice was kept for reporting.

3. Results

3.1. Overall Comparison of Loss Functions

Table 1 lists overall validation performance, ordered by Dice. The losses are remarkably close: Dice spans just 0.007, from 0.9006 for Tversky to 0.9072 for Focal-Dice, with IoU equally tight. Changing the loss therefore barely moves overall accuracy, and all five sit at the level of recent encoder-decoder work on the same benchmark [9]. Figure 4 shows the validation curves converging stably to a similar level, and Figure 5 plots the overall Dice.

Table 1. Overall validation performance of the five loss functions, ordered by Dice.

Loss	Dice	IoU	Sensitivity	Specificity	Precision
Focal + Dice	0.9072	0.8781	0.9316	0.9989	0.9330
Dice + SoftBCE	0.9057	0.8765	0.9279	0.9990	0.9340
Dice	0.9052	0.8759	0.9304	0.9989	0.9307
SoftBCE	0.9049	0.8754	0.9255	0.9991	0.9363
Tversky	0.9006	0.8696	0.9465	0.9984	0.9091

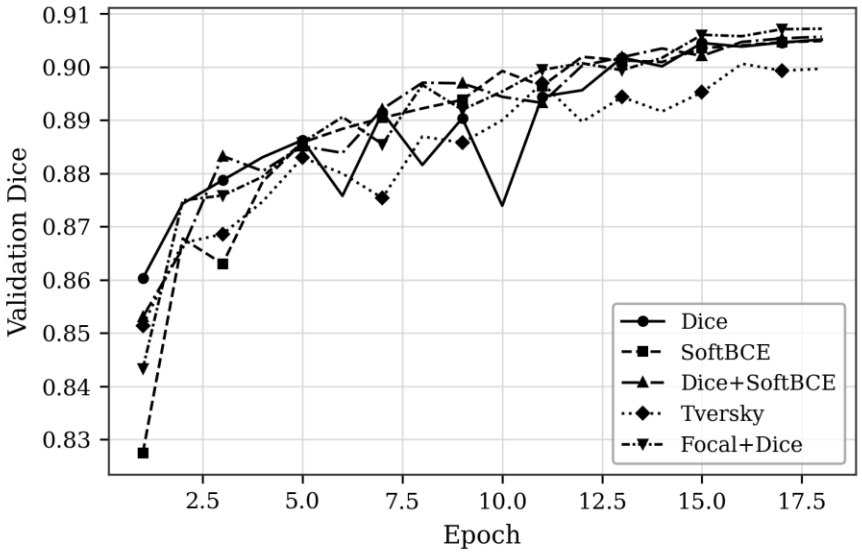

Figure 4. Validation Dice of training epochs for the five loss functions

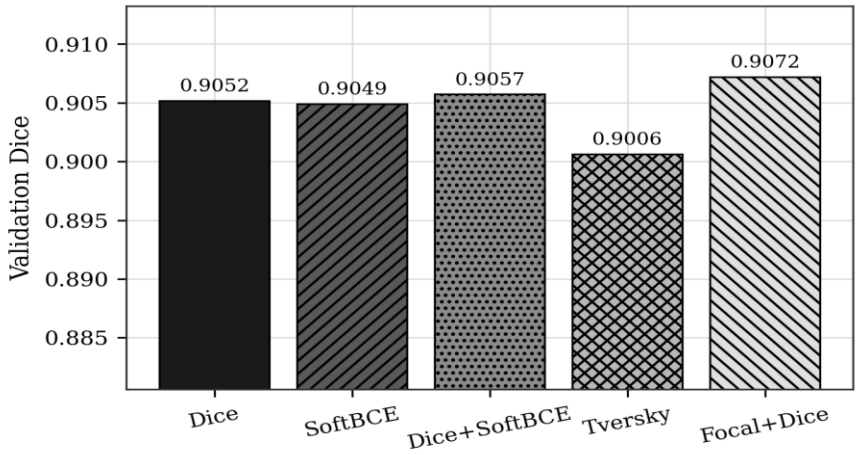

Figure 5. Comparison of Validation Dice Scores for Various Loss Functions

3.2. Per-Organ Performance

Dice score of each organ is presented in Figure 6 and Table 2. This ranking stands up with the expectations of clinician. Stomach is easy to segment organ due to its large size, and it has score between 0.9385 to 0.9442, ad the large bowel falls in the middle. The hardest organ is small bowel, no loss pushed it past 0.8795 which shows its thin and looping shape that varies a lot from patient to patient. The loss shifts each organ by only a few thousandths of a Dice point, which again points to loss-insensitive overall accuracy. Figure 7 shows these small differences as a heatmap.

Table 2. Per-organ validation Dice for each loss function

Loss	Large bowel	Small bowel	Stomach
Focal + Dice	0.8991	0.8783	0.9442
Dice + SoftBCE	0.8971	0.8795	0.9406
Dice	0.8954	0.8788	0.9413

Loss	Large bowel	Small bowel	Stomach
SoftBCE	0.8976	0.8781	0.9390
Tversky	0.8905	0.8728	0.9385

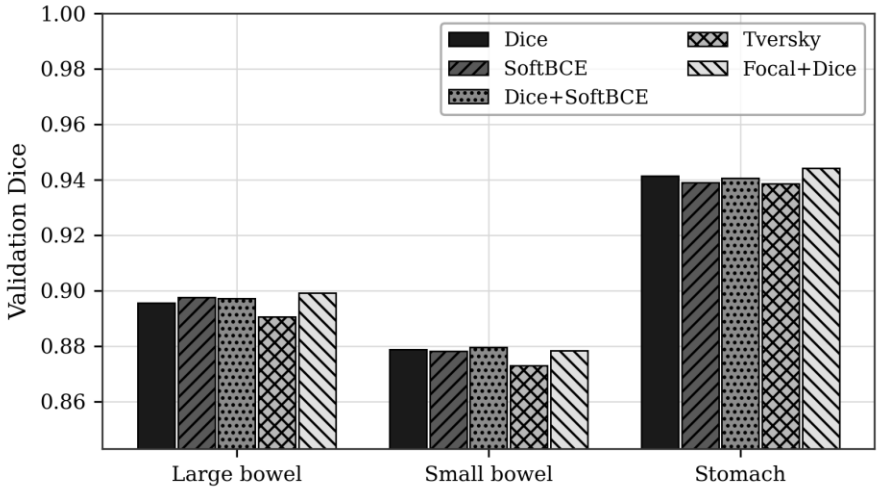

Figure 6. Per-organ validation Dice grouped by organ for the five loss functions

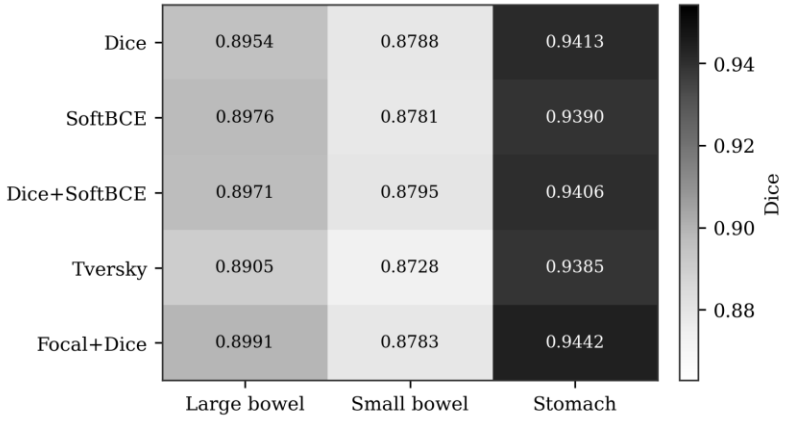

Figure 7. Per-organ Dice for all losses shown as a heatmap

Table 3 gives the full metric set for the best loss, Focal-Dice. It confirms the stomach as both the most accurate and most reliably detected organ, while the small bowel trails on every metric.

Table 3. Full per-organ and overall validation metrics for the best loss

Class	Dice	IoU	Sensitivity	Specificity	Precision
Large bowel	0.8991	0.8645	0.9255	0.9988	0.9254
Small bowel	0.8783	0.8429	0.9085	0.9984	0.9129
Stomach	0.9442	0.9269	0.9607	0.9996	0.9607
Overall	0.9072	0.8781	0.9316	0.9989	0.9330

3.3. Sensitivity-Precision Trade-off

Overall Dice is nearly constant, but the kind of error each loss makes is not, and this is the study's main result (sensitivity and precision columns of Table 1; Figure 8). Tversky, tuned to punish false negatives, reaches the highest sensitivity at 0.9465,

missing the least tissue, but the lowest precision at 0.9091 because it over-segments more. SoftBCE does the reverse, reaching the best precision at 0.9363 but the weakest sensitivity at 0.9255, since it leans toward predicting background. Two extremes were notice between Focal-Dice and plain Dice. There gap tells us the real picture of how sensitivity varies across the losses and precision but the overall Dice moves by only fraction number. So the loss mainly decides which type of error model makes rather how much error there is.

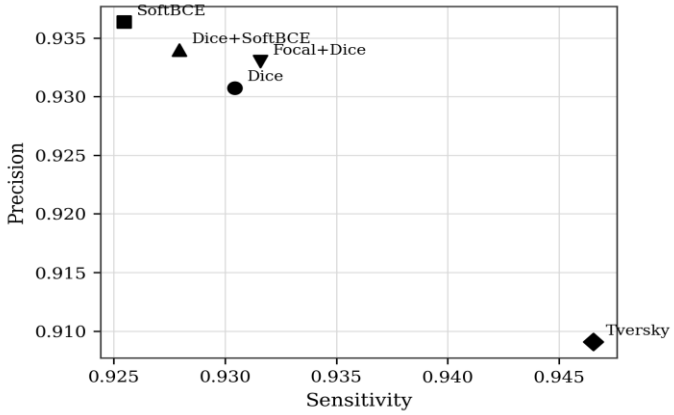

Figure 8. Sensitivity versus precision for the five loss functions

3.4. Qualitative Results

The validation dices with ground truth outlines and predictions from the best model are presented in figure 9. Predicted edges follow the organ walls closely for the stomach and large bowel, although most of the error that remains are on the small bowel. The model sometime merges loops that sit next to each other or misses a thin stretch in small organ which is already proved above in table 2 and Figure 6.

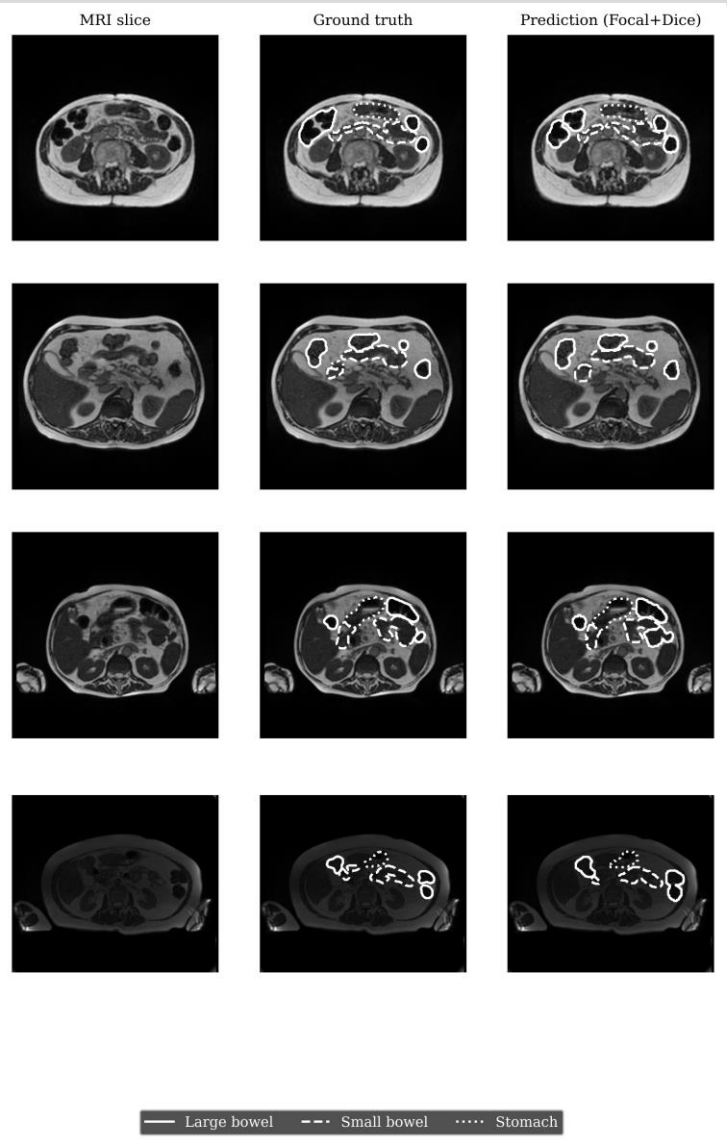

Figure 9. Qualitative results for the best loss Focal & Dice

4. Discussion

The study concluded that on the UW-Madison dataset with the fixed architecture, swapping the loss barely changes overall accuracy. The loss changes type of mistake the model tends to make. The loss no longer drives the accuracy when Dice levels off around 0.90. Instead, it interprets whether the model make a mistake on missing tissue or of drawing in too much. Each loss type acts differently as Tversky shifts it toward high sensitivity while the SoftBCE toward high precision, and the Dice-based combinations stay in the middle and happen to score best overall.

The two types of mistakes do not cost the same in MRI-guided radiotherapy. Missing organ tissue can leave it inside the treatment and expose it to dose while contouring tissue that in not present is usually a irritation corrected at review. Tversky is defensible despite the lower precision when the priority is to avoid treating healthy bowel. And on the other hand, a precision-leaning loss is the better solution where spurious contours are the bigger problem. The contribution is to make that an evidence-based choice rather than a habit, by quantifying the trade-off on a real imbalanced dataset.

The best loss here, Focal-Dice at 0.9072 Dice, matches recent encoder-decoder results on the same data, which reported about 0.908 [9]. The agreement shows the fixed backbone is a fair testbed and that the comparison ran at a competitive level rather than on a weak model. The small bowel's consistent difficulty across losses also agrees with prior reports and with its known anatomical variability.

A few limitations are worth noting. We tested on a single validation fold rather than the full set. The patient-grouped split keeps every patient out of both the training and validation sets, so there is no leakage. The same sensitivity-precision pattern also turns up in all three organs, which makes it unlikely to be a quirk of one split. Even so, running every fold would tighten the small gaps in overall Dice. The five losses are all existing ones, so the paper is a guide for choosing a loss, not a proposal for a new one. The small overall-Dice differences are themselves part of the finding. Because of that, our conclusions rest on sensitivity and precision, and we do not claim that any single loss is better overall. Future work follows directly, by repeating the comparison across all folds to measure variance, extending it to other backbones to test generality, and designing a loss whose false-negative weight can be set to a stated clinical operating point so the balance found here can be chosen deliberately.

5. Conclusions

We compared five losses for class-imbalanced multi-organ GI segmentation on MRI, holding a 2.5D SegFormer-UNet fixed so the loss was the only variable. All five reached comparable overall Dice within 0.007, Focal-Dice best at 0.9072, so overall accuracy is largely loss-insensitive here. The losses differed sharply in the sensitivity-precision balance: Tversky gave the highest sensitivity and SoftBCE the highest precision, a separation several times the overall-Dice spread. The small bowel was hardest under every loss. The practical conclusion is to choose the loss by the clinically preferred balance between missing tissue and over-contouring rather than for a higher overall score, with a recall-leaning loss such as Tversky a sensible default where missing organ tissue is costly.

References

O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015// 2015: Springer International Publishing, pp. 234-241.

F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203-211, 2021/02/01 2021, doi: 10.1038/s41592-020-01008-z.

E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077-12090, 2021.

A. Hatamizadeh *et al.*, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574-584.

A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI brainlesion workshop*, 2021: Springer, pp. 272-284.

J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*, 2022: Springer, pp. 205-218.

U. M. C. C. Center. "UW Madison GI Tract Image Segmentation. Kaggle Competition Dataset." https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation.

N. Sharma, S. Gupta, D. H. Elkamchouchi, and S. Bharany, "Encoder–decoder variant analysis for semantic segmentation of gastrointestinal tract using UW-Madison dataset," *Bioengineering*, vol. 12, no. 3, p. 309, 2025.

M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural networks*, vol. 106, pp. 249-259, 2018.

F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, 2016: Ieee, pp. 565-571.

C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *International Workshop on Deep Learning in Medical Image Analysis*, 2017: Springer, pp. 240-248.

S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *International workshop on machine learning in medical imaging*, 2017: Springer, pp. 379-387.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980-2988.

N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, 2019: IEEE, pp. 683-687.

H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*, 2019: PMLR, pp. 285-296.

J. Ma *et al.*, "Loss odyssey in medical image segmentation," *Medical image analysis*, vol. 71, p. 102035, 2021.

S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, 2020: IEEE, pp. 1-7.

S. A. Taghanaki *et al.*, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24-33, 2019.

S. Asgari Taghanaki *et al.*, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *arXiv e-prints*, p. arXiv: 1805.02798, 2018.

K. C. Wong, M. Moradi, H. Tang, and T. Syeda-Mahmood, "3D segmentation with exponential logarithmic loss for highly unbalanced object sizes," in *International conference on medical image computing and computer-assisted intervention*, 2018: Springer, pp. 612-619.

M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102026, 2022.

Iakubovskii. "Segmentation Models PyTorch." https://github.com/qubvel/segmentation_models.pytorch.

A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009: Ieee, pp. 248-255.

J. Deng, "A large-scale hierarchical image database," *Proc. of IEEE Computer Vision and Pattern Recognition, 2009*, 2009.

A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.

I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.