KSÜ Mühendislik Bilimleri Dergisi, 29(3), 2026
Araştırma Makalesi

KSU J Eng Sci, 29(3), 2026
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 30.03.2026
Kabul Tarihi : 04.06.2026

Received Date : 30.03.2026
Accepted Date : 04.06.2026

INHERENT CLASS IMBALANCE IN LESION SEGMENTATION: THE EFFECT OF LOSS FUNCTIONS

LEZYON BÖLÜTLENMESİNDE DOĞAL SINIF DENGESİZLİĞİ: KAYIP FONKSİYONLARININ ETKİSİ

Umut Kaan KAVAKLI[1] (ORCID: 0009-0003-4968-4124)
Göksu TURAÇ[1,2] (ORCID: 0009-0002-3480-2194)
Sema CANDEMİR[1*] (ORCID: 0000-0001-8619-5619)

[1] Eskisehir Technical University, Artificial Intelligence in Healthcare Laboratory, Computer Engineering Department, Eskisehir, Türkiye
[2] Istanbul Gelisim University, Faculty of Engineering and Architecture, Department of Software Engineering, Istanbul, Türkiye
The first and second authors have contributed equally to the study.
*Sorumlu Yazar / Corresponding Author: Sema CANDEMİR, semacandemir@eskisehir.edu.tr

ABSTRACT

Medical image segmentation plays a critical role in clinical analysis by delineating anatomically or pathologically relevant regions. A major challenge in lesion segmentation is class imbalance where healthy tissues outnumber abnormal regions, biasing learning algorithms toward the majority class and degrading lesion detection. Inter-class and intra-class variability of lesions further complicate the segmentation. Appropriate loss functions can mitigate these challenges. In this study, we evaluate widely used loss functions for COVID-19 lesion segmentation using publicly available CT datasets. A U-Net–based framework is evaluated under varying imbalance ratios and lesion characteristics. The lesion heterogeneity is assessed using entropy, lesion area, gradient magnitude, and intensity differences. Results show that imbalance-aware and region-based losses provide robust performance under heterogeneous lesion distributions. The findings highlight that loss function selection should consider class imbalance, lesion heterogeneity, and structural complexity for reliable segmentation.

Keywords: Data imbalance, loss function, lesion segmentation, inter- and intra-class variability

ÖZET

Tıbbi görüntü segmentasyonu, anatomik veya patolojik açıdan önemli bölgelerin ayrıştırılmasını sağlayarak klinik analizlerde kritik bir rol oynamaktadır. Lezyon segmentasyonundaki temel zorluklardan biri, sağlıklı dokuların anormal bölgelerden fazla olması nedeniyle öğrenme algoritmalarını çoğunluk sınıfına yönlendiren sınıf dengesizliğidir. Lezyonların sınıflar arası ve sınıf içi değişkenliği segmentasyonu daha da zorlaştırmaktadır. Uygun kayıp fonksiyonları bu zorlukları azaltabilir. Bu çalışmada, COVID-19 BT görüntülerinde bu zorlukların etkisini azaltmaya yönelik farklı kayıp fonksiyonları U-Net tabanlı bir model kullanılarak değerlendirilmiştir. Lezyon heterojenitesi; entropi, lezyon alanı, gradyan büyüklüğü ve yoğunluk farkları kullanılarak değerlendirilmiştir. Sonuçlar, dengesizliğe duyarlı ve bölge tabanlı kayıp fonksiyonlarının heterojen lezyon dağılımlarında güçlü performans sağladığını göstermektedir. Bulgular, güvenilir segmentasyon için kayıp fonksiyonu seçiminde sınıf dengesizliği, lezyon heterojenitesi ve yapısal karmaşıklığın dikkate alınması gerektiğini ortaya koymaktadır.

Anahtar Kelimeler: Veri dengesizliği, kayıp fonksiyonu, lezyon bölütlemesi, sınıflar arası ve içi değişkenlik

INTRODUCTION

Image segmentation partitions an image into distinct and meaningful regions for subsequent analysis (Stockman & Shapiro, 2001). A key challenge in image segmentation is data imbalance, where foreground and background pixels differ significantly (Nguyen & Thai, 2023). This issue is particularly visible in lesion segmentation (Ma et al., 2021,

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Taghanaki et al., 2019), in which there is an intrinsic data imbalance due to the relatively small size of anomalous tissue compared to the surrounding healthy tissue. Such imbalance adversely affects the performance of segmentation algorithms by biasing them toward the majority class, thereby limiting their ability to accurately identify and delineate minority regions. Addressing this challenge is essential for obtaining accurate and effective results in medical imaging.

One common strategy to address data imbalance is data augmentation, which increases the number of training samples by altering existing images. However, this approach affects both lesion and non-lesion areas equally, and thus fails to resolve the inherent class imbalance. Another widely adopted strategy is to design loss functions that assign higher penalties to the misclassification of minority classes (Johnson & Khoshgoftaar, 2019). This study focuses on loss functions that mitigate class imbalance and improve the segmentation performance of minority regions. A loss function quantifies the difference between predicted and actual values during training (Géron, 2022). Several loss functions have been proposed in the literature, which can be broadly categorized into four groups: pixel-level, region-level, boundary-level, and hybrid functions (Azad et al., 2023, Jadon, 2020). In this study, we systematically evaluate these loss functions in the context of lesion segmentation in medical imaging.

We analyze the performance of loss functions under class imbalance using two publicly available COVID-19 lesion segmentation datasets based on Computed Tomography (CT) scans. In March 2020, the American College of Radiology recommended the use of medical imaging primarily for monitoring rather than diagnosing COVID-19 (American College of Radiology, 2020). One of the key steps in monitoring COVID-19 patients is the quantitative assessment of lesion segmentation. Manual delineation of lesions is labor-intensive and time-consuming. Therefore, machine learning–based approaches have become valuable tools for automated lesion analysis. Accurate segmentation of infected regions in CT scans enables reliable quantification and monitoring of disease progression (Tang et al., 2021; Fan et al., 2020). However, segmenting infected regions remains a challenging task for ML-based models due to several factors (Li et al., 2020; Sun et al., 2020): (i) low-intensity contrast between infected lesions and normal tissue; (ii) high variability in infection characteristics, including differences in texture, size, and spatial distribution; (iii) the presence of noise and imaging artifacts that may obscure lesion boundaries; (iv) the time-consuming process of obtaining high-quality pixel-level annotations (Sun et al., 2020); and (v) inherent class imbalance between infected lesions and normal tissue.

In this study, we provide a taxonomy of loss functions, analyze them, and examine their strengths and weaknesses in the context of lesion segmentation. Although the loss functions evaluated in this work are well-established in the literature, the primary contribution of our study lies in providing a systematic evaluation specifically for inherent class imbalance in lesion segmentation. To assess the performance of these loss functions under inherent data imbalance, we train a U-Net–based model (Ronneberger et al., 2015) using publicly available COVID-19 CT datasets. This study makes several contributions to the field of medical image segmentation under inherent class imbalance. First, we conduct a systematic evaluation of widely used loss functions specifically under inherent imbalance conditions. Second, beyond the conventional focus on class imbalance between lesion and non-lesion regions, we incorporate intra-class variability through a quantitative assessment using entropy, lesion area, gradient magnitude, and intensity differences. This analysis reveals the complexity of lesion characteristics both within and across datasets and highlights that loss function selection should consider not only class imbalance but also lesion heterogeneity and structural complexity, which is often overlooked in prior work. We evaluate all loss functions on two publicly available COVID-19 CT datasets with significantly different imbalance ratios (83:1 and 25:1), ensuring that the findings are not dataset-specific and demonstrate generalizability. The study shows that Focal Tversky Loss consistently outperforms other loss functions in handling both imbalance and lesion variability. The findings highlight that selecting an appropriate loss function should consider not only class imbalance but also data heterogeneity and lesion characteristics, which has direct implications for real-world clinical applications.

RELATED WORKS

The data imbalance problem is thoroughly addressed in (Johnson & Khoshgoftaar, 2019), a comprehensive survey that categorizes solutions for data imbalance into two main groups: data-level and algorithm-level methods. Data-level methods modify training data to achieve a balanced distribution, while algorithm-level methods adjust the learning algorithm to handle an imbalanced data distribution. One of the data-level solutions is data augmentation, which artificially increases the number of training samples by altering existing images. However, altering the whole image affects both lesion and non-lesion areas equally. Consequently, it fails to address the inherent class imbalance.

A common approach is a patch-based training strategy, where images are divided into smaller patches and some non-lesion patches are discarded to reduce the imbalance between foreground and background regions. However, this limits the model's ability to learn region-of-interest information in a global context (Nguyen & Thai, 2023).

The literature presents various algorithm-level approaches to handle data imbalance by adjusting the learning algorithm itself. Nguyen et al. (Nguyen & Thai, 2023) analyzed the impact of loss functions on crack segmentation, which involves a drastic imbalance between foreground and background regions. The study used four distinct imbalanced datasets. In most cases, the Focal Tversky Loss achieved the best results. In the context of COVID-19 lesion segmentation, Fan et al. (Fan et al., 2020) proposed Inf-Net, a lung infection segmentation algorithm for CT sequences, using the SemiSeg dataset. The model is trained using a semi-supervised approach: first on labeled data, then on pseudo-labels generated from unlabeled data, which are combined to improve performance. This study highlights the challenges of COVID-19 lesion segmentation, particularly the scarcity of high-quality training data with pixel-level annotations of infections on CT slices. However, it does not focus on inherent data imbalance. Shan et al. (Shan et al., 2021) developed a system using a voxel-based network combined with a human-in-the-loop strategy, where radiologists monitored the model's predictions and refined the annotations for each case. This study highlights the challenges of COVID-19 lesion segmentation and demonstrates that results can be improved with a human-in-the-loop strategy. Hashemi et al. (Hashemi et al., 2018) proposed a 3D-CNN trained with an asymmetric similarity loss based on the Tversky index, which penalizes missing pathological pixels more than incorrectly labeling a normal one as pathological. This method was applied to multiple sclerosis lesion segmentation.

The contributions of this study are as follows. We provide a taxonomy of loss functions for the segmentation task, focusing on inherent data imbalance. We systematically evaluate widely used loss functions using the U-Net algorithm to assess their effectiveness in addressing data imbalance in COVID-19 lesion segmentation. Beyond the conventional focus on foreground–background imbalance, we incorporate intra-class heterogeneity by analyzing lesion characteristics using quantitative measures such as entropy, lesion area, gradient magnitude, and intensity differences. We evaluate all loss functions using a U-Net–based framework on two publicly available COVID-19 CT datasets with significantly different imbalance ratios, ensuring that the findings are not dataset-specific. Our experiments provide insights into how different loss functions handle both variability and imbalance within and across datasets of COVID-19 lesions. These results emphasize the importance of selecting appropriate loss functions for training more effective lesion segmentation models.

TAXONOMY OF LOSS FUNCTIONS

One strategy to mitigate data imbalance in segmenting small regions is to adjust the loss functions. Different loss functions provide distinct learning processes and update the model weights based on error feedback. There are various loss functions available in the literature. The loss functions used in this study are listed in Table 1. Broadly, they can be categorized into four groups: pixel-level, region-level, boundary-level, and hybrid functions. In this section, we provide a brief description of each loss function analyzed for lesion segmentation.

Table 1. Comparison Of Loss Functions Based On Their Characteristics

Loss Function	Characteristics
Cross-Entropy Loss	Standard loss for classification tasks; measures the difference between predicted and actual labels.
Weighted Cross-Entropy	Adds weights to classes to handle class imbalance in classification tasks.
Dice Loss	Focuses on overlap; calculated using the Dice coefficient, and is effective in imbalanced segmentation tasks.
Dice Weighted BCE Loss	Combines Dice Loss and Weighted Binary Cross-Entropy to address class imbalance.
Jaccard Loss	Measures similarity based on Intersection over Union (IoU); emphasizes region-level agreement.
Focal Loss	Reduces the impact of well-classified examples; focuses on hard or misclassified cases.
Tversky Loss	Generalizes Dice Loss; allows tuning of penalties for false positives and false negatives.
Focal Tversky Loss	Combines Focal Loss and Tversky Loss to emphasize difficult examples and reduce false negatives.

Weighted Binary Cross-Entropy Loss

The cross-entropy loss function, widely used for classification and segmentation tasks, measures the dissimilarity between the predicted probability and the target label. The cross-entropy loss function treats errors at each pixel with

equal weight, penalizing false positives and false negatives equally. However, in contexts like lesion segmentation, where non-lesion pixels are more abundant, the algorithm can become biased toward non-lesion areas. This imbalance often leads to an increase in false negatives, as the model prioritizes identifying non-lesion areas over identifying areas with lesions. A variant of the BCE loss, called Weighted Binary Cross Entropy (WBCE), assigns weights based on the proportion of imbalanced data. One such method is the inverse frequency weighting method, which assigns a higher weight to the smaller sample (lesion areas). The formula for WBCE:

$$L_{wBCE} = -\sum_{i=1}^{N}\left(\frac{1}{N_l} y_i \log(\hat{y}_i) + \frac{1}{N_{nl}} (1 - y_i)\log(1 - \hat{y}_i)\right) \quad (1)$$

where N_l and N_{nl} represent the number of lesion and non-lesion pixels, respectively.

The Weighted Binary Cross Entropy loss function balances inter-class variance by adjusting the loss according to the inverse proportion of the number of pixels, thus forming a better decision boundary. This approach helps to reduce the impact of the imbalance on the loss error. However, it fails to consider intra-class variances and challenging cases. Lesions may exhibit variability, leading to uneven distributions within the same class. As a result, the model may struggle to make correct assumptions about the samples, ultimately affecting its performance in detecting and segmenting lesions effectively (Zhu et al., 2022).

Focal Loss

The Focal Loss (FL) (Lin et al., 2017) is designed for extreme imbalance between foreground and background pixels. It has two key parameters: a modulating factor $(1 - \hat{y}_i)^{\gamma}$ and a weighting parameter α. The modulating factor penalizes the predictions with low confidence, emphasizing challenging cases and encouraging the model to predict with higher certainty. The modulating factor parameter helps mitigate intra-class variations by focusing on hard-to-classify lesions. The weighting parameter $\alpha \in [0,1]$ balances the number of foreground pixels and the number of background pixels. Similar to the Weighted Binary Cross Entropy, α is set inversely proportional to the frequency of each class. This means that the less frequent class (lesion pixels) receives higher weights. The focal loss effectively addresses inter-class variability with α and provides a more balanced model.

$$L_{FL} = -\sum_{i=1}^{N} \alpha_i (1 - \hat{y}_i)^{\gamma} log(\hat{y}_i), \qquad \hat{y}_i = \begin{cases} \hat{y}_i, & if y = 1 \\ 1 - \hat{y}_i, & otherwise \end{cases} \quad (2)$$

where y_i is the predicted probability of the model, α_i and γ are the loss function parameters taken as 0.25 and 2, respectively, as proposed in (Zhu et al., 2022).

Dice Loss

Dice loss aims to maximize the overlapping regions between ground truth and predicted masks (Ma et al., 2021). This loss function is obtained by subtracting 1.0 from the Dice similarity coefficient.

$$L_{Dice} = 1 - \frac{2Y\hat{Y} + \varepsilon}{Y + \hat{Y} + \varepsilon} \quad (3)$$

Y and $\hat{Y}$ are the flattened ground truth and predicted mask pixels, respectively. They are multiplied element-wise and summed. ε is added to prevent division by zero.

Tversky Loss

The Tversky Loss (TL) is a variation of Dice loss aimed at addressing class imbalance. Dice loss treats false positive and false negative pixels equally (Abraham & Khan, 2019), which is inadequate when dealing with dominant background pixels as in lesion segmentation, often leading to a high number of false negative predictions. To tackle this issue, TL uses a generalization of the Dice score and Jaccard index called the Tversky similarity index, which has two hyperparameters, α and β. These parameters adjust the trade-off between false negatives and false positives (recall and precision), respectively, to handle the imbalance between foreground and background pixels (Abraham & Khan, 2019, Salehi et al., 2017).

$$L_{TL} = 1 - \frac{Y\hat{Y}+\epsilon}{Y\hat{Y}+\alpha(1-Y)\hat{Y}+\beta Y(1-\hat{Y})+\epsilon} \quad (4)$$

The two hyperparameters are used in a combination such that $\alpha + \beta = 1$. The best results were obtained for $\alpha = 0.3$ and $\beta = 0.7$ in the studies (Nguyen & Thai, 2023, Abraham & Khan, 2019, Salehi et al., 2017).

Focal Tversky Loss

An additional enhancement of the Tversky Loss is an effective combination of the advantages of Focal Loss and Tversky Loss, known as Focal Tversky Loss (FTL) (Lin et al., 2017, Salehi et al., 2017). The addition of the focal parameter exponentiates the loss of small and hard-to-detect regions (Abraham & Khan, 2019). Consequently, the overall loss function becomes more sensitive to these areas.

$$L_{FTL} = \left(1 - \frac{Y\hat{Y}+\epsilon}{Y\hat{Y}+\alpha(1-Y)\hat{Y}+\beta Y(1-\hat{Y})+\epsilon}\right)^{\gamma} \quad (5)$$

where γ takes values between 1.25 and 2. The researchers in the study (Abraham & Khan, 2019) reported that they observed the best performance with $\gamma = \frac{4}{3}$.

Combination Of Loss Functions

Some researchers have suggested that integrating multiple loss functions can leverage their individual strengths, thereby improving model performance. Frequently used combined loss functions are Dice loss with Binary Cross-Entropy loss and Dice loss with Focal loss (Ma et al., 2021, Zhu et al., 2019). In this study, we analyze the performance of combinations of Dice with Focal and WBCE loss functions. Each of these loss functions has different advantages: Dice loss is effective for segmentation tasks by measuring overlap, Focal loss addresses intra-class variability by focusing on hard-to-classify examples, and WBCE adjusts for unequal class distributions. Various weighting parameters can be applied to each component independently to adjust their effects. In this study, we use equal weights and combine the loss functions to demonstrate the fundamental impact of each component.

$$L_{Dice+Focal} = L_{Dice} + L_{Focal} \quad (6)$$
$$L_{Dice+WBCE} = L_{Dice} + L_{WBCE} \quad (7)$$

DATASET FOR LESION SEGMENTATION

We use the following datasets to develop and test our COVID-19 lesion segmentation model.

CT Images In COVID-19 Dataset

The dataset consists of CT images of patients who tested positive for COVID-19 along with ground-truth lesion delineations associated with the disease (An et al., 2020). The dataset contains 199 publicly available CT scans. In this study, the dataset was divided on a patient basis to prevent data leakage between the training, validation, and test sets. The dataset was divided into training, validation, and test sets using a 60%-20%-20% split ratio. Table 2 summarizes the dataset partitions used in the experiments. Figure 1 shows example slices with clinician-delineated segmentation overlays.

COVID-SemiSeg Dataset

The COVID-SemiSeg dataset was developed for semi-supervised COVID-19 lesion segmentation studies by combining the COVID-19 CT segmentation dataset presented in (Fan et al., 2020), which contains 100 CT scans with manually delineated lesion masks, and additional CT images collected from the COVID-19 CT collection dataset (Cohen et al., 2020). In this study, the dataset was divided on a patient basis to prevent data leakage between the training, validation, and test sets. Table 2 summarizes the dataset partitions used in the experiments. The datasets were divided into training, validation, and test sets using a 60%-20%-20% split ratio. Table 3 presents the

background-to-foreground (BG) pixel ratio for the study datasets and highlights the inherent class imbalance between lesion and non-lesion regions.

Table 2. Summary Of Datasets Used For Training, Validation, And Testing (Number Of Frames)

Dataset	Training Set (%60)	Validation Set (%20)	Test Set (%20)
CT Images in COVID - 19	2834 CT frames with clinician-delineated lesions	944 CT frames with clinician-delineated lesions	944 CT frames with clinician-delineated lesions
COVID-SemiSeg	1237 CT frames with clinician-delineated lesions	413 CT frames with clinician-delineated lesions	413 CT frames with clinician-delineated lesions

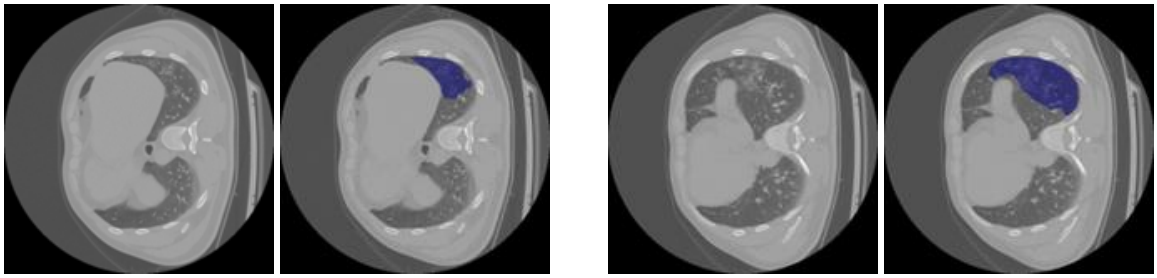

Figure 1. Examples Of CT Slices And Corresponding Ground-truth Lesion Delineations With Colored Overlays

Table 3. Comparison Of Dataset Resolutions And Background-to-Foreground (BG:FG) Pixel Ratios

Dataset	Resolution	BG:FG
CT Images in COVID-19	512 x 512	83:1
COVID-SemiSeg	382 x 276	25:1

LESION SEGMENTATION

Lesion segmentation is the process of delineating abnormal tissue regions affected by the disease. It presents inherent data imbalance due to the relatively small size of anomalous tissue compared to the surrounding healthy tissue. The high variation in infection characteristics within the same class and across different datasets adds an additional challenge for the algorithm. Effectively addressing these challenges is essential for obtaining accurate results in image segmentation (Fan et al., 2020, Hashemi et al., 2018). Figure 1 shows clinician-delineated lesion regions on example CT slices.

We have utilized the U-Net algorithm (Ronneberger et al., 2015) to segment COVID-19 lesions. The U-Net architecture contains two paths: an encoder and a decoder. The encoder consists of sequential convolutional and pooling layers that model the semantic information in the image. The decoder recovers the location information, which is lost during the pooling process. The decoder layer employs feature maps from the corresponding encoder layer that help achieve a more accurate expansion process, thus providing more accurate segmentation and localization. We have trained the U-Net model with the COVID-19 lesion datasets reported in Table 2. The primary objective of the loss function in a U-Net architecture is to quantify the difference between the predicted segmentation masks and the ground truth masks. This difference is a measure of how accurately the model segments the lesions in the training data. Through continuous iterations, the model adjusts its weights in a way that minimizes the loss function. After model prediction, the probability values of each pixel range between 0 and 1. With a predetermined threshold value, each pixel is assigned to either the foreground or the background.

Analysis Of Inter- And Intra- Variability Of Lesions

Accurate segmentation of infected regions remains a challenging task, particularly when loss functions fail to adequately account for both inter-class imbalance and intra-class variability. While the imbalance between foreground and background regions (Table 3) is a well-recognized issue, an equally critical yet often overlooked challenge is the substantial heterogeneity in lesion characteristics within and across datasets. These variations manifest in differences in lesion size, texture, intensity distribution, and spatial localization, all of which significantly influence segmentation performance. Figure 2 shows examples of intra-class variability in lesion areas from the

COVID-SemiSeg dataset. As observed, lesion regions range from very small and sparse patterns to large and diffuse structures occupying a significant portion of the lung.

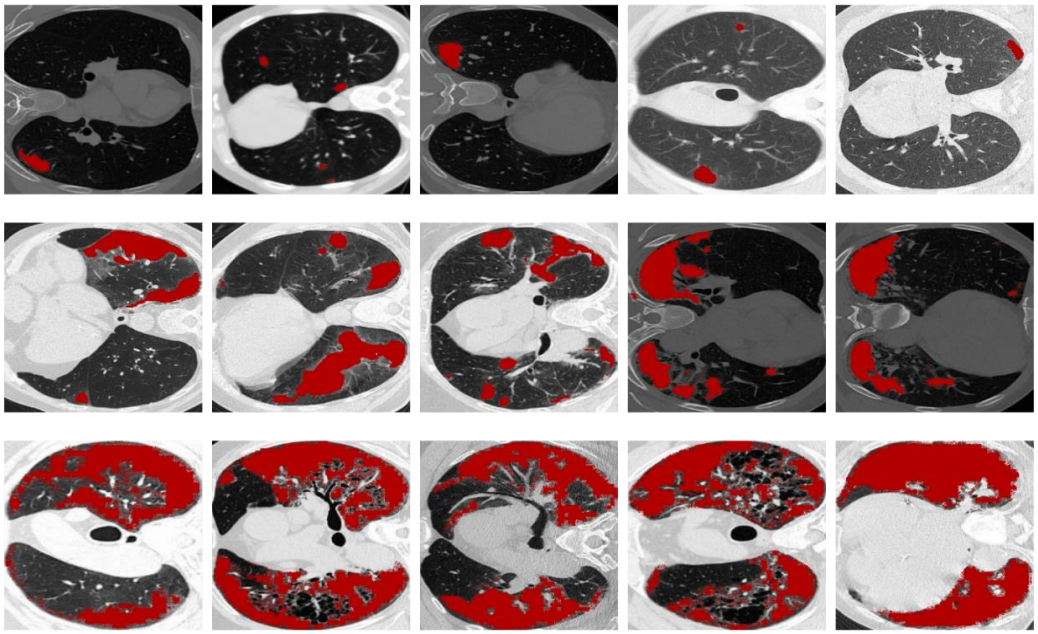

Figure 2. Examples Of Intra-Class Variability In Lesion Areas From The CT Images In COVID-SemiSeg Dataset. The Top Row Shows Images With Very Small Lesion Areas, The Middle Row Presents Images With Moderate Lesion Areas, And The Bottom Row Displays Images Where Lesions Occupy A Large Portion Of The Lung Area

To illustrate the inter- and intra-class variability of lesion areas within and across datasets, we computed the following features: entropy, area, gradient magnitude, and intensity differences of lesions from the SemiSeg dataset and the COVID-19 dataset. Entropy quantifies the structural complexity and randomness within lesion regions, where higher values indicate more heterogeneous and irregular patterns. Lesion area reflects the scale of infection, with smaller lesions posing greater difficulty due to the dominance of non-lesion pixels. Gradient magnitude captures boundary sharpness, where lower gradients correspond to blurred or poorly defined lesion boundaries, making segmentation more challenging. Intensity difference measures the contrast between lesion and surrounding tissue; lower contrast levels further obscure lesion boundaries and complicate detection. The distributions of these features are visualized through histograms in Figure 3 for both datasets. The histograms reveal that lesions exhibit substantial variability in terms of complexity, size, and boundary sharpness, confirming that lesion characteristics differ significantly both within a single dataset and across different datasets. The variability observed in these features highlights the necessity for loss functions to tune their parameters in a way that handles different lesion types both within and across datasets.

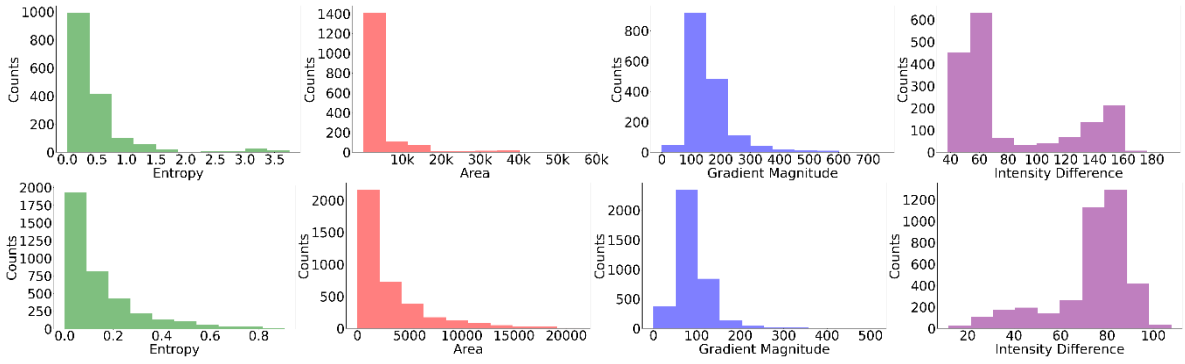

Figure 3. Distribution Histograms Of Entropy, Area, Gradient Magnitude, And Intensity Difference Of Lesions In The SemiSeg (Top) And COVID-19 Dataset (Bottom)

EXPERIMENTAL RESULTS AND ANALYSIS

Implementation Details

The development of the lesion segmentation algorithm and loss function comparison experiments were conducted in Python using the TensorFlow Keras API and trained on an NVIDIA RTX 4090 system with 16 GB of graphics memory with CUDA v11 dependencies for GPU acceleration. The data were organized into 60% training, 20% validation, and 20% test subsets. We set 150 epochs for training iterations and monitored the learning curves to save the best weights, considering the maximum validation Dice score to prevent overfitting. During training, the CT frames were resized to 256 × 256. No data augmentation techniques were applied, as the primary objective of this study is to provide a controlled comparison of loss functions without introducing additional variability from augmented data.

Hyperparameter Tuning

We systematically searched for the best combination of loss hyperparameters. For each loss function, we used a range of values. We varied the α parameter from 0.10 to 0.70, the β parameter as $1-\alpha$, and the γ parameter from 1.25 to 2.00. The α and β values were incremented by 0.05 at each step. The segmentation network and training protocol parameters were determined empirically. Hyperparameter tuning was performed on the validation set to avoid any data leakage. The best hyperparameter values reported in Table 4 were selected based on the highest validation Dice score. For each combination, early stopping criteria was applied. The test set was strictly kept unseen during this process.

Table 4. Summary Of Hyperparameters Used In The Experiments

Hyperparameter	Value
Learning Rate	0.001
Batch Size	32
Epoch	150
Optimizer	Adam, $\beta_1 = 0.9$, $\beta_2 = 0.999$
Tversky	$\alpha = 0.15$
Focal	$\alpha = 0.3$, $\gamma = 1.25$
Focal Tversky	$\alpha = 0.45$, $\gamma = 1.25$
Adaptive	$\alpha = 0.65$, $\gamma = 1.25$
Generalized Dice	$\alpha = 0.6$, $\gamma = 1.25$
Threshold Value	th = 0.5

Comparative Performance Analysis

The performance of each model with different loss functions was evaluated on the test data using standard metrics: Dice Score, Intersection over Union (IoU), Accuracy, Precision, Sensitivity (Recall), Specificity, F1 Score, and Area Under the Curve (AUC). To improve statistical reliability, all models were trained and evaluated using multiple independent runs with different random seeds. We report the mean and standard deviation (mean ± std) of all evaluation metrics across these runs. The performance of each loss function on the datasets is reported in Table 5 for the COVID-SemiSeg dataset and Table 6 for the COVID-19 dataset. The results demonstrate that the effectiveness of the loss functions varies with dataset characteristics and evaluation metrics. For both datasets, the Focal Tversky loss generally achieves better performance, demonstrating its robustness in handling intra- and inter-class variability of lesion areas within and across datasets. The results also show that no single loss function consistently outperforms the others across all evaluation metrics. However, imbalance-aware and region-based loss functions such as WBCE, Tversky, Focal, Focal Tversky, and Dice-based hybrid losses generally provide more robust segmentation performance under varying imbalance ratios and lesion characteristics. Figures 4 and 5 support these findings. The ground truth infected regions were annotated by medical experts and are presented in the final column. The intermediate columns display the predicted infected regions obtained using various loss functions. This comparative analysis demonstrates the effectiveness of different loss functions in identifying infected regions.

Table 5. Performance Comparison Of Loss Functions On The COVID-SemiSeg Dataset. Results Are Reported As Mean ± Standard Deviation Across Multiple Independent Runs With Different Random Seeds

Metric	BCE	Dice	WBCE	Focal	Tversky	Focal Tversky	Dice + Focal	Dice + WBCE
Dice Score	0.745± 0.021	0.766± 0.014	0.771± 0.006	0.718± 0.014	0.740± 0.009	0.766± 0.020	0.762± 0.010	0.765± 0.010
IOU	0.594± 0.026	0.621± 0.018	0.627± 0.008	0.557± 0.018	0.590± 0.016	0.616± 0.020	0.616± 0.013	0.619± 0.013
Accuracy	0.924± 0.007	0.930± 0.005	0.927± 0.003	0.915± 0.004	0.927± 0.002	0.929± 0.005	0.928± 0.004	0.928± 0.003
Precision	0.811± 0.029	0.825± 0.021	0.772± 0.015	0.777± 0.017	0.860± 0.008	0.822± 0.021	0.806± 0.029	0.804± 0.017
Sensitivity	0.689± 0.015	0.716± 0.008	0.711± 0.020	0.663± 0.020	0.650± 0.011	0.770± 0.003	0.724± 0.010	0.729± 0.011
Specificity	0.969± 0.005	0.971± 0.004	0.959± 0.009	0.964± 0.004	0.983± 0.006	0.971± 0.004	0.967± 0.007	0.966± 0.004
F1	0.745± 0.021	0.766± 0.014	0.771± 0.006	0.718± 0.014	0.740± 0.009	0.766± 0.020	0.762± 0.010	0.765± 0.010
AUC	0.887± 0.013	0.850± 0.006	0.911± 0.007	0.937± 0.006	0.822± 0.005	0.854± 0.006	0.879± 0.006	0.881± 0.008

Table 6. Performance Comparison Of Loss Functions On The COVID-19 Dataset. Results Are Reported As Mean ± Standard Deviation Across Multiple Independent Runs With Different Random Seeds

Metric	BCE	Dice	WBCE	Focal	Tversky	Focal Tversky	Dice + Focal	Dice + WBCE
Dice Score	0.505 ±0.004	0.616± 0.023	0.613± 0.005	0.581± 0.019	0.582± 0.019	0.617± 0.012	0.624± 0.032	0.605± 0.016
IOU	0.340± 0.003	0.448± 0.024	0.444± 0.005	0.456± 0.015	0.413± 0.019	0.455± 0.024	0.453± 0.028	0.437± 0.017
Accuracy	0.992± 0.001	0.993± 0.001	0.993± 0.001	0.991± 0.001	0.993± 0.001	0.993± 0.001	0.993± 0.001	0.993± 0.001
Precision	0.752± 0.038	0.759± 0.018	0.717± 0.026	0.734± 0.002	0.815± 0.017	0.745± 0.005	0.759± 0.009	0.760± 0.006
Sensitivity	0.385± 0.010	0.524± 0.037	0.541± 0.010	0.575± 0.006	0.458± 0.029	0.573± 0.019	0.531± 0.040	0.508± 0.023
Specificity	0.999± 0.001	0.998± 0.001	0.997± 0.001	0.997± 0.001	0.999± 0.001	0.999± 0.001	0.998± 0.001	0.998± 0.001
F1	0.505± 0.004	0.616± 0.023	0.613± 0.005	0.581± 0.019	0.582± 0.020	0.617± 0.012	0.624± 0.022	0.605± 0.016
AUC	0.752± 0.004	0.769± 0.018	0.819± 0.009	0.747± 0.007	0.736± 0.014	0.796± 0.008	0.793± 0.024	0.793± 0.024

In addition to segmentation performance, we also evaluated the computational cost of different loss functions. The reported values are presented as mean ± standard deviation across multiple independent runs with different random seeds. Tables 7 and 8 present the average training time per epoch and inference time per image. The results indicate that the differences in computational cost across loss functions are relatively limited. On both datasets, more computationally complex loss functions such as Focal Loss, Focal Tversky Loss, and Dice + Focal slightly increased the training time due to their additional weighting and modulation operations during optimization. The inference time per image remained highly similar across all loss functions, as they do not affect the forward pass of the network.

CONCLUSIONS

This study systematically evaluated the effectiveness of widely used loss functions in addressing inherent class imbalance and lesion variability in COVID-19 lesion segmentation using CT images. Lesion segmentation presents significant challenges due to the significantly smaller size of anomalous tissue compared to healthy tissue, which leads models to be biased toward the majority class. In addition to foreground–background imbalance, lesion regions also exhibit substantial intra-class and inter-class variability in terms of size, texture, boundary sharpness, and

intensity distribution, further complicating the segmentation process. These challenges highlight the importance of selecting and optimizing appropriate loss functions in image segmentation tasks. In this work, we provided a taxonomy of commonly used loss functions and analyzed their strengths and limitations for lesion segmentation under imbalanced conditions. Unlike conventional studies that mainly focus on foreground–background imbalance, this study explicitly incorporated both inter-class imbalance and intra-class heterogeneity into the evaluation framework. By integrating quantitative measures such as entropy, lesion area, gradient magnitude, and intensity differences, the study provided a more comprehensive analysis of lesion structural complexity within and across datasets.

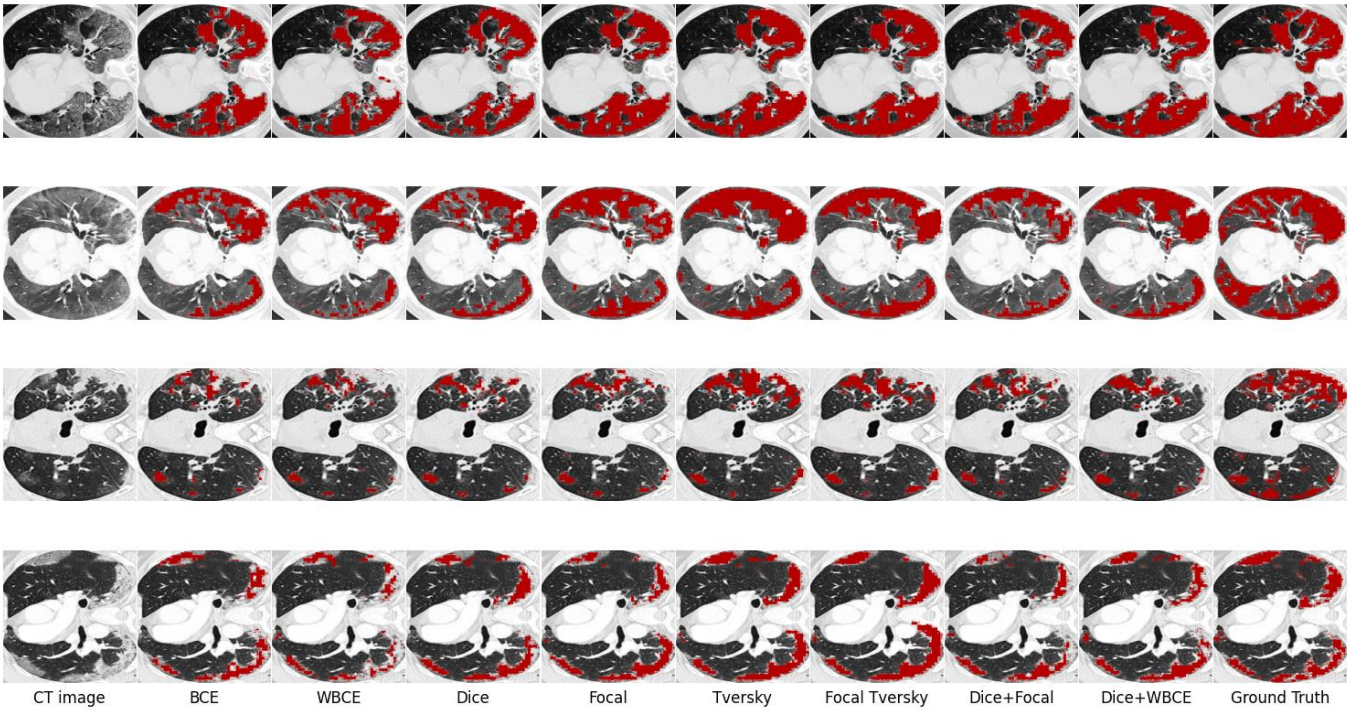

Figure 4. Comparison Of CT Scans From Four Different Patients (First Column) From COVID-SemiSeg Dataset With Predicted Infected Regions Using Various Loss Functions (Middle Columns), And Ground Truth (Last Column)

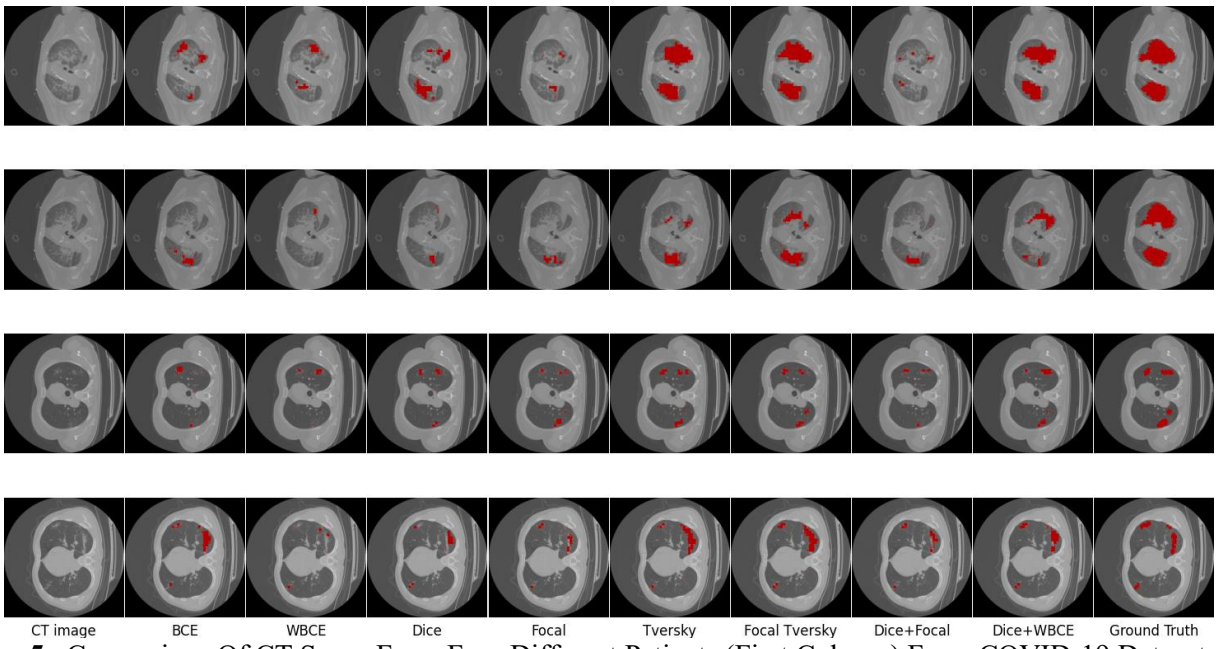

Figure 5. Comparison Of CT Scans From Four Different Patients (First Column) From COVID-19 Dataset, With Predicted Infected Regions Using Various Loss Functions (Middle Columns), And Ground Truth (Last Column)

Table 7. Average Training And Inference Time Comparison Across Different Loss Functions In COVID-SemiSeg Dataset. Results Are Reported As Mean ± Standard Deviation Across Multiple Independent Runs With Different Random Seeds

	BCE	Dice	WBCE	Focal	Tversky	Focal Tversky	Dice + Focal	Dice + WBCE
Training Time (second /epoch)	2.870 ± 0.021	2.881 ± 0.003	2.887 ± 0.003	3.468 ± 0.007	2.888 ± 0.005	2.900 ± 0.006	3.531 ± 0.068	2.924 ± 0.006
Inference Time (second/image)	0.001482 ± 0.000016	0.001543 ± 0.000029	0.001520 ± 0.000036	0.001568 ± 0.000047	0.001554 ± 0.000034	0.001478 ± 0.000024	0.001550 ± 0.000015	0.001568 ± 0.000005

Table 8. Average Training And Inference Time Comparison Across Different Loss Functions in COVID-19 Dataset. Results Are Reported As Mean ± Standard Deviation Across Multiple Independent Runs With Different Random Seeds

	BCE	Dice	WBCE	Focal	Tversky	Focal Tversky	Dice + Focal	Dice + WBCE
Training Time (second /epoch)	6.664 ± 0.069	6.721 ± 0.026	6.737 ± 0.051	7.926 ± 0.006	6.718 ± 0.037	6.881 ± 0.034	8.224 ± 0.015	6.742 ± 0.036
Inference Time (second/image)	0.002204 ± 0.000297	0.002414 ± 0.000018	0.002465 ± 0.000020	0.002493 ± 0.000070	0.002281 ± 0.000166	0.002284 ± 0.000250	0.002473 ± 0.000044	0.002517 ± 0.000069

All models were evaluated using multiple independent runs with different random seeds, and the results were reported as mean ± standard deviation. Experimental results on two publicly available datasets demonstrated that the effectiveness of loss functions varies depending on the dataset characteristics and evaluation metrics. The findings showed that no single loss function consistently outperformed the others across all metrics. However, imbalance-aware and region-based loss functions such as WBCE, Tversky, Focal, Focal Tversky, and Dice-based hybrid losses generally provided more robust performance under varying imbalance ratios and lesion characteristics. The Focal Tversky Loss generally outperforms other methods, highlighting its ability to simultaneously address class imbalance and emphasize hard-to-segment regions. In addition to segmentation performance, we also analyzed the computational cost of different loss functions by reporting average training and inference times. The results showed that more complex loss functions slightly increased training time due to additional weighting and modulation operations. However, inference times remained similar across all methods, indicating that the choice of loss function has minimal impact on deployment-time efficiency.

This study highlights the importance of carefully choosing and optimizing loss functions to align with the specific characteristics of the dataset and the nature of the lesions being segmented. Effectively addressing the inherent class imbalance between lesion and non-lesion areas, as well as lesion variability, contributes to the development of more accurate and reliable lesion segmentation algorithms.

ACKNOWLEDGEMENTS

This research is supported by the TUBITAK-BIDEB 2232 International Fellowship program under the grant number 121C085.

CONFLICT OF INTEREST

The author(s) stated that there are no conflicts of interest regarding the publication of this article.

Artıfıcıal Intellıgence Contrıbutıon Statement

This manuscript was entirely written, edited, analyzed, and prepared without the assistance of any artificial intelligence tools. All content, including text, data analysis, and figures, was solely generated by the authors.

REFERENCES

Abraham, N., & Khan, N. M. (2019, April). A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI) (pp. 683–687). IEEE. https://doi.org/10.1109/ISBI.2019.8759329

American College of Radiology. (2020). ACR recommendations for the use of chest radiography and computed tomography for suspected COVID-19 infection. Retrieved March 28, 2026, from American College of Radiology

An, P., Xu, S., Harmon, S. A., Turkbey, E. B., Sanford, T. H., Amalou, A., Kassin, M., Varble, N., Blain, M., Anderson, V., & Patella, F. (2020). CT images in COVID-19 dataset. Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2020.GQRY-NC81

Azad, R., Heidary, M., Yilmaz, K., Hüttemann, M., Karimijafarbigloo, S., Wu, Y., Schmeink, A., & Merhof, D. (2023). Loss functions in the era of semantic segmentation: A survey and outlook. arXiv:2312.05391. https://arxiv.org/abs/2312.05391

Cohen, J. P., Morrison, P., & Dao, L. (2020). COVID-19 image data collection. arXiv:2003.11597. https://arxiv.org/abs/2003.11597

Fan, D.-P., Zhou, T., Ji, G.-P., Zhou, Y., Chen, G., Fu, H., Shen, J., & Shao, L. (2020). Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. IEEE Transactions on Medical Imaging, 39, 2626–2637. https://doi.org/10.1109/TMI.2020.2996645

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media. https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/

Hashemi, S. R., Salehi, S. S., Erdogmus, D., Prabhu, S. P., Warfield, S. K., & Gholipour, A. (2018). Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access, 7, 1721–1735. https://doi.org/10.1109/ACCESS.2018.2886371

Jadon, S. (2020). A survey of loss functions for semantic segmentation. In 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (pp. 1–7). IEEE. https://doi.org/10.1109/CIBCB48159.2020.9277638

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6, 1–54. https://doi.org/10.1186/s40537-019-0192-5

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., & Cao, K. (2020). Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy. Radiology, 296, E65–E71. https://doi.org/10.1148/radiol.2020200905

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017, October). Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 2980–2988). IEEE. https://doi.org/10.1109/ICCV.2017.324

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., & Martel, A. (2021). Loss odyssey in medical image segmentation. Medical Image Analysis, 71, 102035. https://doi.org/10.1016/j.media.2021.102035

Nguyen, Q., & Thai, H. (2023). Crack segmentation of imbalanced data: The role of loss functions. Engineering Structures, 297, 116988. https://doi.org/10.1016/j.engstruct.2023.116988

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation.

In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (Vol. 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Salehi, S. S., Erdogmus, D., & Gholipour, A. (2017). Tversky loss function for image segmentation using 3D fully convolutional deep networks. In the International Workshop on Machine Learning in Medical Imaging (MLMI). Springer. https://doi.org/10.1007/978-3-319-67389-9_44

Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z., Shen, D., & Shi, Y. (2021). Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction. Medical Physics, 48(4), 1633–1645. https://doi.org/10.1002/mp.14609

Stockman, G., & Shapiro, L. (2001). Computer Vision. Prentice Hall, pp.305-351.

Sun, D., Li, X., Guo, D., Wu, L., Chen, T., Fang, Z., Chen, L., Zeng, W., & Yang, R. (2020). CT quantitative analysis and its relationship with clinical features for assessing the severity of patients with COVID-19. Korean Journal of Radiology, 21(7), 859–868. https://doi.org/10.3348/kjr.2020.0293

Taghanaki, S. A., Zheng, Y., Zhou, S., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., & Hamarneh, G. (2019). Combo loss: Handling input and output imbalance in multi-organ segmentation. Computerized Medical Imaging and Graphics, 75, 24–33. https://doi.org/10.1016/j.compmedimag.2019.04.005

Tang, Z., Zhao, W., Xie, X., Zhong, Z., Shi, F., Ma, T., Liu, J., & Shen, D. (2021). Severity assessment of COVID-19 using CT image features and laboratory indices. Physics in Medicine & Biology, 66, 035015. https://doi.org/10.1088/1361-6560/abbf9e

Zhu, W., Huang, Y., Zeng, L., Chen, X., Liu, Y., Qian, Z., Du, N., Fan, W., & Xie, X. (2019). AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. Medical Physics, 46, 576–589. https://doi.org/10.1002/mp.13300

Zhu, X., Men, J., Yang, L., & Li, K. (2022). Imbalanced driving scene recognition with class focal loss and data augmentation. International Journal of Machine Learning and Cybernetics, 13(6). https://doi.org/10.1007/s13042-022-01575-x