Biomedical Physics &
Engineering Express

Endorsed by IPEM Institute of Physics and Engineering in Medicine
PURPOSE-LED PUBLISHING™

ACCEPTED MANUSCRIPT

Comparative Analysis of Loss Functions for Uncertainty Quantification in Medical Image Segmentation

To cite this article before publication: Sivalal Kethavath *et al* 2025 *Biomed. Phys. Eng. Express* in press https://doi.org/10.1088/2057-1976/ae2b73

Comparative Analysis of Loss Functions for Uncertainty Quantification in Medical Image Segmentation

Sivalal Kethavath[1], **Vasundhara**[1,*], **and Phaneendra K. Yalavarthy**[2]

[1]Department of Electronics and Communication Engineering,
National Institute of Technology Warangal,
Hanumakonda, Warangal, Telangana, 506004, India

[2]Department of Computational and Data Sciences,
Indian Institute of Science,
C.V. Raman Avenue, Bangalore, 560012, India

∗Corresponding author: vasundhara@nitw.ac.in

Email: sk21ecrep03@student.nitw.ac.in, phaneendra.k.yalavarthy@gmail.com

Abstract

Accurate and reliable medical image segmentation is crucial for diagnosis and treatment planning, yet current deep learning methods often lack calibrated uncertainty estimates, limiting their clinical trustworthiness. To address this challenge, we systematically evaluate the effect of seven loss functions—Dice, Log-Cosh, Binary Cross-Entropy, Tversky, Focal Tversky, Log Dice, and a Hybrid loss—on both segmentation performance and uncertainty quantification. Using U-Net architectures trained with Monte Carlo Dropout, we assess pancreas and spleen segmentation on two benchmark CT datasets from the Medical Segmentation Decathlon. Our results show that Dice loss consistently achieves the highest segmentation accuracy (Dice score up to 0.87 for spleen, 0.68 for pancreas, 0.92 for liver) while maintaining well-calibrated uncertainty estimates across diverse training regimes. In contrast, alternative loss functions such as Focal Tversky and Hybrid losses demonstrated potential in specific scenarios but lacked robustness. These findings highlight the importance of loss-function selection for balancing segmentation accuracy and uncertainty, with Dice loss emerging as a particularly reliable choice. This work provides novel insights into how loss functions influence uncertainty–error alignment, offering guidance for the development of more trustworthy segmentation models in clinical workflows.

Keywords: Segmentation, Uncertainty Quantification, Deep Learning, Loss Function, Monte Carlo Dropout

1. Introduction

Accurate medical image segmentation is an important image-based task that enables precise diagnosis, treatment planning, and disease monitoring. In particular, the pancreas and spleen

present significant challenges for segmentation owing to their complex anatomical shapes, variability among patients, and proximity to other organs. These complexities make reliable segmentation essential, yet difficult to achieve, particularly when using automated methods.

Deep learning models[1], notably those based on the U-Net architecture[2], have revolutionized medical image segmentation by leveraging convolutional neural networks to capture intricate spatial hierarchies in imaging data. In clinical applications, understanding the confidence level of model predictions is vital for ensuring patient safety and building trust in automated systems. Uncertainty quantification [3–7] allows clinicians to assess the reliability of segmentation results, guide decision-making processes, and identify areas that may require further expert review.

The aim of this study is to provide insights into selecting appropriate loss functions that can simultaneously optimize the segmentation performance and produce reliable uncertainty estimates. This knowledge is crucial for the development of more trustworthy and clinically applicable deep learning models for medical image segmentation. By enhancing both the accuracy and interpretability of automated segmentation tools, this study contributes to bridging the gap between advanced computational methods and their practical deployment in healthcare settings, ultimately aiming to improve patient care through reliable diagnostic support.

This study investigates the impact of different loss functions on both segmentation performance and uncertainty estimation in deep-learning-based pancreas and spleen segmentation. We systematically evaluate seven loss functions—including Dice loss, log-cosh loss, binary cross-entropy, Tversky loss, focal Tversky loss, log Dice loss, and a hybrid loss—using a U-Net architecture and Monte Carlo dropout for uncertainty estimation [8]. Experiments were conducted on two distinct datasets for pancreas and spleen segmentation, with variations in training epochs and learning rates, to assess the robustness of each loss function. The insights gained regarding the relationship between the loss functions, segmentation accuracy, and uncertainty estimation can inform the development of deep learning models for other medical imaging tasks.

This work makes the following contributions:

- Comprehensive Comparative Study: We provide a systematic comparison of seven loss functions to evaluate their effect on both segmentation performance and uncertainty calibration.
- Cross-Organ Robustness Analysis: The experiments are conducted on pancreas and spleen datasets under varying training regimes, allowing us to assess the generalizability of each loss function.
- Key Empirical Findings: Our results indicate that Dice loss consistently delivers high segmentation accuracy while maintaining low predictive uncertainty, making it a strong candidate for clinical use.

- Insights for Clinical Interpretability: Beyond numerical results, we analyze the relationship between uncertainty estimates and segmentation reliability, emphasizing their role in supporting safer, AI-assisted clinical decision-making.

2. Related Work

2.1. Loss Functions for Medical Image Segmentation

Loss function design plays a central role in medical image segmentation, particularly in handling class imbalance and boundary ambiguity. Traditional pixel-wise losses such as Binary Cross-Entropy (BCE) are widely used but often perform poorly in imbalanced settings, where the foreground organ occupies a small fraction of the image. Region-overlap losses such as Dice and Jaccard have been shown to improve robustness in these cases by directly optimizing similarity between predicted and ground-truth masks [8, 9]. Extensions such as the Tversky and Focal Tversky losses introduce tunable parameters to control false positives and negatives, offering flexibility in highly imbalanced datasets. Hybrid approaches that combine pixel-wise, region-based, and structural similarity metrics have also been proposed to better capture fine anatomical boundaries [8, 9]. Despite these advances, most prior studies focus on segmentation accuracy alone, without systematically investigating how different loss functions affect model uncertainty.

2.2. Uncertainty Quantification in Segmentation

In parallel, uncertainty estimation has become increasingly important for clinical adoption of deep learning models. Monte Carlo Dropout [4, 10] is a widely used Bayesian approximation technique that captures model (epistemic) uncertainty by performing multiple stochastic forward passes during inference. Extensions include probabilistic U-Net models [11], ensemble methods, and confidence calibration approaches [12–14]. These methods provide pixel-wise uncertainty maps that highlight regions where predictions are unreliable, enabling clinicians to identify areas that require manual review. Previous work has applied uncertainty estimation to tumor segmentation, radiotherapy planning, and organ delineation [15–17], but most studies evaluate uncertainty in isolation rather than jointly with segmentation accuracy under different loss functions.

2.3. Research Gap

While prior work has independently explored segmentation losses [18] and uncertainty quantification techniques [19, 20], there is limited systematic evaluation of how the choice of loss function influences both segmentation accuracy and the quality of uncertainty estimates. This gap is particularly relevant in medical imaging, where reliable uncertainty maps are essential for clinical interpretability. Our study addresses this gap by performing a comparative evaluation of seven widely used and emerging loss functions, analyzing their impact

not only on segmentation performance but also on uncertainty calibration across two organ datasets.

3. Methods

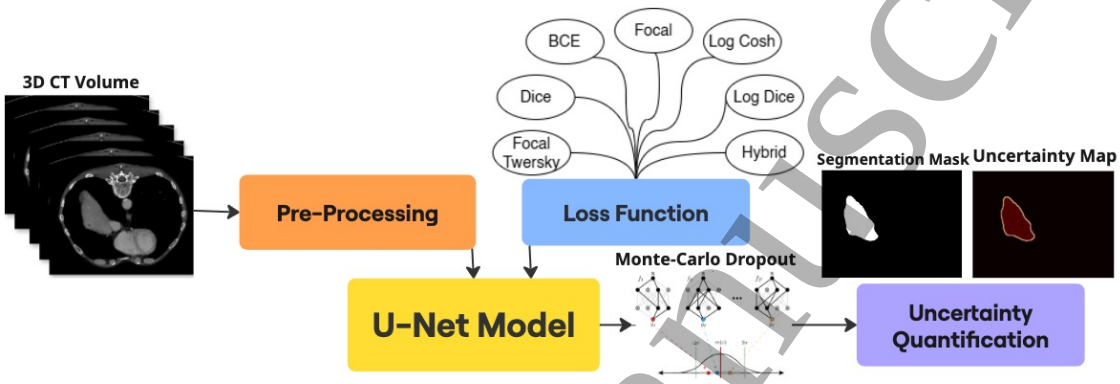


Figure 1: Workflow for Uncertainty Quantification in Medical Image Segmentation. The workflow starts with data collection, followed by preprocessing and training the U-Net model using various loss functions. After training, Monte Carlo Dropout is applied for uncertainty estimation, and the results are visualized.

As illustrated in Figure 1, our overall methodology comprises several key stages. The process begins with data collection and preprocessing, continues with training the U-Net based segmentation model using a variety of loss functions, and concludes with uncertainty estimation using Monte Carlo Dropout. This comprehensive workflow ensures that both segmentation accuracy and predictive uncertainty are rigorously quantified.

3.1. Datasets

The datasets used in this study were obtained from the Medical Segmentation Decathlon (MSD) , a collection curated to promote the development of generalizable algorithms across diverse medical imaging tasks. All data were provided in the Neuroimaging Informatics Technology Initiative (NIfTI) format, which is widely adopted for storing medical imaging data owing to its ability to handle multidimensional datasets and associated metadata.

3.1.1. Pancreas Dataset

The pancreatic dataset focuses on the segmentation of the pancreas and pancreatic tumors using abdominal computed tomography (CT) scans. It comprises 420 three-dimensional (3D) volumes (282 training, 138 testing) acquired in the portal venous phase and sourced from the Memorial Sloan Kettering Cancer Center. To enable efficient model training, each 3D

volume was decomposed into axial 2D slices, resulting in a total of 26,719 annotated samples for training and testing. A significant challenge presented by this dataset is pronounced label imbalance, characterized by large background regions, medium-sized pancreatic structures, and small tumor areas. This imbalance poses difficulties for segmentation algorithms, particularly for accurately identifying and delineating small tumor regions in the predominant background.

3.1.2. Spleen Dataset

The spleen dataset consists of CT scans aimed at spleen segmentation, comprising 61 3D volumes (41 training, 20 testing), also sourced from the Memorial Sloan Kettering Cancer Center. Similar to the pancreas dataset, the 3D volumes were decomposed into 2D slices, yielding 3,650 annotated samples in total. This dataset presents the challenge of a wide range of foreground sizes owing to significant variability in spleen size among patients. Such variability requires the segmentation model to be robust to anatomical and scale variations, ensuring accurate segmentation across diverse patient anatomies.

3.1.3. Atlas23 Dataset

The Atlas23 (often referred to as the ATLAS liver dataset) dataset was additionally used to evaluate the generalization capability of the proposed approach. This dataset is used for benchmarking of automatic liver segmentation algorithms, particularly in the context of hepatocellular carcinoma. This is a contrast-enhanced magnetic resonance imaging (CE-MRI) dataset with 60 volumetric scans, comprising a total of 4,744 axial slices. Among these, 48 volumes (3,795 slices) were utilized for training and 12 volumes (949 slices) for testing. This data was acquired primarily on Siemens 1.5T and 3T MRI machines, with some on GE equipment. The dataset provides high-quality anatomical annotations that enable the assessment of segmentation performance across diverse imaging modalities, complementing the CT-based pancreas and spleen datasets used in this study.

3.2. Data Preprocessing

Preprocessing is a critical step in medical image segmentation with the aim of enhancing image quality, reducing noise, and standardizing data for efficient model training. Appropriate preprocessing improves model convergence and contributes to more accurate and reliable segmentation outcomes, which are essential for clinical applicability.

3.2.1. Conversion from 3D Volumes to 2D Slices

To reduce the computational complexity and memory requirements, 3D CT volumes were decomposed into two-dimensional (2D) axial slices. This approach simplifies the segmentation task by enabling the use of 2D convolutional neural networks (CNNs), which are less

resource-intensive than their 3D counterparts. Each 2D slice was paired with its corresponding segmentation mask to ensure the spatial alignment and consistency. This slicing strategy facilitates more manageable data handling and accelerates the training process without significantly compromising the spatial context required for an accurate segmentation.

3.2.2. Intensity Windowing

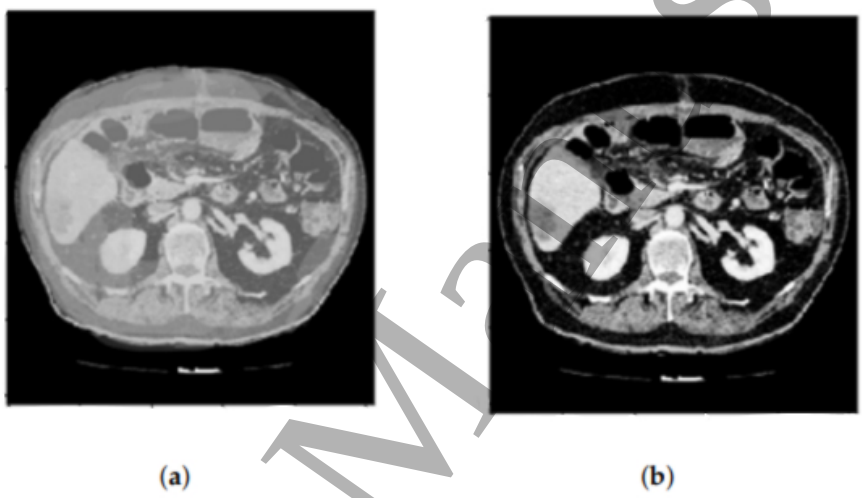


Figure 2: Comparison between (a) normal and (b) organ-specific windowed visualization of a CT image. The window width (WW) range of 0–300 HU optimizes contrast between organ tissue and surrounding structures, and the window level (WL) of 150 HU enhances organ detail visibility.

As demonstrated in Figure 2, intensity windowing—specifically Hounsfield Unit (HU) windowing—was applied to focus on the relevant intensity ranges that best represent the target organs. CT images encompass a broad spectrum of HU values, and restricting this range enhances the contrast between the organs of interest and the surrounding tissues. While the HU value ranges for specific organs can vary depending on the CT scanner and patient-specific factors, windowing was utilized without specifying exact ranges to generalize the method and avoid scanner-specific dependence. This step is crucial for differentiating the pancreas and spleen from adjacent anatomical structures and mitigating scanner-dependent intensity variations, thereby improving the model's ability to learn meaningful features.

3.2.3. Median Filtering

To further improve image quality, a median filter was applied to each 2D slice. Median filtering is effective in reducing noise while preserving important edge information, which is vital for accurate segmentation. By replacing each pixel value with the median of its neighboring pixels within a specified kernel size (typically 3×3), this nonlinear technique diminishes the impact of outliers and random noise, resulting in smoother images with well-preserved anatomical boundaries. The smoothing effect aids the model by focusing on significant structural features, rather than being distracted by high-frequency noise.

3.3. U-Net Architecture

The U-Net architecture [2] was employed for the segmentation tasks due to its proven effectiveness in biomedical image segmentation. The network follows a symmetric encoder–decoder design with skip connections that fuse low-level and high-level feature maps, enabling both precise localization and global context understanding.

The encoder (contracting path) consists of four downsampling stages, each comprising two successive 3×3 convolutional layers with He-normal initialization, batch normalization, and rectified linear unit (ReLU) activation. Each stage is followed by a 2×2 max-pooling layer for downsampling and a dropout layer (rate = 0.15–0.3) to reduce overfitting. The number of filters doubles at each stage, progressing from 16 to 32, 64, 128, and finally 256 at the bottleneck.

The decoder (expanding path) mirrors the encoder with four upsampling stages. Each stage begins with a transposed convolution (3×3, stride = 2) followed by concatenation with the corresponding encoder feature maps via skip connections. Two 3×3 convolutional layers with batch normalization and ReLU activation are then applied, along with dropout after concatenation to preserve robustness.

At the output, a 1×1 convolution with sigmoid activation produces a single-channel probability map for binary segmentation. This design ensures that both global anatomical context and fine structural details are preserved.

To balance efficiency and performance, we empirically selected an initial filter count of 16 and limited the network depth to five levels. Preliminary tests with smaller (8 filters) and larger (32 filters) initial filter counts showed that 16 provided the best trade-off, achieving comparable segmentation accuracy while avoiding unnecessary computational overhead. This configuration reduced the memory footprint and training cost without sacrificing segmentation quality, aligning with sustainable computing practices.

3.4. Loss Functions

Seven different loss functions were explored to evaluate their impact on segmentation performance and uncertainty estimation. Each loss function addresses specific challenges in medical image segmentation, such as class imbalance, boundary precision, and optimization stability

[8, 9]. The loss functions employed include Dice Loss, Log-Cosh Loss, Binary Cross-Entropy Loss, Tversky Loss, Focal Tversky Loss, Log Dice Loss, and a Hybrid Loss combining Jaccard Loss, Binary Cross-Entropy Loss, and Structural Similarity Index Measure (SSIM) Loss.

- p_i is the predicted probability for pixel i.
- g_i is the ground truth label for pixel i.
- ϵ is a small constant to prevent division by zero.

3.4.1. Dice Loss

Dice Loss [8, 9] is derived from the Dice Similarity Coefficient (DSC) and is particularly effective for handling class imbalance by focusing on the overlap between the predicted segmentation P and the ground truth G:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \tag{1}$$

where p_i is the predicted probability for pixel i, g_i is the ground-truth label, ϵ is a small constant to prevent division by zero, and the summation runs over all pixels. This loss function emphasizes the correct prediction of the minority class, which is critical in medical segmentation tasks with imbalanced data.

3.4.2. Log-Cosh Loss

Log-Cosh Loss [8, 9] combines the advantages of Mean Squared Error (MSE) and Mean Absolute Error (MAE), providing a smooth and robust alternative:

$$\mathcal{L}_{\text{Log-Cosh}} = \sum_i \log\left(\cosh(p_i - g_i)\right) \tag{2}$$

This loss function is less sensitive to outliers and can improve convergence during training. By smoothly penalizing larger errors, the model focuses on reducing overall prediction discrepancies without being disproportionately influenced by extreme values.

3.4.3. Binary Cross-Entropy Loss

Binary Cross-Entropy (BCE) [8, 9] Loss measures the pixel-wise discrepancy between the predicted probabilities and ground truth labels:

$$\mathcal{L}_{\text{BCE}} = -\sum_i \left[g_i \log(p_i) + (1 - g_i)\log(1 - p_i)\right] \tag{3}$$

BCE is commonly used in binary classification tasks, and is effective for segmentation when combined with other loss functions. It provides a straightforward measure of the difference between predicted probabilities and actual labels, encouraging the model to produce confident and accurate predictions.

3.4.4. Tversky Loss

Tversky Loss [8, 9] generalizes Dice Loss by introducing adjustable parameters to balance false positives and false negatives:

$$\mathcal{L}_{\text{Tversky}} = 1 - \frac{\sum_i p_i g_i + \epsilon}{\sum_i p_i g_i + \alpha \sum_i p_i (1 - g_i) + \beta \sum_i (1 - p_i) g_i + \epsilon} \tag{4}$$

where α and β are weighting factors that control the penalties for false positives and false negatives, respectively. By adjusting these parameters, the loss function can be tailored to prioritize the reduction of specific types of errors, which is beneficial in medical contexts where the costs of false negatives and false positives may differ significantly.

3.4.5. Focal Tversky Loss

Focal Tversky Loss [8, 9] extends Tversky Loss by incorporating a focusing parameter γ, enhancing model training on hard examples:

$$\mathcal{L}_{\text{Focal Tversky}} = \left(1 - \frac{\sum_i p_i g_i + \epsilon}{\sum_i p_i g_i + \alpha \sum_i p_i (1 - g_i) + \beta \sum_i (1 - p_i) g_i + \epsilon}\right)^{\gamma} \tag{5}$$

This loss function is particularly useful for highly imbalanced datasets, as it focuses on the learning process of difficult-to-classify pixels and improves the model's ability to correctly segment small or complex structures.

3.4.6. Log Dice Loss

Log Dice Loss [8, 9] applies a logarithmic transformation to the Dice Coefficient, emphasizing smaller overlap regions and providing stronger gradients when the overlap is minimal:

$$\mathcal{L}_{\text{Log Dice}} = -\log\left(\frac{2 \sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon}\right) \tag{6}$$

This transformation makes the loss function more sensitive when the Dice Coefficient is low, thereby encouraging the model to improve in areas where it performs poorly.

3.4.7. Hybrid Loss

Hybrid Loss [8, 9] combines multiple loss functions to exploit their strengths. In this study, the Jaccard Loss, Binary Cross-Entropy Loss, and Structural Similarity Index Measure

(SSIM) loss are defined as follows:

$$\mathcal{L}_{\text{Hybrid}} = \mathcal{L}_{\text{Jaccard}} + \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{SSIM}} \tag{7}$$

Jaccard Loss: Jaccard Loss measures the similarity between predicted and ground truth regions by computing the ratio of their intersection over union. It penalizes mismatched regions and is particularly effective in handling class imbalance by focusing on overlap rather than pixel-wise differences. The formulation is given by:

$$\mathcal{L}_{\text{Jaccard}} = 1 - \frac{\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i - \sum_i p_i g_i} \tag{8}$$

Structural Similarity Index (SSIM) Loss: SSIM Loss focuses on capturing structural information and preserving image quality at the boundary, which is crucial for ensuring accurate contour detection in medical applications. The SSIM Loss function is formulated as follows:

$$L_{\text{SSIM}} = 1 - \text{SSIM}(P, G) \tag{9}$$

where $\text{SSIM}(P, G)$ is the structural similarity index computed for the predicted image P and the ground truth image G. The SSIM function is defined as

$$\text{SSIM}(P, G) = \frac{(2\mu_P\mu_G + C_1)(2\sigma_{PG} + C_2)}{(\mu_P^2 + \mu_G^2 + C_1)(\sigma_P^2 + \sigma_G^2 + C_2)} \tag{10}$$

where: μ_P and μ_G are the mean intensities of P and G, σ_P and σ_G are their standard deviations, σ_{PG} is the covariance between P and G, $C_1 = 0.012$ and $C_2 = 0.032$ are small constants that prevent division by zero;

The inclusion of SSIM Loss encourages the model to consider the structural integrity of images, preserve spatial details, and ensure perceptually accurate segmentation. This is particularly beneficial in medical imaging applications where fine anatomical details are critical for diagnosis and treatment planning.

3.5. Uncertainty Calculation Using Monte Carlo Dropout Method

Uncertainty quantification [4] is pivotal in medical image analysis to evaluate the reliability of model predictions, thereby enhancing clinical decision-making. This provides insights into areas where the model is less confident, indicating the need for cautious interpretation or further analysis.

In this study, uncertainty was estimated using the Monte Carlo Dropout method[4].

Monte Carlo Dropout (MC Dropout) is a practical Bayesian approximation method that enables uncertainty estimation without altering the training process. In conventional dropout, units are randomly deactivated during training to reduce overfitting, but disabled

Table 1: Summary of the loss functions evaluated in this study, their key properties, and intended advantages.

Loss Function	**Key Property**	**Intended Advantage**
Dice Loss	Based on Dice Similarity Coefficient	Handles class imbalance by maximizing overlap between predicted and ground truth regions.
Log-Cosh Loss	Smooth variant of MSE/MAE	Robust to outliers and stabilizes optimization during training.
Binary Cross-Entropy (BCE) Loss	Pixel-wise probabilistic loss	Simple and widely used; encourages confident predictions but may suffer from imbalance.
Tversky Loss	Weighted generalization of Dice	Allows tuning trade-off between false positives and false negatives; useful when one type of error is more critical.
Focal Tversky Loss	Adds focusing parameter to Tversky	Emphasizes hard-to-classify pixels; improves segmentation of small or complex structures.
Log Dice Loss	Logarithmic transformation of Dice	Provides stronger gradients in low-overlap regions; improves convergence in difficult cases.
Hybrid Loss	Combination of Jaccard, BCE, and SSIM	Balances region overlap, pixel-level accuracy, and structural integrity for fine anatomical details.

at inference. In MC Dropout, the dropout layers are instead kept active during inference, and the same input is passed through the network multiple times, each time with a different dropout mask applied. This produces a set of stochastic predictions that approximate a posterior distribution over the model's outputs .

In this study, each input slice was passed through the trained U-Net $T = 20$ times with dropout enabled, and the resulting probability maps were aggregated to compute:

Predictive mean: the average probability across all passes, used as the final segmentation prediction.

Predictive uncertainty: pixel-wise entropy and standard deviation across passes, quantifying epistemic uncertainty (model-related uncertainty due to limited data).

The entropy-based uncertainty quantifies the randomness of the predicted probability distribution for each pixel, as follows:

$$U_i = -\sum_c p_{i,c} \log p_{i,c} \tag{11}$$

where U_i is the uncertainty of pixel i, $p_{i,c}$ is the predicted probability of pixel i belonging to class c, and the summation runs over all the classes.

This method is particularly suitable for medical image segmentation because it provides not only a segmentation mask but also an uncertainty map that highlights regions where predictions are unreliable, often corresponding to blurred anatomical boundaries or ambiguous tumor regions.

3.6. Evaluation Metrics

Several metrics were employed to comprehensively assess the segmentation models and reliability of the uncertainty estimates. Each metric evaluates different aspects of the model's performance and predictive confidence.

3.6.1. Test Dice Coefficient

The Test Dice Coefficient measures the overlap between the predicted segmentation and the ground truth on the test set. It is defined as:

$$\text{Dice} = \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \tag{12}$$

The Dice Coefficient ranges from 0 (no overlap) to 1 (perfect overlap), providing a quantitative measure of segmentation accuracy. A higher Dice score indicated better performance, reflecting the ability of the model to accurately segment the target organ.

3.6.2. Uncertainty Mean and Standard Deviation

The Mean and Standard Deviation of the uncertainty quantify the average predictive uncertainty across all pixels and their variability. These metrics are derived from entropy-based uncertainty calculations using the Monte Carlo Dropout method.

Uncertainty for Each Pixel: The entropy-based uncertainty for pixel i was calculated as follows:

$$U_i = -\sum_c p_{i,c} \log p_{i,c} \tag{13}$$

where $p_{i,c}$ is the predicted probability of pixel i belonging to class c and the summation runs over all classes.

Uncertainty Mean:

$$\text{Uncertainty Mean} = \frac{1}{N} \sum_{i=1}^{N} U_i \tag{14}$$

where N denotes the total number of pixels. This metric provides an overall measure of the model's predictive uncertainty.

Uncertainty Standard Deviation:

$$\text{Uncertainty Std} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (U_i - \text{Uncertainty Mean})^2} \tag{15}$$

This metric assesses the variability of uncertainty across the image, indicating how consistently the model predicts the uncertainty.

3.7. Mean Prediction Mean and Standard Deviation

The mean Prediction Mean and Standard Deviation assess the average predicted probabilities and their variability, reflecting the confidence and consistency of the model in its predictions.

3.7.1. Mean Prediction for Each Pixel

The mean predicted probability for pixel i across multiple stochastic forward passes is defined as

$$\bar{p}_i = \frac{1}{T} \sum_{t=1}^{T} p_i^{(t)} \tag{16}$$

where $p_i^{(t)}$ is the predicted probability at the t-th forward pass and T is the total number of forward passes.

3.7.2. Mean Prediction Mean

The Mean Prediction for all pixels was calculated as follows:

$$\text{Mean Prediction} = \frac{1}{N}\sum_{i=1}^{N}\bar{p}_i \tag{17}$$

where N is the total number of pixels. This metric indicates the average confidence level of the model across all the pixels.

3.7.3. Mean Prediction Standard Deviation

The Mean Prediction Standard Deviation is calculated as follows:

$$\text{Mean Prediction Std} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{p}_i - \text{Mean Prediction})^2} \tag{18}$$

This metric measures the variability of the model's predicted probabilities, highlighting the regions in which the confidence of the model fluctuates.

3.8. Training Setup

The training configurations were meticulously designed to ensure robust model performance and facilitate a thorough analysis of the loss functions. All experiments were conducted using TensorFlow as the deep-learning framework, leveraging its advanced features for building and training neural networks. The models were trained on a workstation equipped with a NVIDIA RTX A5000 graphics card.

The Adam optimizer was employed owing to its adaptive learning rate capabilities and efficient handling of sparse gradients. An initial learning rate of 0.01 was set for both datasets.

A learning rate scheduler callback was implemented, reducing the learning rate to 10% of its previous value every four epochs. This strategy helped prevent vanishing gradients and promoted smooth convergence by enabling larger updates during the early stages of training and finer adjustments as training progressed. The models were trained for 20 epochs, with an early stopping mechanism monitoring the validation Dice Coefficient to prevent overfitting. Across all experiments, the model weights corresponding to the highest validation Dice Coefficient were saved for evaluation and deployment.

In addition to the primary training setup, an ablation study was conducted on the spleen dataset to examine the influence of learning rate variations and extended training duration on the performance and consistency of loss functions. Specifically, models were trained for 40 epochs with learning rates of 0.1 and 0.001. This exploration aimed to study the stability and behavior of different loss functions under varying training regimes, providing deeper insights into their impact on model performance.

Furthermore, an additional ablation study was performed where models were selected based on the lowest validation loss, rather than the highest validation Dice Coefficient. This study was designed to investigate whether there exists any inherent bias toward Dice Loss when optimizing solely for Dice-based metrics during model selection.

4. Results

Table 2: Segmentation performance on the Pancreas dataset using different loss functions. The table reports the Dice Score, Uncertainty Mean, and Uncertainty Standard Deviation across models trained with U-Net architecture.

Loss Function	Dice Score	Uncertainty Mean	Uncertainty Std
Log Cosh Loss	0.4747	0.008646	0.002048
Dice Loss	**0.6838**	**0.000446**	**0.000732**
Focal Tversky Loss	0.6699	0.000542	0.000867
Tversky Loss	0.6784	0.000555	0.000878
BCE Loss	0.6270	0.001356	0.001828
Log Dice Loss	0.6681	0.000424	0.000690
Hybrid Loss	0.5680	0.000704	0.000969

Table 3: Segmentation performance on the Spleen dataset using different loss functions. The table reports the Dice Score, Uncertainty Mean, and Uncertainty Standard Deviation across models trained with U-Net architecture.

Loss Function	Dice Score	Uncertainty Mean	Uncertainty Std
Log Cosh Loss	0.2987	0.3712132	0.002048
Dice Loss	**0.8733**	**0.0003376**	0.000732
Focal Tversky Loss	0.8583	0.0004006	0.000867
Tversky Loss	0.8603	0.0004455	0.000878
BCE Loss	0.7671	0.0041139	0.001828
Log Dice Loss	0.7853	0.0004672	**0.000690**
Hybrid Loss	0.8000	0.0004426	0.000969

Tables 2–7 present the quantitative performance metrics for the different loss functions evaluated in this study. Table 2 summarizes the segmentation performance on the pancreas dataset, whereas Tables 3–7 detail the results on the spleen and atlas23 datasets under varying training regimes and learning rates. As shown in Table 2, the Dice Loss consistently achieves a higher Dice Score (0.6838) compared to other loss functions, accompanied by notably lower uncertainty measures, indicating better segmentation consistency.

For the spleen dataset, Table 3 highlights that Dice Loss outperforms alternative loss functions in terms of both segmentation accuracy and uncertainty calibration, establishing

Table 4: Segmentation performance on the Atlas23 dataset using different loss functions. The table reports the Dice Score, Uncertainty Mean, and Uncertainty Standard Deviation across models trained with U-Net architecture.

Loss Function	**Dice Score**	**Uncertainty Mean**	**Uncertainty Std**
Log Cosh Loss	0.8658	0.039621	0.043659
Dice Loss	0.9241	**0.003053**	0.027146
Focal Tversky Loss	**0.9277**	0.003350	0.027972
Tversky Loss	0.9179	0.003146	0.027306
BCE Loss	0.8962	0.016933	0.052711
Log Dice Loss	0.9174	0.004144	**0.027622**
Hybrid Loss	0.9268	0.005281	0.032427

Table 5: Ablation study on the Spleen dataset with a learning rate of 1×10^{-1} for 40 epochs. Dice Score, Uncertainty Mean, Uncertainty Standard Deviation, Mean Prediction, and Mean Prediction Standard Deviation are reported.

Loss Function	Dice Score	Uncertainty Mean	Uncertainty Std	Mean Prediction	Mean Prediction Std
Log Cosh Loss	0.00002	0.000033	0.000015	0.000003	0.000001
Dice Loss	**0.8698**	**0.000327**	**0.000553**	**0.0040**	**0.0086**
Focal Tversky Loss	0.8185	0.000726	0.001031	0.0055	0.0095
Tversky Loss	0.1704	0.004741	0.005464	0.0050	0.0344
BCE Loss	0.0654	0.013086	0.009289	0.0041	0.0036
Log Dice Loss	0.7732	0.000636	0.000817	0.0045	0.0080
Hybrid Loss	0.8181	0.000908	0.001211	0.0049	0.0088

Table 6: Ablation study on the Spleen dataset with a learning rate of 1×10^{-3} for 40 epochs. Dice Score, Uncertainty Mean, Uncertainty Standard Deviation, Mean Prediction, and Mean Prediction Standard Deviation are reported.

Loss Function	Dice Score	Uncertainty Mean	Uncertainty Std	Mean Prediction	Mean Prediction Std
Log Cosh Loss	0.1852	0.111076	0.000765	0.022004	0.008392
Dice Loss	**0.8726**	**0.001892**	**0.000432**	**0.004636**	**0.008829**
Focal Tversky Loss	0.8432	0.002733	0.000663	0.005209	0.009439
Tversky Loss	0.7867	0.006703	0.000724	0.005831	0.009353
BCE Loss	0.2822	0.071881	0.001007	0.022004	0.008392
Log Dice Loss	0.8307	0.005613	0.000521	0.004984	0.008756
Hybrid Loss	0.7946	0.006085	0.000662	0.005754	0.009247

Table 7: Evaluation on the Spleen dataset using validation loss criteria with a learning rate of 0.01 for 20 epochs.

Loss Function	Dice Score	Uncertainty Mean	Uncertainty Std	Mean Prediction	Mean Prediction Std
Log Cosh Loss	0.0361	0.047324	0.0200	0.0107	0.0079
Dice Loss	0.8697	**0.000239**	**0.0074**	0.0033	0.0562
Focal Tversky	0.7815	0.000552	0.0106	0.0035	0.0576
Tversky Loss	**0.8858**	0.000636	0.0103	0.0039	0.0103
BCE Loss	0.6134	0.013980	0.0168	0.0050	0.0460
Log Dice Loss	0.7695	0.000423	0.0092	0.0029	0.0519
Hybrid Loss	0.8028	0.000956	0.0098	0.0033	0.0548

its robustness across different datasets. This observation is further strengthened by the ablation studies presented in Tables 5 and 6, which investigate the impact of different learning rates and training durations. The confidence metrics (Mean Prediction and its Standard Deviation) further indicate that Dice Loss consistently yields superior performance, demonstrating higher segmentation accuracy and more stable prediction confidence across various experimental setups. In addition to the pancreas and spleen datasets, the proposed approach was also evaluated on the Atlas23 dataset (Table 4). The results exhibit a consistent trend, where Dice Loss continues to deliver superior segmentation accuracy and stable uncertainty calibration, further confirming its robustness and generalizability across different anatomical structures.

Finally, Table 7 demonstrates that even when using validation loss as the model selection criterion, the Dice Loss maintains its dominance, as evidenced by lower uncertainty values and improved segmentation performance. This additional evaluation using validation loss was specifically conducted to mitigate potential bias toward Dice-based metrics. Notably, even with this alternative model selection strategy, Dice Loss continues to outperform other loss functions, reaffirming its effectiveness and stability for medical image segmentation tasks.

4.1. Violin Plots of Different Datasets Under Various Training Regimes

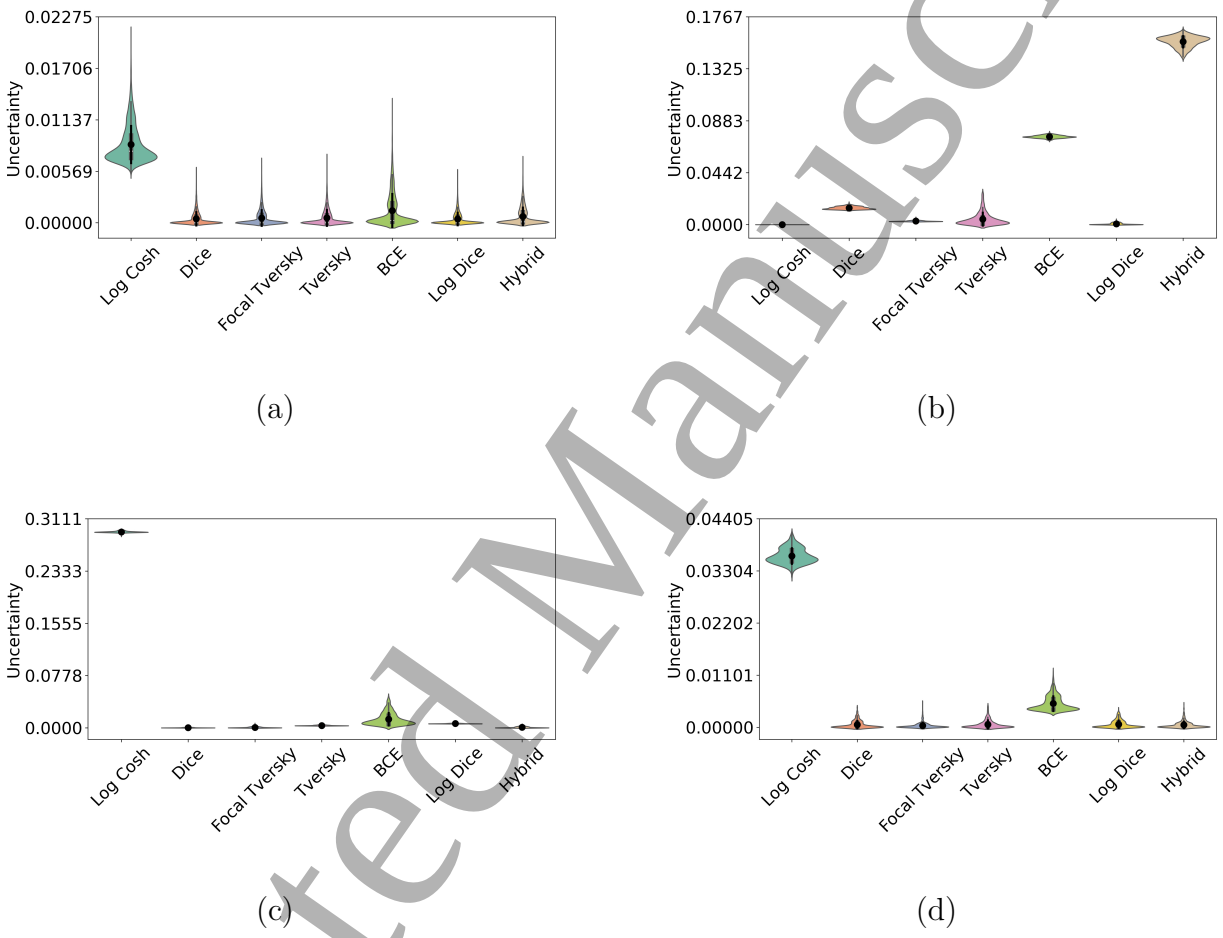
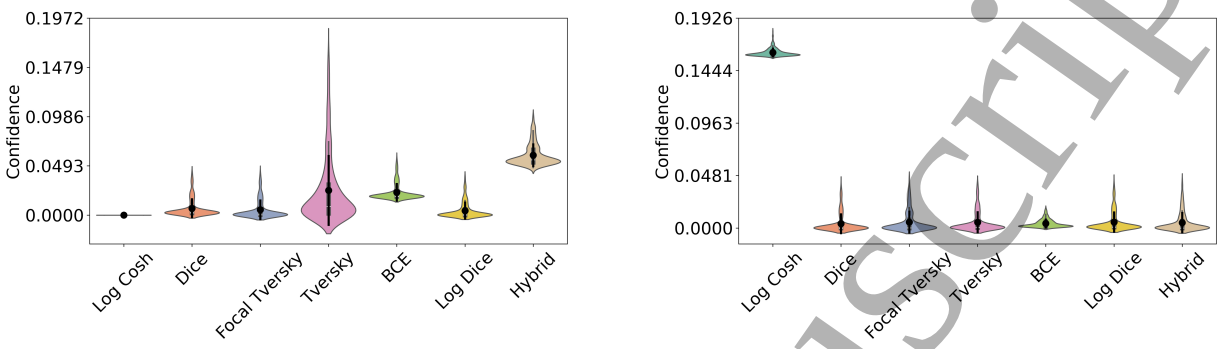


Figure 3: Uncertainty violin plots for different datasets under varying training regimes. (a) Uncertainty violin plot for the pancreatic dataset. (b) Uncertainty violin plot for the spleen dataset with a learning rate of 1×10^{-1}. (c) Uncertainty violin plot for the spleen dataset with a learning rate of 1×10^{-3}. (d) shows the uncertainty violin plot for the Spleen dataset over 20 epochs.

4.2. Confidence Plots for Spleen Dataset with Different Learning Rates



(a) (b)

Figure 4: Confidence plots for the Spleen dataset with different learning rates. (a) Confidence plot for the spleen dataset with a learning rate of 1×10^{-3}. (b) Confidence plot for the spleen dataset with a learning rate of 1×10^{-1}.

4.3. Statistical Analysis of Uncertainty Metrics

Analysis of Variance (ANOVA) is a statistical method used to determine whether there are significant differences between the means of three or more groups. It assesses whether observed variations among group means are likely to be due to actual differences or random chance. In a one-way ANOVA, a single independent variable (factor) was analyzed to determine its effect on a dependent variable. The test calculates an F-statistic, which is the ratio of variance between groups to the variance within groups. A higher F-value suggests a greater disparity among group means, and a corresponding p-value indicates the probability that such differences occurred by chance.

To assess whether the choice of loss function significantly influences the model's uncertainty estimates, we performed a one-way analysis of variance (ANOVA) on the mean uncertainty values obtained from the models trained with different loss functions. The uncertainty values were derived from Monte Carlo dropout predictions and aggregated across the validation set.

The ANOVA test yielded an F-statistic of 2.14×10^8 , with a corresponding p-value of 0 ($p < 0.001$) on the spleen dataset for 20 epochs using a learning rate of 0.01. This highly significant result indicates that there are statistically significant differences in the mean uncertainty estimates among the seven loss functions tested: BCE, Dice, Focal Tversky, Tversky, Log Dice, Log Cosh, and Hybrid.

These findings suggest that the selection of a loss function not only affects segmentation accuracy but also has a profound impact on the model's uncertainty quantification. This

also reaffirms the importance of carefully choosing loss functions in applications in which reliable uncertainty estimation is critical.

5. Discussion

This study provided a comprehensive analysis of various loss functions for pancreas and spleen segmentation using a U-Net architecture, along with an evaluation of uncertainty quantification in medical image segmentation. The quantitative results reported in Tables 2, 3, 5, 6, and 7 consistently demonstrate that Dice Loss achieves the highest Dice Scores and the lowest uncertainty measures across different datasets and training regimes.

The uncertainty violin plots presented in Figure 3 further corroborate these findings. Subfigure (a) illustrates the uncertainty distribution for the pancreas dataset, while subfigures (b), (c), and (d) display the spleen dataset under varying learning rates and epoch settings. In all cases, models optimized with Dice Loss exhibit lower variability in uncertainty estimates compared to those trained with alternative loss functions.

Moreover, the confidence plots in Figure 4 provide additional support by highlighting that the predictions made by models using Dice Loss show higher mean prediction values and reduced prediction variability. This implies that Dice Loss not only enhances segmentation accuracy but also leads to more reliable and consistent predictions—a crucial factor in clinical decision-making.

Visual evidence from the final segmentation results further strengthens our discussion. In the best-case example (Figure 5), the Dice Loss-based model produces a segmentation with well-defined boundaries and a corresponding uncertainty map that aligns closely with the ground truth. Conversely, the worst-case example (Figure 6) reveals that alternative loss functions tend to yield higher segmentation errors and elevated uncertainty levels in challenging regions.

Together, the integration of these quantitative metrics, uncertainty and confidence visualizations, and direct comparisons of segmentation outputs underscores the robustness and generalizability of Dice Loss across various scenarios. These comprehensive findings provide valuable insights for developing high-performance segmentation models where both accuracy and reliable uncertainty estimation are paramount.

5.1. Performance in Segmentation Accuracy

From the segmentation accuracy metrics provided in Tables 2–7, it is evident that Dice loss consistently achieves the highest Dice Coefficient (DICE) across both the pancreas and spleen datasets. In contrast, other loss functions, such as Binary Cross-Entropy (BCE) and Tversky, demonstrated lower DICE values, indicating a suboptimal overlap between the predicted and ground-truth segmentation masks. The performance gap between Dice and these loss functions became particularly noticeable when evaluated under different learning rates and training epochs, as illustrated by the validation Dice Coefficient plots (Figure 3).

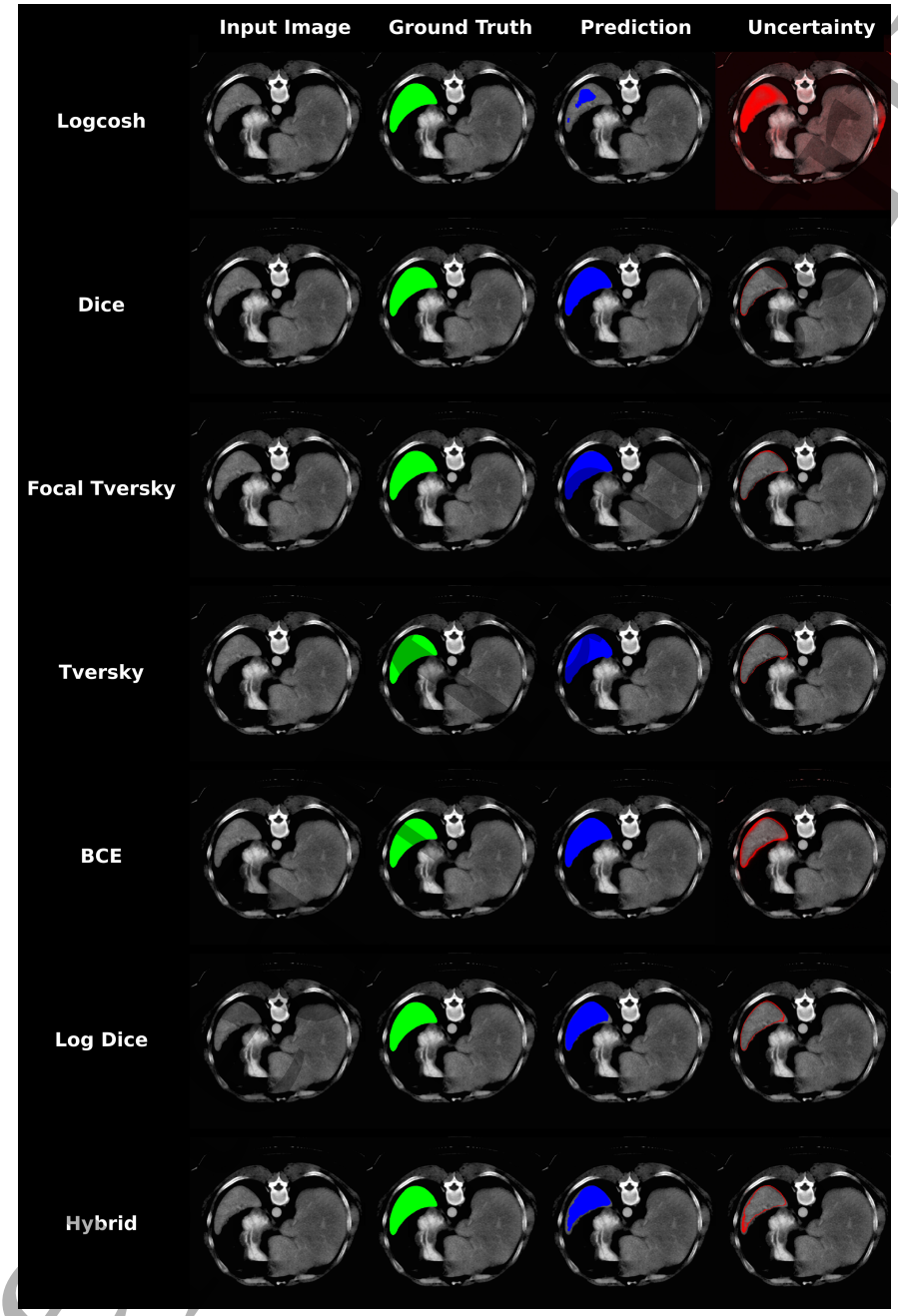


Figure 5: Segmentation results using different loss functions on an example slice of patient 62 in the spleen dataset. The rows represent the performance of different loss functions (BCE, Tversky, Dice, Focal Tversky, Hybrid, Log Dice, Log Cosh) on a medical image segmentation task. Columns display the original image, ground truth, predicted segmentation, and an overlay of the uncertainty map on the original image.

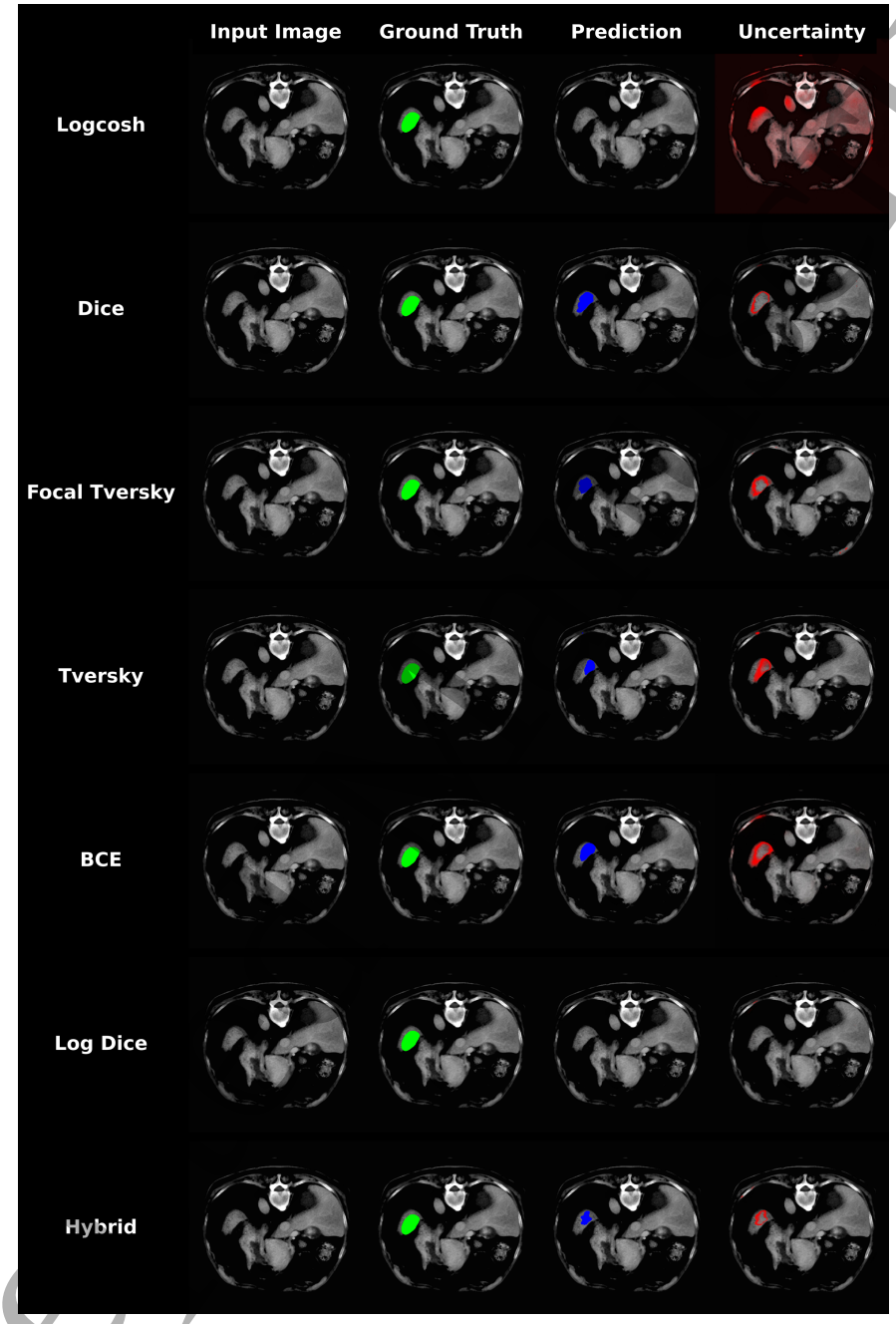


Figure 6: Segmentation results using different loss functions on an example slice of patient 51 in the spleen dataset. The rows represent the performance of different loss functions (BCE, Tversky, Dice, Focal Tversky, Hybrid, Log Dice, Log Cosh) on a medical image segmentation task. Columns display the original image, ground truth, predicted segmentation, and an overlay of the uncertainty map on the original image.

Our experiments consistently demonstrated that Dice loss achieved superior segmentation accuracy and uncertainty calibration across both pancreas and spleen datasets. This can be explained by the characteristics of the Dice formulation and the imaging challenges. Specifically, Dice directly optimizes for region overlap, making it robust to class imbalance where background pixels dominate foreground organ pixels. In CT imaging of the pancreas and spleen, where organs occupy relatively small portions of the field of view and boundaries are often blurred, this property ensures that the model focuses on the foreground organ rather than being overwhelmed by the background.

However, Dice loss also has limitations. In the pancreatic dataset, small tumor regions are extremely underrepresented relative to the organ and background. In such cases, Dice loss can become unstable when the true positive set is very small, leading to poor gradient flow. Alternative loss functions can provide advantages in these scenarios:

- **Tversky and Focal Tversky Losses:** By weighting false positives and false negatives differently, they allow tuning toward higher sensitivity for small tumors, which is critical in early cancer detection.
- **Hybrid Loss:** By combining region-based (Jaccard), pixel-wise (BCE), and structural (SSIM) terms, this loss can better preserve fine anatomical boundaries, particularly when organ contours are blurred.
- **Log Dice Loss:** By applying a logarithmic penalty in low-overlap regions, this loss provides stronger gradients where Dice is weakest, which may help in segmenting very small or irregular structures.

Thus, while Dice loss proved to be a robust baseline across organs and training regimes, other loss functions may offer complementary strengths in specialized scenarios such as tumor segmentation, fine boundary delineation, or highly imbalanced datasets. This suggests a potential avenue for future work: hybrid training schemes that combine Dice with task-specific losses to balance overall accuracy with sensitivity to clinically critical structures.

5.2. Uncertainty Analysis

The violin plots (Figure 4) further highlight the superiority of Dice loss in terms of uncertainty quantification. Models trained with Dice loss exhibit lower uncertainty means and standard deviations than the other loss functions. This suggests that the predictions made by the dice-optimized models are accurate, reliable, and consistent. In contrast, loss functions such as BCE and Log Cosh display higher uncertainty values, indicating greater variability in their predictions.

Notably, the focal tversky loss also performs relatively well in some instances, but tends to produce slightly higher uncertainty and variability compared to Dice, as observed in the spleen dataset experiments with different learning rates. This may be attributed to

the nature of Focal Tversky loss, which emphasizes difficult-to-segment regions, potentially increasing prediction variability in less ambiguous regions.

An important aspect of clinical interpretability is whether uncertainty estimates align with segmentation errors. Visual inspection of the spleen dataset results (Figures 5 and 6) suggests that regions of high predictive uncertainty frequently coincide with boundaries where the model misclassifies organ pixels, as well as with anatomically ambiguous or low-contrast areas. This uncertainty–error overlap indicates that the Monte Carlo Dropout method is not only capturing model variability but also highlighting areas where the segmentation is less reliable. From a clinical perspective, this property is crucial, as it allows uncertainty maps to function as reliability indicators: areas of high uncertainty can alert clinicians to potential segmentation errors, guiding them to review or correct specific regions rather than relying solely on automated outputs.

5.3. Confidence and Mean Prediction

The confidence plots (Figure 4) offer a deeper understanding of how well the model trained with different loss functions generalizes its predictions. Specifically, the Mean Prediction and its Standard Deviation provide insight into the model's confidence in its output. Dice loss, in both learning rate regimes (1×10^{-1} and 1×10^{-3}), exhibits the highest confidence levels with relatively low prediction variability, indicating that it provides accurate, stable, and reliable predictions. This is crucial for medical image segmentation tasks, where confidence in the predicted boundaries is as important as segmentation accuracy itself.

Other loss functions, particularly BCE and Hybrid loss, exhibit higher standard deviations in Mean Prediction, indicating that their predictions fluctuate more across multiple stochastic forward passes. This variability can potentially lead to uncertainty in clinical decisions, making Dice loss a more suitable choice for tasks in which both accuracy and confidence are required.

Throughout the experiments, the Dice loss function consistently emerged as the most robust option across different settings, outperforming other loss functions in terms of higher accuracy, lower uncertainty, and higher confidence. The consistent performance of Dice loss across different datasets, learning rates, and epochs suggests that it can be generalized well to various medical segmentation tasks, providing reliable outcomes in both well-defined and ambiguous segmentation cases.

Note that Dice Loss has been widely used owing to its robustness and simplicity; however, its limitations, such as sensitivity to class imbalance, poor calibration, and lack of region-specific focus, were not studied in this work. On the otherhand, enhanced variants (Focal Tversky and Log Dice) or combining Dice Loss with other loss functions (hybrid loss) is believed to improve segmentation accuracy and reduce uncertainty, but the results shown here shows that the dice loss is superior compared to others. Dice Loss directly optimizes the overlap between predicted and ground truth regions, making it inherently aligned with the evaluation metric (DSC). This simplicity reduces the complexity and potential sources

of error introduced by the additional parameters in the enhanced variants. By focusing solely on overlap, Dice Loss avoids introducing biases from additional weighting factors or hyperparameters, which can sometimes amplify the uncertainty in predictions. It is well known that Dice Loss works well with batch normalization, stabilizing training, and reducing model overconfidence [21].

While other loss functions such as Focal Tversky and Log Dice perform relatively well in certain instances, they often fall short in terms of confidence and prediction stability, which are crucial factors in sensitive applications like medical diagnostics. This robustness of Dice loss can be particularly useful in clinical workflows, where decision making relies on both the accuracy and reliability of predictions.

This study provides a comprehensive evaluation of both segmentation accuracy and uncertainty, which encourages the adaptation of a more holistic approach when assessing the performance of deep learning algorithms, considering not only accuracy, but also uncertainty and confidence in predictions. Usage of uncertainty using entropy has also utility in terms of test time adaptation [22] as well as inference time correction [10] to provide better generalization of models and also mitigate biases. Future algorithms may be designed to output both segmentation results and the associated uncertainty/confidence levels, thereby enhancing their utility in clinical workflows. This could lead to the development of more trustworthy automated decision support systems.

5.4. Generalizability to Other Tasks and Modalities

Although this study focuses on pancreas, spleen segmentation from CT scans and liver segmentation from MRI scans, the findings have broader implications for other anatomical structures and imaging modalities. The consistent performance of Dice loss across datasets and training regimes suggests that its robustness may extend to organs with different anatomical characteristics, such as the liver, kidneys, or lungs, where class imbalance and blurred boundaries are also common challenges. Furthermore, uncertainty quantification using Monte Carlo Dropout is modality-agnostic and can be readily applied to MRI, PET, or ultrasound imaging, where variations in contrast and acquisition parameters introduce additional ambiguity. Recent studies in liver and tumor segmentation, as well as boundary-aware segmentation approaches, similarly emphasize the importance of jointly considering segmentation accuracy and uncertainty calibration. Hence, while the pancreas and spleen serve as representative case studies, the principles demonstrated here, particularly the advantages of Dice loss for stable performance and the interpretive value of uncertainty maps are expected to generalize to a wide range of medical image segmentation tasks. Future work should empirically validate this generalizability across diverse organ systems and modalities.

6. Conclusion

This study provided a comprehensive analysis of various loss functions in the context of uncertainty quantification for pancreas and spleen segmentation using U-Net architecture. Across both datasets and under different training regimes, whether varying the number of epochs or the learning rate, Dice Loss consistently emerged as the most robust performer. Notably, the Dice Loss achieved the highest Dice Coefficient along with low uncertainty, demonstrating superior accuracy and reliability in predictive confidence. This consistent performance underscores the versatility and robustness of Dice Loss, making it a strong candidate for clinical applications where both accuracy and confidence in segmentation are paramount. In contrast, other loss functions, including Log-Cosh, Focal Tversky, and Hybrid Loss, displayed greater variability in both segmentation performance and uncertainty calibration. While these functions showed potential in specific configurations, they lacked the consistency exhibited by Dice Loss across a range of learning rates and epoch settings. Furthermore, the ability of the model to maintain high performance despite adjustments in hyperparameters suggests its robustness and suitability in diverse clinical environments. Future work will focus on integrating uncertainty measures into clinical workflows and enhancing trust in automated decision-support systems that can output both segmentation results and associated uncertainty/confidence levels. These advancements could significantly improve the reliability and applicability of automated segmentation tools in medical practice.

Data Availability

The Pancreas and Spleen datasets used in this study are publicly available as part of the *Medical Segmentation Decathlon (MSD)* challenge at http://medicaldecathlon.com/.

The Atlas23 dataset used for additional validation is not publicly available but can be accessed upon reasonable request from the corresponding authors of the paper titled *"A Tumour and Liver Automatic Segmentation (ATLAS) Dataset on Contrast-Enhanced Magnetic Resonance Imaging for Hepatocellular Carcinoma"* by Félix Quinton *et al.* [23]. The dataset was released as part of the ATLAS challenge. Further details can be found at https://doi.org/10.3390/data8050079.

Ethical Statement

This study used publicly available, anonymized medical imaging datasets (pancreas and spleen segmentation datasets from the Medical Segmentation Decathlon). The Institutional Review Board confirmed that no additional approval was required, and a waiver of ethical approval was granted. The research was conducted in accordance with the principles of the Declaration of Helsinki and complied with all relevant institutional and national regulations.

References

[1] M. Kim, J. Yun, Y. Cho, K. Shin, R. Jang, H.-J. Bae, and N. Kim, "Deep learning in medical imaging," Neurospine, vol. 16, no. 4, pp. 657–668, Dec. 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31905454/

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241. [Online]. Available: https://arxiv.org/abs/1505.04597

[3] D. Park, "Uncertainty estimation in volumetric image segmentation," Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2023.

[4] M. C. Krygier, T. LaBonte, C. Martinez, C. Norris, K. Sharma, L. N. Collins, P. P. Mukherjee, and S. A. Roberts, "Quantifying the unknown impact of segmentation uncertainty on image-based simulations," Nature Communications, vol. 12, no. 5414, pp. 1–13, 2021. [Online]. Available: https://doi.org/10.1038/s41467-021-25493-8

[5] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," Information Fusion, vol. 76, pp. 243–297, 2021. [Online]. Available: https://doi.org/10.1016/j.inffus.2021.05.008

[6] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher et al., "A survey of uncertainty in deep neural networks," Artificial Intelligence Review, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.

[7] S. Faghani, M. Moassefi, P. Rouzrokh, B. Khosravi, F. I. Baffour, M. D. Ringler, and B. J. Erickson, "Quantifying uncertainty in deep learning of radiologic images," Radiology, vol. 308, no. 2, p. e222217, 2023.

[8] S. Jadon, "A survey of loss functions for semantic segmentation," in IEEE International Conference on E-Learning, E-Management and E-Services (IC3e), 2020, pp. 1–5. [Online]. Available: https://github.com/shruti-jadon/Semantic-Segmentation-Loss-Functions

[9] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, and A. L. Martel, "Loss odyssey in medical image segmentation," Medical Image Analysis, vol. 71, p. 102035, 2021. [Online]. Available: https://doi.org/10.1016/j.media.2021.102035

[10] J. Jeffrey, A. RajKumar, S. Pandey, L. Bathala, and P. K. Yalavarthy, "Inference time correction based on confidence and uncertainty for improved deep-learning model per-

formance and explainability in medical image classification," Computerized Medical Imaging and Graphics, vol. 125, p. 102630, 2025.

[11] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proceedings of the 22nd international conference on Machine learning, 2005, pp. 625–632.

[12] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue, vol. 16, no. 3, pp. 31–57, 2018.

[13] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," cell, vol. 172, no. 5, pp. 1122–1131, 2018.

[14] C. Syrykh, A. Abreu, N. Amara, A. Siegfried, V. Maisongrosse, F. X. Frenois, L. Martin, C. Rossi, C. Laurent, and P. Brousset, "Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning," NPJ digital medicine, vol. 3, no. 1, p. 63, 2020.

[15] A. E. Flanders, L. M. Prevedello, G. Shih, S. S. Halabi, J. Kalpathy-Cramer, R. Ball, J. T. Mongan, A. Stein, F. C. Kitamura, M. P. Lungren et al., "Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge," Radiology: Artificial Intelligence, vol. 2, no. 3, p. e190211, 2020.

[16] Z. Senousy, M. M. Abdelsamea, M. M. Mohamed, and M. M. Gaber, "3e-net: Entropy-based elastic ensemble of deep convolutional neural networks for grading of invasive breast carcinoma histopathological microscopic images," Entropy, vol. 23, no. 5, p. 620, 2021.

[17] E. A. Krupinski, "The importance of perception research in medical imaging," Radiation medicine, vol. 18, no. 6, pp. 329–334, 2000.

[18] 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.

[19] S. Balasubramaniam, Y. Velmurugan, D. Jaganathan, and S. Dhanasekaran, "A modified lenet cnn for breast cancer diagnosis in ultrasound images," Diagnostics, vol. 13, no. 17, p. 2746, 2023.

[20] A. Titoriya and S. Sachdeva, "Breast cancer histopathology image classification using alexnet," in 2019 4th International conference on information systems and computer networks (ISCON). IEEE, 2019, pp. 708–712.

[21] A. Mehrtash, W. M. W. III, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," IEEE Transactions on Medical Imaging, vol. 39, no. 12, pp. 3868–3879, 2020. [Online]. Available: https://doi.org/10.1109/TMI.2020.3006437

[22] H. Ravishankar, N. Paluru, P. Sudhakar, and P. K. Yalavarthy, "Information geometric approaches for patient-specific test-time adaptation of deep learning models for semantic segmentation," IEEE Transactions on Medical Imaging, vol. 44, no. 6, pp. 2553–2567, 2025.

[23] F. Quinton, R. Popoff, B. Presles, S. Leclerc, F. Meriaudeau, G. Nodari, O. Lopez, J. Pellegrinelli, O. Chevallier, D. Ginhac, J.-M. Vrigneaud, and J.-L. Alberini, "A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma," Data, vol. 8, no. 5, p. 79, 2023.