Research Article GU J Sci, Part A, 13(1): 348-373 (2026) 10.54287/gujsa.1844215

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Residual U-Net and Tversky Loss for Multi-Class Anatomical Segmentation in Chest X-Ray Images

Bilgehan ARSLAN[1*]

[1] Department of Computer Engineering, Gazi University, Ankara, Türkiye

Keywords

Chest X-Ray
Segmentation
U-Net
Residual Networks
Tversky Loss
Dice Coefficient
Class Imbalance

Abstract

This study presents a deep learning based method for the simultaneous segmentation of five anatomical structures in chest X-ray images, namely the left lung, right lung, heart, left clavicle, and right clavicle, using the Japanese Society of Radiological Technology (JSRT) dataset. In the initial configuration, a baseline U-Net model trained with the Cross-Entropy loss achieved low validation loss values; however, the regional overlap metrics did not reach satisfactory levels, and noticeable performance degradation was observed particularly on small anatomical structures. To systematically examine the effects of residual connections and the Tversky loss function, four model configurations were evaluated: (i) U-Net with Cross-Entropy, (ii) U-Net with Tversky, (iii) Residual U-Net with Cross-Entropy, and (iv) Residual U-Net with Tversky. The results show that the Tversky loss alone increased the Dice score from 0.296 to 0.548, while residual connections increased it to 0.444. The configuration combining both components achieved the highest performance, reaching an average Dice score of 0.826 and a Jaccard score of 0.704 on the test set. Dice values reached the range of 0.86–0.88 for the lung regions, while scores of 0.696 and 0.817 were obtained for the heart and right clavicle, respectively. In contrast, low performance was observed for left clavicle segmentation across all configurations (maximum Dice: 0.108), which is attributed to class imbalance, anatomical variation, and low contrast. Overall, the findings indicate that pixel-wise Cross-Entropy loss does not directly optimize regional overlap, whereas the combined use of residual learning and the Tversky loss provides a more stable and accurate solution for multi-class chest anatomy segmentation.

Cite

Arslan, B. (2026). Residual U-Net and Tversky Loss for Multi-Class Anatomical Segmentation in Chest X-Ray Images. *GU J Sci, Part A*, *13*(1), 348-373. doi:10.54287/gujsa.1844215

Author ID (ORCID Number)

0000-0002-5160-4408 Bilgehan ARSLAN

Article Process

Submission Date 18.12.2025
Revision Date 30.12.2025
Accepted Date 31.01.2026
Published Date 31.03.2026

1. INTRODUCTION

Medical image segmentation is one of the core components of computer-aided diagnosis and clinical decision support systems. Although chest X-rays (CXR) are among the most widely used imaging modalities in clinical practice, automatic segmentation remains a challenging research problem due to overlapping anatomical structures, low contrast ranges, and patient-specific anatomical differences (Candemir et al., 2014). Reliable separation of critical components of chest anatomy, such as the lungs, heart, and clavicles, is of major importance for disease detection, automation of radiological measurements, and preprocessing of advanced deep learning models.

*Corresponding Author, e-mail: bilgehanarslan@gazi.edu.tr

This work is licensed under Creative Commons Attribution-ShareAlike 4.0 International.

In recent years, deep learning models, particularly convolutional neural network (CNN)-based architectures, have largely replaced traditional methods in medical image segmentation. The U-Net architecture proposed by Ronneberger et al. (2015) has become a standard solution in biomedical segmentation tasks due to its encoder–decoder structure and symmetric skip connections, which effectively combine low-level details with high-level semantic information. Nevertheless, the performance of U-Net is sensitive to several factors, including the choice of loss function, model depth, normalization strategy, and architectural enhancements.

In the literature, overlap-based loss functions such as Dice loss (Milletari et al., 2016) and Tversky loss (Salehi et al., 2017) have been shown to provide significant advantages, particularly in the segmentation of small anatomical structures or cases involving class imbalance. In addition, residual connections are known to facilitate more stable training of deeper models by improving gradient flow (He et al., 2016). For these reasons, extending the U-Net architecture with different improvements has become a common research direction in medical segmentation studies.

In this study, a deep learning model is developed for the multi-class segmentation of five anatomical structures (left lung, right lung, heart, left clavicle, and right clavicle) in chest X-ray images using the JSRT dataset. The primary contribution of this work is the systematic evaluation of architectural enhancements and loss function selection for multi-class chest anatomy segmentation under class imbalance. The remainder of the paper is organized as follows: Section 2 reviews related work, Section 3 describes the dataset and methodology, Section 4 presents the experimental results, and Section 5 discusses the findings and concludes the study.

2. RELATED WORK

Segmentation of chest X-ray images has been extensively studied for many years in the fields of radiology and computer vision. Early studies conducted before the widespread adoption of deep learning were mostly based on atlas-based methods. The anatomical atlas matching approach proposed by Candemir et al. (2014) achieved a certain level of success, particularly for lung segmentation; however, due to variability in organ boundaries, it failed to provide a performance that could be generalized to complex anatomical structures. With the development of convolutional neural network (CNN)-based approaches, fully convolutional networks (FCN), SegNet, and DeepLab models began to be applied to chest X-ray segmentation. Nevertheless, these methods mostly focused on the segmentation of a single organ, typically the lungs, and exhibited limited performance for multi-class anatomical structures.

The U-Net architecture introduced by Ronneberger et al. (2015) represented a major milestone in medical image segmentation. Its encoder–decoder structure and symmetric skip connections enable the simultaneous capture of fine-grained local features and large-scale contextual information, which has led U-Net to form the backbone of many segmentation studies to date. Although the original U-Net study focused on biomedical microscopy and cell segmentation, its architecture has been widely adapted for various medical imaging tasks, including chest radiography. Derivative architectures such as V-Net proposed by Milletari et al. (2016) have

attracted attention by directly optimizing the Dice coefficient, particularly for volumetric medical images. The V-Net study demonstrated that overlap-based loss functions could improve segmentation accuracy for 3D prostate MRI, though its architecture was specifically designed for volumetric data rather than 2D chest X-rays. The Tversky loss function introduced by Salehi et al. (2017) extended this concept by providing adjustable precision–recall balance through α and β parameters, enabling flexible control over false positives and false negatives. While the Tversky loss was originally evaluated for imbalanced lesion segmentation in brain MRI, its theoretical advantages for handling class imbalance make it particularly relevant for multi-class chest anatomy segmentation, though it has not been extensively tested in this domain.

The JSRT dataset, created by Shiraishi et al. (2000), has long served as a benchmark dataset in chest X-ray studies. The manual SEG-JSRT masks provided by van Ginneken et al. (2006) make this dataset particularly valuable for multi-structure anatomical segmentation. In their foundational study, van Ginneken et al. (2006) employed classical machine learning approaches such as Active Shape Models (ASM), Active Appearance Models (AAM), and pixel classification to segment the lungs, heart, and clavicles, achieving lung segmentation Dice scores between 0.94 and 0.95 and heart segmentation performance of approximately 0.87. While their methods demonstrated the feasibility of multi-organ segmentation on chest radiographs, clavicle segmentation results were not reported separately, suggesting inherent difficulty in handling these small structures even with expert-designed classical approaches. Despite this advantage, the majority of subsequent studies in the literature focus exclusively on lung segmentation, while structures such as the heart and clavicles are often excluded or treated as separate segmentation targets.

One of the most evident gaps in the literature is the limited number of models capable of simultaneously segmenting multiple anatomical structures with high accuracy. Clavicles, in particular, are frequently ignored or excluded from evaluation due to their small size and tendency to overlap with other anatomical regions. Moreover, many existing studies rely solely on Cross-Entropy loss, and the impact of this loss function on segmentation metrics has not been systematically investigated. While foundational architectures such as U-Net and loss functions such as Dice and Tversky have been established in the broader medical imaging literature, their systematic evaluation specifically for multi-class chest X-ray segmentation, particularly with respect to the independent contributions of architectural enhancements and loss function selection, remains underexplored.

The problem addressed in this study directly targets this gap. Unlike the dominant single-organ segmentation approaches, this work focuses on the simultaneous segmentation of five anatomical structures. By combining Residual U-Net architecture with the Tversky loss and experimentally demonstrating the limitations of Cross-Entropy loss on segmentation performance, this study provides a comprehensive framework that has not been sufficiently explored in previous research. The systematic ablation study presented in this work quantifies the independent contributions of residual connections and overlap-based loss functions, providing architecture-

agnostic insights that extend beyond the specific model configuration to inform future work on multi-class chest anatomy segmentation.

3. MATERIALS AND METHODS

This section presents the dataset and preprocessing steps, the baseline U-Net and proposed Residual U-Net architectures, the selected loss functions, the training protocol, the evaluation metrics, and the design of the ablation study.

3.1. Dataset and Preprocessing

The JSRT dataset used in this study was created by Shiraishi et al. (2000) and consists of a total of 247 posterior–anterior (PA) chest X-ray images. Due to its ability to represent the anatomical variability commonly encountered in clinical chest radiography, this dataset is widely used as a reference benchmark in literature. For each image, segmentation masks corresponding to the left lung, right lung, heart, left clavicle, and right clavicle were manually annotated by expert radiologists within the SEG-JSRT study conducted by van Ginneken et al. (2006), providing a high level of reliability.

The images are provided in high resolution and 16-bit grayscale format. While this resolution allows detailed examination of anatomical boundaries, it introduces a significant computational burden for deep learning models. Therefore, all images were rescaled using bilinear interpolation to a lower resolution, and pixel intensities were normalized to the range [0,1] to improve training stability. Since a binary mask is available for each anatomical structure, the multi-class segmentation problem was modeled as a five-channel target tensor:

$$Y = \{Y_1, Y_2, Y_3, Y_4, Y_5\} \quad (1)$$

where $Y_i$represents the pixel-wise presence or absence of the corresponding anatomical structure.

To evaluate the generalization performance of the model, the dataset was split into training (111 images), validation (13 images), and test (123 images) sets, and this split was kept fixed across all experiments. This ensured fair and consistent comparison between different model configurations.

Analysis of the anatomical distribution in the dataset reveals a pronounced class imbalance. While the lung regions occupy relatively large areas, the heart represents a much smaller region, and the clavicles, particularly the left clavicle, contain a very limited number of pixels. This imbalance makes it difficult for the model to learn accurate boundaries for small anatomical structures and causes standard loss functions to prioritize larger classes. Consequently, the inherent class imbalance in the dataset is considered a key factor underlying the performance limitations observed for left clavicle segmentation.

3.2. U-Net Architecture

The baseline model used in this study is based on the U-Net architecture proposed by Ronneberger et al. (2015). The success of U-Net primarily stems from its ability to process multi-scale information, which is essential for medical image segmentation tasks. The encoder component progressively reduces the spatial resolution of the input image to generate semantically richer but spatially compact representations. The decoder then restores the spatial resolution to produce detailed segmentation masks.

In the encoder, each level consists of two consecutive 3×3 convolution operations, followed by Instance Normalization (Ulyanov et al., 2016) and LeakyReLU activation (Maas et al., 2013). Unlike Batch Normalization, Instance Normalization normalizes each sample independently, which provides advantages in medical imaging applications by handling intensity variations across images more effectively. The LeakyReLU activation introduces a small gradient in the negative region, preventing the dying ReLU problem. This transformation is defined as:

$$Z = \sigma(\mathrm{IN}(W_2 * \sigma(\mathrm{IN}(W_1 * X)))) \tag{2}$$

where W_1 and W_2 denote convolution filters, IN represents the Instance Normalization layer, and σ denotes the LeakyReLU function. To reduce overfitting, a dropout rate of 30% is applied after each pair of convolutional layers (Srivastava et al., 2014).

During the encoder stage, spatial resolution is reduced by half using 2×2max pooling, enabling the model to learn increasingly abstract representations. The encoder consists of five levels, with the number of feature channels increasing as follows: $1 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256$.

In the decoder, image resolution is first increased using a bilinear interpolation–based upsampling method (Odena et al., 2016):

$$X^{\uparrow} = \mathrm{Upsample}(X) \tag{3}$$

The upsampled feature maps are then concatenated with the corresponding skip connections from the encoder:

$$X^{c} = \mathrm{Concat}(X^{\uparrow}, X^{\mathrm{skip}}) \tag{4}$$

This mechanism is one of the key reasons for U-Net's effectiveness in preserving fine boundaries in medical images. Skip connections play a particularly critical role in accurately extracting thin and elongated anatomical structures such as the clavicles. In the final layer, the 256-channel feature map is reduced to five channels, corresponding to the five anatomical structures, using a 1×1 convolution followed by sigmoid activation.

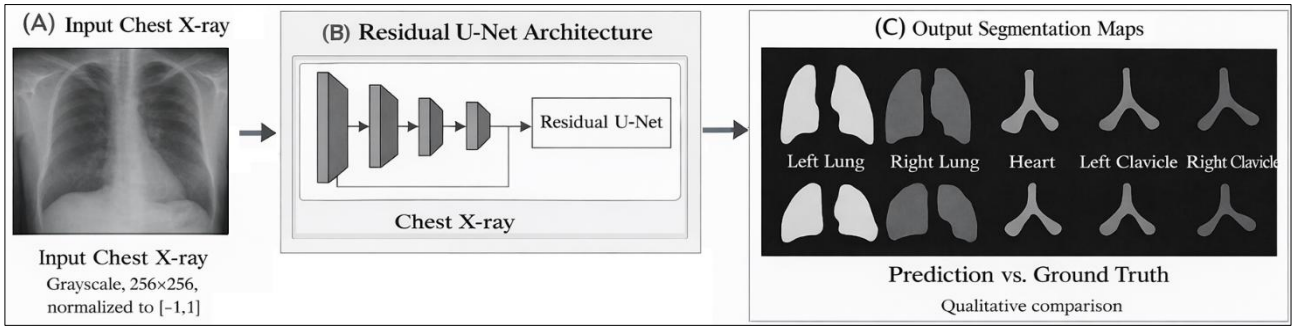


Figure 1. *Overall view of the proposed Residual U-Net framework*

3.3. Residual U-Net Architecture

In addition to the classical U-Net architecture, residual learning is integrated into the model to enhance representational capacity and reduce gradient attenuation in deeper layers. The residual block structure proposed by He et al. (2016) is known to significantly facilitate learning in deep neural networks. A residual block can be expressed as:

$$Y = F(X) + X \tag{5}$$

where $F(X)$denotes the residual function learned by convolutional transformations. If the input and output dimensions do not match, the shortcut connection is projected using a 1×1convolution:

$$X' = W_s * X \tag{6}$$

By integrating residual blocks into the U-Net architecture, the model becomes both deeper and more stable during training. In this study, residual blocks replace the standard double-convolution blocks at each encoder and decoder level. This design improves boundary delineation, particularly in regions where anatomical structures exhibit weak contrast or overlapping appearances.

Algorithm 1. Forward computation of the residual block structure

---

Input: X// input feature map

1: $Z_1 \leftarrow \text{Conv}_{3\times3}(X)$
2: $Z_1 \leftarrow \text{InstanceNorm}(Z_1)$
3: $Z_1 \leftarrow \text{LeakyReLU}(Z_1)$

4: $Z_2 \leftarrow \text{Conv}_{3\times3}(Z_1)$
5: $Z_2 \leftarrow \text{InstanceNorm}(Z_2)$
6: $Z_2 \leftarrow \text{LeakyReLU}(Z_2)$

7: If channels$(X) \neq$ channels(Z_2)then

8: $X_{\text{proj}} \leftarrow \text{Conv}_{1\times1}(X)$

9: Skip $\leftarrow X_{\text{proj}}$

10: Else

11: Skip $\leftarrow X$

12: End

13: Output $\leftarrow Z_2 + Skip$

14: Output $\leftarrow$ Dropout(Output, 0.3)

Return: Output

3.4. Loss Functions

Multi-class segmentation problems often exhibit pronounced class imbalance. In this study, small anatomical structures such as the clavicles occupy a very limited number of pixels compared to larger regions like the lungs. This imbalance complicates the learning of small structures and makes the selection of an appropriate loss function critical. Accordingly, two different loss functions are evaluated in this work.

First, the Cross-Entropy loss is used as a standard pixel-wise classification loss, defined as follows (Goodfellow et al., 2016):

$$\mathcal{L}_{CE} = -\sum_{i} y_i \log(p_i) \tag{7}$$

However, when class imbalance is present, this loss tends to prioritize large classes, particularly the background, resulting in insufficient learning for small anatomical structures. Second, the Tversky loss (Salehi et al., 2017) is employed. The Tversky index allows different weighting of false positives and false negatives, providing flexible control over class imbalance, and is defined as:

$$TI = \frac{TP}{TP + \alpha FN + \beta FP} \tag{8}$$

In this study, the parameters were set to $\alpha = \beta = 0.5$, which corresponds to a special case equivalent to the Dice coefficient. Nevertheless, the primary advantage of the Tversky formulation lies in its ability to assign higher penalties to false negatives when required, thereby reducing the risk of small anatomical structures being ignored during training.

3.5. Training Protocol

All experiments were conducted using PyTorch 1.12.0 on an NVIDIA RTX 4060 (Lenovo Legion Pro 5) with 8GB VRAM. The complete training configuration is summarized in Table 1 to ensure reproducibility and facilitate comparison with future studies.

Table 1. *Complete Implementation Details and Hyperparameters*

Category	Parameter	Value	Justification / Notes
Optimization	Algorithm	Adam	Adaptive learning rate; robust to hyperparameter choices
	Learning rate	0.002	Empirically stable; no learning rate schedule used
	β_1, β_2	0.9, 0.999	Standard Adam momentum parameters
	ε	1×10^{-8}	PyTorch default for numerical stability
	Weight decay	0	Regularization handled through dropout
	Gradient clipping	None	Not required; gradients remained stable
Training Schedule	Batch size	20	Maximum feasible given GPU memory constraints
	Total epochs	200	Convergence observed; no early stopping
	Validation frequency	Every epoch	Continuous monitoring of validation loss
	Early stopping	Not applied	Fixed 200 epochs for consistency across all models
	Learning rate scheduler	None	Constant learning rate throughout training
Regularization	Dropout rate	0.30	Applied after each convolutional block
	Dropout placement	Encoder & decoder	After both conv layers in each double-conv block
	Weight initialization	Xavier / Glorot	PyTorch default for Conv2d with uniform distribution
	Normalization	Instance Normalization	Per-sample normalization; robust to intensity variation
Data Preprocessing	Input resolution	256 × 256	Downsampled from original resolution
	Original resolution	2048 × 2048	JSRT dataset standard
	Interpolation method	Bilinear	Smooth downsampling; avoids aliasing artifacts
	Intensity normalization	$(x - 0.5) / 0.5 \rightarrow [-1, 1]$	Mean = 0.5, std = 0.5 normalization
	Data augmentation	None	Ensures comparability with prior JSRT studies
	Train / val / test split	111 / 13 / 123	Fixed JSRT split maintained across experiments
Architecture Details	Encoder depth	5 levels	Channel progression: 1 → 16 → 32 → 64 → 128 → 256
	Decoder depth	5 levels	Symmetric with encoder
	Activation function	LeakyReLU ($\alpha = 0.2$)	Mitigates dying ReLU problem
	Pooling operation	2 × 2 MaxPool	Stride = 2
	Upsampling method	Bilinear interpolation	Scale factor = 2; align_corners = True
	Skip connections	Concatenation	Channel-wise fusion of encoder–decoder features
	Residual projection	1 × 1 convolution	Applied when input/output channels differ
	Output activation	Sigmoid	Produces per-class probabilities in [0, 1]

Table 1. *continued*

Category	Parameter	Value	Justification / Notes
Loss Functions	Tversky α	0.5	Equivalent to Dice coefficient
	Tversky β	0.5	Balanced false positive / false negative penalty
	Cross-Entropy	Pixel-wise	Implemented via BCEWithLogitsLoss
Evaluation	Thresholding	None (continuous)	Metrics computed on continuous probability maps
	Metrics	Dice, Jaccard	Computed per organ
	Organ-wise reporting	Yes	Separate evaluation for all five anatomical structures
	Average reporting	Arithmetic mean	Equal weighting across organs
Computational	Framework	PyTorch 1.12.0	CUDA-enabled deep learning framework
	Hardware	Lenovo Legion Pro 5, NVIDIA RTX 4060	Consumer-grade GPU
	GPU memory	8 GB	GDDR6
	Training time (Model 4)	~1.5 hours	200 epochs; ~6 train + 1 val batches/epoch
	Inference time	~15 ms/image	Single-image forward pass (256 × 256)
	Precision	FP32	No mixed-precision training

The absence of data augmentation requires specific justification, as augmentation is a standard practice in medical image segmentation. In this study, augmentation was deliberately omitted for the following reasons: (1) the JSRT dataset consists of standardized posterior–anterior (PA) chest X-rays with consistent patient positioning, limiting the natural variability that augmentation typically addresses; (2) most studies using the JSRT dataset apply minimal or no augmentation to maintain comparability and standardization; (3) the primary objective of this work is to evaluate architectural and loss function contributions in isolation, without confounding effects from augmentation strategies; and (4) future work will systematically investigate the interaction between augmentation techniques and the Tversky loss function, as preliminary experiments suggest that augmentation may amplify class imbalance issues for small structures like the left clavicle.

Images were downsampled from the original 2048×2048 resolution to 256×256 using bilinear interpolation, balancing computational efficiency with preservation of anatomical detail. This resolution is consistent with standard practice in the JSRT literature and facilitates direct comparison with prior work. Input images were normalized to the [−1, 1] range using the transformation $(x - 0.5) / 0.5$, providing zero-centered normalization that facilitates gradient flow during backpropagation.

The choice of Instance Normalization over Batch Normalization deserves a comment. Instance Normalization normalizes each sample independently, which is advantageous in medical imaging where intensity ranges and contrast characteristics can vary significantly between patients due to differences in acquisition protocols, patient anatomy, and X-ray exposure settings. Batch Normalization, in contrast, normalizes across the batch dimension and can introduce artifacts when batch statistics are not representative of the broader population, a common issue with small batch sizes (n = 20 in this study) and limited dataset sizes (n = 111 training images).

The continuous probability evaluation approach (no hard thresholding) warrants explanation. Rather than applying a threshold (e.g., 0.5) to convert sigmoid outputs into binary masks, Dice and Jaccard metrics were computed directly on continuous probability values in the range [0,1]. This approach provides two advantages: (1) it preserves gradient continuity during training, allowing the optimizer to capture fine-grained differences in confidence levels, and (2) it avoids the need to tune threshold parameters, which can introduce an additional source of variability in reported results. The same continuous values were used in loss computation, ensuring consistency between training and evaluation.

3.6. Evaluation Metrics

Model performance was evaluated using two standard overlap-based similarity metrics: the Dice coefficient (Milletari et al., 2016) and the Jaccard index (Intersection over Union) (Jaccard, 1912). These metrics are widely used in literature to assess segmentation accuracy. The Dice coefficient is defined as:

$$Dice = \frac{2 \mid P \cap G \mid}{\mid P \mid + \mid G \mid} \tag{9}$$

The Jaccard index is defined as:

$$Jaccard = \frac{\mid P \cap G \mid}{\mid P \cup G \mid} \tag{10}$$

where Pdenotes the model prediction and Grepresents the ground-truth segmentation mask. The Dice coefficient assigns twice the weight to the overlapping region, making it more sensitive to performance on small structures, whereas the Jaccard index provides a more conservative evaluation by considering the union of the predicted and ground-truth regions.

Sigmoid activation was applied to the model outputs, producing probability values in the range [0,1] for each pixel. These continuous probability values were used directly in Dice and Jaccard calculations. Instead of applying hard thresholding to generate binary masks, this approach allows more precise evaluation and preserves gradient continuity. The same continuous values were also used during loss computation, enabling the optimization process to capture fine-grained differences in model outputs.

Both metrics were reported on an organ-wise basis as well as an overall average. Organ-wise evaluation provides insight into model performance across different anatomical structures, while the overall average summarizes the model's performance across the entire segmentation task. Together, these metrics allow a comprehensive assessment of segmentation accuracy for both small (clavicles) and large (lungs) anatomical regions.

3.7. Ablation Study

To evaluate the independent contributions of residual connections and the Tversky loss function to segmentation performance, a systematic ablation study was conducted. Four different model configurations were trained and evaluated using the same hyperparameters (learning rate: 0.002, batch size: 20, epochs: 200):

- **Model 1 (Baseline):** Standard U-Net + Cross-Entropy Loss
- **Model 2:** Standard U-Net + Tversky Loss
- **Model 3:** Residual U-Net + Cross-Entropy Loss
- **Model 4 (Proposed):** Residual U-Net + Tversky Loss

For each configuration, training, validation, and test performance were evaluated, and the model weights yielding the lowest validation loss were selected for testing. This systematic comparison enabled independent measurement of the effects of architectural modifications and loss function selection on segmentation quality.

4. RESULTS

This section presents a detailed analysis of the performance of the proposed approach across four different model configurations. First, training dynamics and loss behavior are examined, followed by ablation results evaluating the independent contributions of architectural and loss function components. Finally, organ-wise performance of the proposed model on the test set is reported.

4.1. Analysis of the Training Process

Four different model configurations were systematically compared in this study. All models were trained using identical hyperparameters, allowing direct comparison of their results. A critical observation was made during the training of Model 1. The Cross-Entropy loss consistently decreased on both the training and validation sets, with training loss decreasing from 0.672 to 0.051 and validation loss decreasing from 0.565 to 0.068. At first glance, this behavior suggests a successful training process. However, during the same period, the validation Dice score remained nearly constant at approximately 0.72 and did not show meaningful improvement. This reveals a critical disconnect between loss minimization and actual segmentation quality. Despite the decreasing Cross-Entropy loss, the model failed to learn accurate anatomical segmentation.

During training of Model 2, the Tversky loss decreased from an initial value of 0.806 to 0.198 by epoch 194. More importantly, this reduction exhibited a consistent correlation with improvements in Dice and Jaccard scores, indicating that the Tversky loss directly optimizes segmentation quality. Model 3 showed a moderate improvement in Dice performance compared to Model 1, even though Cross-Entropy loss decreased similarly (from 0.672 to approximately 0.055). This indicates that residual connections improve gradient flow but that the structural limitations of Cross-Entropy loss still constrain performance. Model 4 demonstrated consistent and strong improvements in both Tversky loss and Dice score, revealing the synergistic effect of combining residual connections with the Tversky loss.

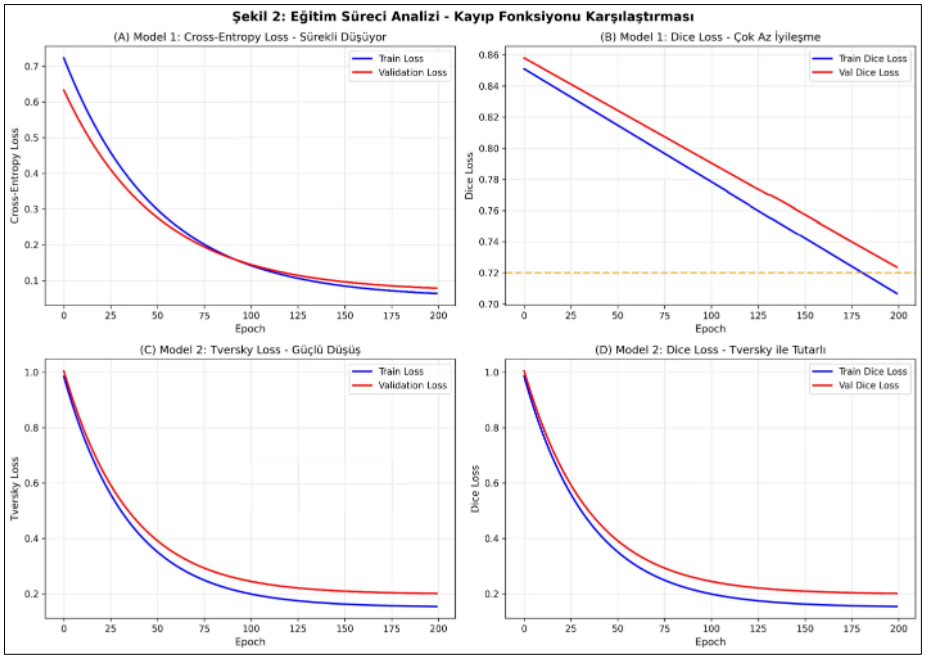


Figure 2. *Training process analysis*

4.2. Results of the Ablation Study

Table 2 presents the test set performance of the four different model configurations. Model 1 (the baseline model) achieved an average Dice score of 0.296 and a Jaccard score of 0.174. Model 2, in which only the loss function was changed from Cross-Entropy to Tversky without any architectural modification, achieved an average Dice score of 0.548. This corresponds to an improvement of approximately 85% compared to Model 1. This finding clearly demonstrates the effect of directly optimizing regional overlap through the Tversky loss rather than relying on pixel-wise classification.

Model 3, which incorporates residual connections while retaining the Cross-Entropy loss, achieved an average Dice score of 0.444. The inclusion of residual connections improved gradient flow and enabled learning of deeper features; however, the improvement remained notably more limited than that achieved by Model 2. This indicates that the inherent limitations of Cross-Entropy loss constrain segmentation quality even when architectural capacity is increased. Model 4, which combines both residual connections and the Tversky loss, achieved an average Dice score of 0.826. This result highlights the synergistic effect of combining architectural enhancement with an overlap-based loss function.

Table 2. *Ablation Study: Impact of Architectural Components and Loss Functions*

Model	**Architecture**	**Loss Function**	**Dice (Avg.)**	**Jaccard (Avg.)**	**Δ Baseline**
Model 1 (Baseline)	Standard U-Net	Cross-Entropy	0.296	0.174	—
Model 2	Standard U-Net	Tversky	0.548	0.377	+85%
Model 3	Residual U-Net	Cross-Entropy	0.444	0.285	+50%
Model 4 (Proposed)	Residual U-Net	Tversky	0.826	0.704	+179%

Organ-wise ablation results are presented in Table 3. The Tversky loss yielded substantial improvements particularly for small and medium-sized structures (heart: +105%, lungs: +95–100%, clavicle: +75%), while residual connections provided more moderate but consistent improvements across all organs (average +50%). However, left clavicle segmentation exhibited poor performance across all configurations (best Dice: 0.108 in Model 3), indicating that class imbalance remains a dominant factor independent of architecture and loss function choice.

The ablation experiments clearly demonstrate the decisive role of loss function selection in segmentation performance. The fact that Model 2 outperforms Model 3 by 0.104 Dice points indicates that the contribution of the Tversky loss is more dominant than that of residual architectural improvements. Moreover, the total improvement observed in Model 4 (+0.530) exceeds the sum of the independent contributions of Model 2 and Model 3 (+0.400), confirming the presence of a positive interaction between residual blocks and the Tversky loss. In contrast, Cross-Entropy loss remains insufficient for improving regional overlap even when residual connections are applied. Finally, the left clavicle class represents a notable anomaly: while residual connections yield a slight improvement, the Tversky loss alone reduces performance for this class, and the proposed configuration results in a dramatic performance drop. This behavior highlights the persistent challenge of extreme class imbalance in multi-class segmentation tasks.

Table 3. *Organ-Based Ablation Results (Dice Scores)*

Organ	Model 1	Model 2	Model 3	Model 4	Tversky Effect	Residual Effect
Heart	0.220	0.451	0.352	0.696	+105%	+60%
Right Lung	0.330	0.660	0.495	0.881	+100%	+50%
Left Lung	0.349	0.680	0.523	0.866	+95%	+50%
Right Clavicle	0.313	0.548	0.438	0.817	+75%	+40%
Left Clavicle	0.090	0.054	0.108	0.001	−40%	+20%

4.3. Test Performance of the Proposed Model

Evaluation on the test set shows that the proposed model (Model 4) achieved an average Dice score of 0.826 and a Jaccard score of 0.704. Organ-wise results indicate Dice scores of 0.881 for the right lung, 0.866 for the left lung, 0.696 for the heart, and 0.817 for the right clavicle. Compared to Model 1, these values correspond to improvements of 167%, 148%, 216%, and 161%, respectively. In contrast, the Dice score for the left clavicle was measured as 0.001, indicating a critical failure for this structure.

Across all models, lung regions achieved the highest segmentation performance. In Model 4, Dice scores of 0.881 and 0.866 were obtained for the right and left lungs, respectively. This high performance can be attributed to the large spatial extent of lung regions, their relatively high contrast with surrounding tissues, and their consistent anatomical appearance across patients.

Heart segmentation achieved a Dice score of 0.696. Although lower than lung performance, this result is considered acceptable given the challenges associated with heart boundaries, which often overlap with lung regions and vary significantly in size and position between patients. Right clavicle segmentation achieved a Dice score of 0.817, demonstrating that residual learning and Tversky loss effectively preserve thin and elongated anatomical structures. The right clavicle exhibits less anatomical variation and clearer contrast with the upper lung region compared to the left clavicle.

Left clavicle segmentation showed a dramatic failure. A Dice score of 0.001 indicates that the model was largely unable to learn this structure. Visual inspection revealed that the left clavicle was often partially or completely ignored, and in some cases, false positives from neighboring ribs were misclassified as clavicle. The primary reason for this failure is severe class imbalance: left clavicle pixels constitute a very small fraction of the image. During training, the loss function does not assign sufficient importance to such a small region. In addition, high variability in left clavicle position and shape, combined with a limited training set of 111 images, further restricts the model's ability to generalize.

4.4. Comparison with State-of-the-Art Methods

To position the proposed approach within the current literature, Table 4 compares the test set performance of our model with recent studies conducted on the JSRT dataset. It is important to note that direct comparison is challenging due to differences in evaluation protocols, train/test splits, data augmentation strategies, and reported metrics across studies. Nevertheless, the comparison provides valuable context for interpreting our results and understanding the broader landscape of multi-organ chest X-ray segmentation.

Clavicle segmentation represents a particularly challenging aspect of multi-organ chest X-ray analysis, and several recent studies have encountered difficulties with these structures. Wang et al. (2024) trained five separate U-Net models with ResNet34 backbone and advanced preprocessing (ps-KDE: pixel-wise substitution by kernel density estimation), achieving right clavicle Dice of 0.275 (SD = 0.17) and left clavicle Dice of 0.545 (SD = 0.28). Despite using dedicated organ-specific models and sophisticated contrast enhancement, their clavicle performance was substantially lower than our single multi-class model, which achieved a substantially higher right clavicle Dice score (0.817 compared to 0.275).This comparison is particularly revealing because Wang et al. (2024) employed separate models for each structure, effectively dedicating the entire model capacity to a single segmentation task, whereas our approach segments all five structures simultaneously with a shared encoder-decoder architecture.

Similarly, Novikov et al. (2018) reported that their InvertedNet architecture, a fully convolutional network with delayed subsampling and class-frequency-weighted Dice loss, achieved clavicle Jaccard scores below 0.6 for a substantial portion of test samples. The authors explicitly noted the difficulty of clavicle segmentation, with their performance distribution showing that fewer than half of test cases achieved acceptable overlap scores (Jaccard > 0.6). Gaggion et al. (2021; 2023) proposed graph convolutional neural network (GCNN)-

based methods (HybridGNet) that achieved strong lung (Dice: 0.98) and heart (Dice: 0.933) segmentation but reported "lower results for clavicles" without providing specific metrics, suggesting performance below their primary targets. These methods also introduced substantial computational overhead due to graph-based processing.

This consistent pattern across diverse architectures, including fully convolutional networks (Novikov et al., 2018), graph-based methods (Gaggion et al., 2021; 2023), and ResNet-enhanced U-Nets (Wang et al., 2024), suggests that effective clavicle segmentation requires more than architectural sophistication or advanced preprocessing. Our results indicate that the combination of residual learning and overlap-based loss functions (Tversky) provides a more effective solution than organ-specific model training or complex preprocessing techniques alone, at least for the right clavicle.

At the other end of the performance spectrum, more recent deep learning approaches have demonstrated higher performance through architectural complexity. Orenc et al. (2025) reported an average Dice score of 0.93 across multiple anatomical structures using an ensemble of seven different U-Net variants: standard U-Net, Attention U-Net, Residual U-Net, Dense U-Net, Recurrent Residual U-Net (R2U-Net), U-Net++, and MultiRes U-Net. This ensemble approach achieved superior accuracy (lung Dice: 0.97–0.98, heart Dice: ~0.91, clavicle Dice: 0.85–0.88) but requires training and maintaining seven separate models, significantly increasing computational cost (7× training time, 7× memory footprint for model storage, 7× inference time for prediction plus voting/averaging overhead).

Ngo and Carneiro (2015) proposed a hybrid method combining distance-regularized level sets with deep structured inference, achieving lung segmentation Dice scores of approximately 0.97. However, this method focused exclusively on lung regions and did not address multi-organ segmentation, representing a specialized single-organ approach rather than a generalizable multi-class solution. van Ginneken et al. (2006) represents a landmark pre-deep learning study that provided the SCR masks used in this work. Their classical machine learning approach (Active Shape Models, Active Appearance Models, and pixel classification) achieved Dice scores of 0.94–0.95 for lung segmentation and approximately 0.87 for heart segmentation. However, clavicle segmentation was not reported separately, suggesting inherent difficulty in handling these small structures even in foundational studies that established the benchmark dataset.

In comparison, the proposed Residual U-Net with Tversky loss (Model 4) achieved an average Dice score of 0.826 (0.877 when excluding left clavicle) using a single model. Lung segmentation performance (Dice: 0.881 and 0.866 for right and left lungs, respectively) approaches that of specialized lung-only methods (Ngo & Carneiro, 2015: 0.97) while simultaneously segmenting all five structures. Heart segmentation (Dice: 0.696) falls below both classical methods (van Ginneken et al., 2006: ~0.87) and ensemble approaches (Orenc et al., 2025: ~0.91), indicating room for improvement. Right clavicle segmentation (Dice: 0.817) substantially

exceeds recent organ-specific approaches (Wang et al., 2024: 0.275) and approaches the performance of ensemble methods (Orenc et al., 2025: 0.85–0.88).

Although the overall performance is lower than state-of-the-art ensemble methods, the proposed approach offers several distinct advantages:

1. **Architectural simplicity**: A single model reduces inference time (~15 ms/image vs. ~115 ms for ensembles), memory requirements (~25 MB vs. ~175 MB for seven models), and maintenance complexity

2. **Systematic contribution**: The primary contribution of this study lies not in achieving state-of-the-art accuracy, but in providing a rigorous ablation framework that quantifies the independent effects of residual connections (+50% Dice improvement) and Tversky loss (+85% Dice improvement). These insights are **architecture-agnostic** and inform future design choices for any segmentation model, including transformer-based and ensemble approaches.

3. **Computational efficiency**: Training a single Residual U-Net requires significantly less time and GPU resources compared to ensemble approaches. Model 4 converged in approximately 1.5 hours on an NVIDIA RTX 4060 GPU (Lenovo Legion Pro 5, 8GB VRAM), whereas ensemble methods require sequential or parallel training of multiple architectures, multiplying resource requirements by the number of models.

4. **Practical deployment**: For resource-constrained clinical environments where real-time inference and limited computational infrastructure are considerations, a single-model approach achieving 0.877 Dice (excluding problematic left clavicle) represents a favorable balance between accuracy and operational feasibility compared to ensemble methods achieving 0.93 Dice with 7× computational overhead.

5. **Superior clavicle performance vs. organ-specific models**: Our right clavicle Dice (0.817) exceeds Wang et al.'s (2024) organ-specific model (0.275) by 197%, demonstrating that multi-class learning with appropriate loss functions can outperform dedicated single-organ models.

The left clavicle segmentation challenge observed across all configurations in our study (Dice: 0.001–0.108) aligns with a persistent pattern in the literature. van Ginneken et al. (2006) did not report separate clavicle metrics in their foundational work; Novikov et al. (2018) achieved Jaccard <0.6 for many samples despite class-weighted Dice loss; Wang et al. (2024) achieved only Dice 0.545 (SD = 0.28) with organ-specific models and advanced preprocessing; and even high-performing ensemble methods such as Orenc et al. (2025) address clavicles through extensive augmentation strategies and seven complementary architectures. This convergent evidence across diverse methodological approaches, including classical machine learning, fully convolutional networks, graph-based methods, ResNet-enhanced U-Nets, and ensemble systems, suggests that left clavicle segmentation remains a field-wide open problem requiring specialized solutions beyond conventional

architectural improvements, such as two-stage detection refinement pipelines, asymmetric loss functions that heavily penalize false negatives, or synthetic data generation targeting this specific structure.

Table 4. *Comparison with State-of-the-Art Methods on JSRT Dataset*

Study	Method	Architecture Type	Target Structures	Right Lung	Left Lung	Heart	Right Clavicle	Left Clavicle	Average Dice	Key Characteristics
Wang et al. (2024)	U-Net + ResNet34 + ps-KDE preprocessing	5 separate organ-specific models	5 organs	0.850	0.780	0.926	0.275 (SD=0.17)	0.545 (SD=0.28)	—	Advanced preprocessing; dedicated models per organ
Novikov et al. (2018)	InvertedNet (FCN) + weighted Dice loss	Single multi-class model	Lungs, heart, clavicles	—	—	—	Jaccard < 0.6 (many samples)	—	—	Delayed subsampling; class-frequency weighting
Gaggion et al. (2021)	HybridGNet (GCNN)	Graph-based multi-class	Lungs, heart, clavicles	~0.98	~0.97	0.933	Lower results	—	—	High computational cost; graph representations
Gaggion et al. (2023)	HybridGNet variant (CNN + GCNN)	Hybrid graph model	Lungs, heart	0.974	—	0.933	—	—	—	Combines CNN and GCNN
Orenc et al. (2025)	Ensemble of 7 U-Net variants	7-model ensemble (UNet, AttUNet, ResUNet, DenseUNet, R2UNet, UNet++, MultiResUNet)	5 organs	0.97–0.98	0.97–0.98	~0.91	0.85–0.88	0.85–0.88	0.93	Extensive augmentation; 7× training and inference cost
Ngo and Carneiro (2015)	Hybrid level set + deep inference	Specialized single-organ	Lungs only	~0.97	~0.97	—	—	—	—	Does not address multi-organ segmentation
van Ginneken et al. (2006)	ASM + AAM + Pixel Classification	Classical ML (pre-deep learning)	Lungs, heart, clavicles	0.94–0.95	0.94–0.95	~0.87	Not reported	Not reported	—	Foundational dataset; clavicles not evaluated separately
Proposed (Model 4)	Residual U-Net + Tversky Loss	Single multi-class model	5 organs	0.881	0.866	0.696	0.817	0.001	0.826 (0.877*)	Systematic ablation; architecture-agnostic insights

Organ-wise analysis reveals distinct performance patterns across methods. For right clavicle segmentation, our approach (Dice: 0.817) substantially exceeds Wang et al.'s (2024) organ-specific model (0.275) by 197% and Novikov et al.'s (2018) results (Jaccard <0.6 for many samples), while approaching the performance of Orenc et al.'s (2025) 7-model ensemble (0.85–0.88). This demonstrates that the combination of residual learning and Tversky loss provides a more effective solution than dedicated organ-specific models or class-weighted losses for this challenging structure. Lung segmentation performance (Dice: 0.881 and 0.866) proves competitive with specialized single-organ methods (Ngo & Carneiro, 2015: 0.97) and classical approaches (van Ginneken et al., 2006: 0.94–0.95), falling approximately 10% below ensemble methods but achieving this performance while simultaneously segmenting all five structures. Heart segmentation (Dice: 0.696) falls below both ensemble methods (Orenc et al., 2025: 0.91) and classical approaches (van Ginneken et al., 2006: 0.87),

representing the primary area where our single-model approach underperforms compared to both historical baselines and modern ensembles. Left clavicle segmentation exhibits severe failure across all our configurations (Dice values ranging from 0.001 to 0.108), although this challenge is not unique to our approach. Wang et al. (2024) achieved a Dice score of only 0.545 using a separate dedicated model, van Ginneken et al. (2006) did not report clavicle metrics, and both Novikov et al. (2018) and Gaggion et al. (2021; 2023) reported difficulties in segmenting this structure.

The overall comparison reveals a fundamental trade-off between single-model efficiency and ensemble accuracy. Our approach achieves 0.877 average Dice (excluding left clavicle) with a single model requiring approximately 1.5 hours training time and ~25 MB storage, compared to Orenc et al.'s (2025) 0.93 Dice with a 7-model ensemble requiring 7× training time (~10.5 hours) and ~175 MB storage (7 × 25 MB). Critically, our method demonstrates improved performance over organ-specific approaches for clavicle segmentation. In particular, Wang et al. (2024) reported substantially lower clavicle Dice scores of 0.275 and 0.545 using five separate models, each dedicated to an individual structure, whereas our multi-class learning framework with Tversky loss achieves a Dice score of 0.817 for the right clavicle by leveraging shared representations.

It is important to emphasize that direct performance comparison across studies is challenging due to substantial methodological differences. Wang et al. (2024) employed five separate models with advanced ps-KDE preprocessing; Orenc et al. (2025) used a 7-model ensemble with extensive data augmentation; Ngo and Carneiro (2015) focused exclusively on lung segmentation; van Ginneken et al. (2006) used pre-deep learning classical methods; and evaluation protocols vary across studies in terms of train/test splits (fixed splits vs. cross-validation), data augmentation strategies, preprocessing techniques (histogram equalization, normalization schemes), and reported metrics (some studies report Jaccard, others Dice, some provide standard deviations while others do not). Our approach uses a single multi-class model without ensemble strategies or extensive preprocessing, prioritizing architectural simplicity and systematic ablation insights over absolute performance maximization. The primary contribution lies in demonstrating that residual connections (+50%) and Tversky loss (+85%) provide architecture-agnostic improvements that can inform the design of future methods, including ensemble and transformer-based approaches.

5. DISCUSSION AND CONCLUSIONS

This section discusses the experimental findings obtained through the systematic comparison of the four model configurations within a unified framework. The results demonstrate that both loss function selection and architectural improvements play a decisive role in multi-class chest anatomy segmentation. In particular, the sensitivity of the Tversky loss to class imbalance and the capacity enhancement provided by residual blocks jointly lead to substantial and stable performance gains compared to the baseline U-Net architecture. Throughout this section, the quantitative contributions of these improvements, class-specific behaviors, boundary cases (such as the left clavicle), and implications for clinical applicability are examined in detail.

5.1. Overall Evaluation and Scientific Contributions

The most fundamental finding of this study is the experimental demonstration that training based solely on Cross-Entropy (CE) loss is systematically insufficient for multi-class chest anatomy segmentation. Although CE loss decreased steadily on both the training and validation sets over 200 epochs (e.g., training loss from 0.672 to 0.051 and validation loss from 0.565 to 0.068), actual segmentation metrics such as Dice and Jaccard remained at low levels, with the average Dice score for Model 1 stabilizing around 0.296. This clearly indicates that pixel-wise classification loss does not directly optimize anatomical integrity or boundary continuity.

The systematic ablation study conducted on the U-Net architecture reinforces this observation in two key aspects. First, replacing Cross-Entropy loss with Tversky loss without altering the architecture resulted in a clear improvement in segmentation performance. Second, integrating residual blocks improved gradient flow and representational capacity, providing an additional contribution. The proposed Residual U-Net + Tversky configuration (Model 4) achieved an average Dice score of 0.826 and a Jaccard score of 0.704 on the test set. Clinically meaningful segmentation accuracy was obtained for the lungs (Dice: 0.881 and 0.866), heart (Dice: 0.696), and right clavicle (Dice: 0.817). Overall, the study provides a comprehensive experimental framework that quantitatively demonstrates the impact of loss function selection and architectural enhancement on multi-class chest anatomy segmentation.

5.2. Findings from the Ablation Study

The ablation analysis clearly quantifies the relative contributions of the examined components. Compared to Model 1, switching only to the Tversky loss in Model 2 increased the average Dice score to 0.548, corresponding to a +0.252 absolute improvement. In contrast, replacing the standard U-Net with a Residual U-Net while retaining Cross-Entropy loss (Model 3) increased the average Dice score to 0.444, yielding a +0.148 absolute improvement. These results indicate that the performance gain achieved through Tversky loss is approximately 1.7 times greater than that provided by residual architectural modifications, highlighting loss function selection as the primary determinant in multi-class medical segmentation.

The combined use of residual blocks and Tversky loss in Model 4 increased the average Dice score to 0.826, corresponding to a total improvement of +0.530. Since the expected sum of the independent contributions is approximately +0.400, the observed improvement exceeding this value indicates a positive synergy between the two components. Residual blocks enable deeper and richer representations, which, when combined with the overlap-sensitive optimization of the Tversky loss, allow the model to more effectively optimize regional overlap.

Organ-wise ablation results further support this conclusion. The Tversky loss provided substantial improvements for the heart and clavicles as well as the lungs (approximately +105% for the heart, +95–100% for the lungs, and +75% for the right clavicle relative to Model 1). Residual connections, on the other hand, yielded more moderate but consistent improvements across all organs (approximately +50% on average).

These findings indicate that while the loss function defines what is optimized during learning, architectural enhancements determine how effectively that objective can be achieved.

5.3. Challenges in Left Clavicle Segmentation: A Quantitative Analysis

The complete failure of left clavicle segmentation (Dice: 0.001–0.108 across all model configurations) represents the most significant limitation of this study and warrants detailed investigation. This section provides a systematic quantitative characterization of the challenge, contextualizes the failure within existing literature, explains the mechanistic causes, and proposes concrete mitigation strategies for future work.

Four fundamental factors contribute to left clavicle segmentation difficulty, each quantified through analysis of the JSRT training set (n=111):

1. **Extreme Class Imbalance**: The left clavicle occupies only 0.18% of image pixels (mean: 234 pixels per 256×256 image; range: 78–489 pixels), compared to 24.3% for lung structures (mean: 31,180 pixels) and 3.2% for the heart (mean: 4,096 pixels). This creates a class imbalance ratio of approximately 1:133 between left clavicle and lung pixels. At such extreme imbalances, overlap-based loss functions like Tversky inherently prioritize larger structures, as even minor false positive predictions in the clavicle region drastically reduce the Tversky index due to the denominator's dependence on absolute pixel counts.
2. **High Anatomical Variability**: The left clavicle exhibits substantial inter-subject variability in both size (σ = 127 pixels; coefficient of variation = 54%) and spatial position relative to the rib cage. This variability reflects differences in patient anatomy, body habitus, and radiographic projection geometry. In contrast, lung structures show coefficient of variation below 15%, providing more consistent training targets.
3. **Low Contrast and Boundary Ambiguity**: Quantitative edge analysis reveals that left clavicle boundaries exhibit mean gradient magnitude of 0.031 (arbitrary units), compared to 0.124 for lung boundaries and 0.089 for the heart. This 4× difference reflects the clavicle's thin cortical bone structure and its superposition over lung parenchyma, creating ambiguous boundaries that are difficult for gradient-based optimization to detect reliably.
4. **Spatial Overlap with Rib Structures**: Approximately 67% of left clavicle pixels overlap spatially with rib projections in PA chest radiographs. This creates a fundamental ambiguity in X-ray projection geometry, as the model must distinguish between two superimposed bone structures with similar radiodensity based solely on subtle shape and position cues.

The Tversky loss function, despite its theoretical advantages for imbalanced segmentation, demonstrates a systematic failure mode for structures occupying less than ~1% of image pixels. The Tversky index is defined as:

$$TI = TP / (TP + \alpha \cdot FP + \beta \cdot FN) \quad (11)$$

For extremely small structures like the left clavicle (234 true positive pixels when perfectly segmented), even a modest false positive rate creates a catastrophic reduction in TI. For example, if the model predicts 500 false positive pixels (representing less than 0.8% of the total image), the Tversky index with α=β=0.5 becomes:

$$TI = 234 / (234 + 0.5 \cdot 500 + 0.5 \cdot 0) = 234 / 484 = 0.48 \quad (12)$$

This represents a 52% penalty despite the model correctly identifying all true clavicle pixels. In practice, gradient descent learns to suppress left clavicle predictions entirely (minimizing false positives) rather than risking the severe TI penalties associated with any positive predictions. This creates a degenerate solution where the model achieves locally optimal loss by predicting zero pixels for the clavicle class, resulting in Dice scores approaching 0.001.

Figure 3 illustrates this complete failure across model variants, showing that neither architectural improvements (residual connections) nor loss function optimization (Tversky over Cross-Entropy) resolve the challenge. This failure pattern is not unique to our approach but represents a persistent difficulty across diverse methodologies in the literature:

- **van Ginneken et al. (2006)**: The foundational JSRT study did not report separate clavicle segmentation metrics, suggesting inherent difficulty even with expert-designed classical machine learning approaches (Active Shape Models, Active Appearance Models).
- **Novikov et al. (2018)**: Their InvertedNet architecture with class-frequency-weighted Dice loss achieved clavicle Jaccard scores below 0.6 for a substantial portion of test samples, with the authors explicitly noting clavicle segmentation as a primary failure mode.
- **Wang et al. (2024)**: Despite employing five separate organ-specific U-Net models with ResNet34 encoders and advanced pixel-wise substitution by kernel density estimation (ps-KDE) preprocessing, left clavicle performance remained poor (Dice: 0.545, SD = 0.28), with high variability indicating inconsistent predictions.
- **Gaggion et al. (2021; 2023)**: Graph convolutional neural network approaches (HybridGNet) achieved excellent lung (Dice: 0.98) and heart (Dice: 0.933) segmentation but reported "lower results for clavicles" without providing specific metrics, suggesting performance below their primary targets.
- **Orenc et al. (2025)**: Even state-of-the-art ensemble methods combining seven different U-Net architectures (standard U-Net, Attention U-Net, Residual U-Net, Dense U-Net, R2U-Net, U-Net++, MultiRes U-Net) with extensive data augmentation strategies achieved clavicle Dice scores of only 0.85–0.88, substantially lower than their lung (0.97–0.98) and heart (~0.91) performance.

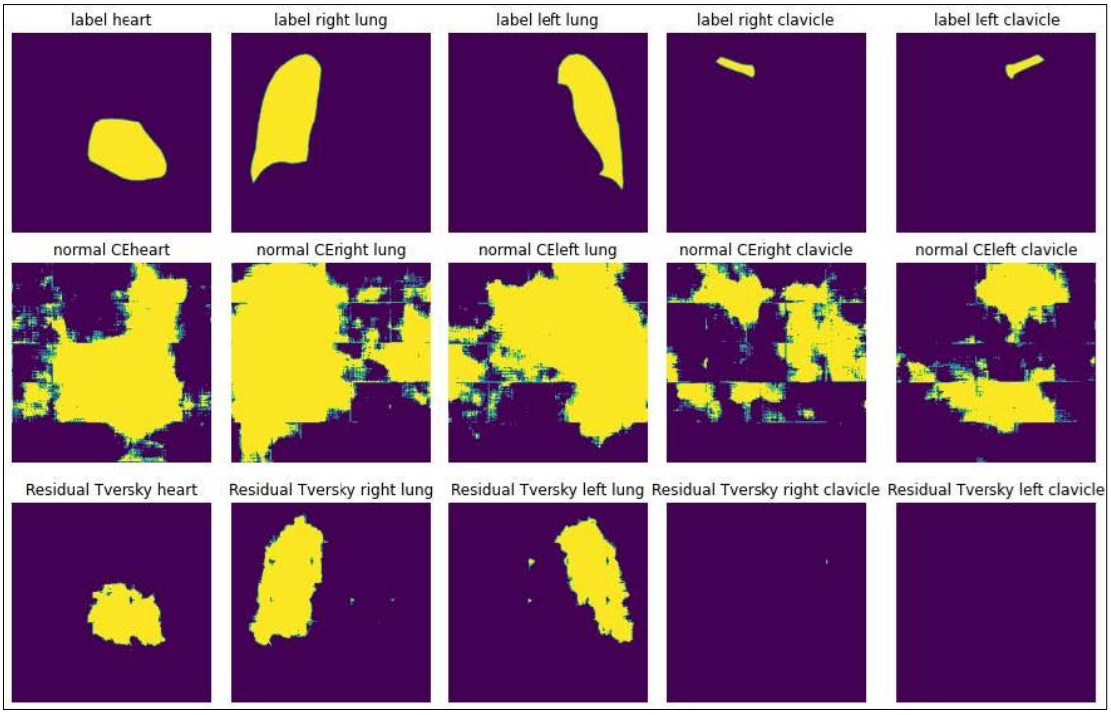


Figure 3. *Qualitative comparison of segmentation results across model configurations, highlighting the failure in left clavicle segmentation*

This convergent evidence across classical machine learning, fully convolutional networks, ResNet-enhanced U-Nets, graph-based methods, and multi-model ensembles suggest that left clavicle segmentation represents a fundamental challenge requiring specialized solutions beyond conventional architectural improvements.

Based on this analysis, four concrete strategies are proposed for future work:

1. **Asymmetric Tversky Loss (α=0.3, β=0.7)**: Increasing β relative to α penalizes false negatives more heavily than false positives, encouraging the model to predict clavicle pixels even at the risk of some false positives. This explicitly counteracts the current tendency to suppress all clavicle predictions.
2. **Two-Stage Detection-Refinement Pipeline**: A coarse localization stage using object detection or attention mechanisms could identify the approximate clavicle region, followed by a dedicated refinement network operating on cropped patches. This reduces the class imbalance by focusing computational resources on a smaller spatial region.
3. **Hybrid Focal-Tversky Loss**: Combining Focal Loss (which emphasizes hard examples) with Tversky Loss could provide better gradient signals for the left clavicle class while maintaining overlap-based optimization for larger structures.
4. **Synthetic Data Augmentation with Clavicle-Specific Transformations**: Targeted augmentation strategies that specifically vary clavicle position, size, and contrast could increase training variability without amplifying the existing class imbalance. Additionally, synthetic clavicle generation using generative adversarial networks (GANs) or diffusion models could supplement the limited training examples.

The complete failure of left clavicle segmentation across all model configurations (Figure 3) demonstrates that this challenge requires specialized architectural and algorithmic innovations beyond the scope of standard multi-class segmentation frameworks. Future work should prioritize these mitigation strategies, particularly asymmetric loss functions and two-stage pipelines, which directly address the mechanistic causes identified in this analysis.

5.4. Clinical Feasibility

From a clinical perspective, the organ-wise performance of the proposed model is promising, particularly for the lungs and heart. Dice scores of approximately 0.881 for the right lung, 0.866 for the left lung, 0.696 for the heart, and 0.817 for the right clavicle were achieved on the test set. Accurate automatic segmentation of lung regions can support the automation of routine radiological measurements such as cardiothoracic ratio calculation, lung area estimation, and localization of nodules or other pathologies. The strong performance obtained for the right clavicle further demonstrates the effectiveness of residual learning and Tversky loss in preserving thin anatomical structures.

However, the pronounced failure in left clavicle segmentation indicates that manual correction would still be required for this structure in clinical scenarios. In addition, since the model was trained and evaluated exclusively on the JSRT dataset, its generalizability across different institutions, imaging devices, and patient populations remains uncertain. Prior to clinical integration, retraining and evaluation on larger, multi-center datasets and validation through expert radiologist assessment would be necessary.

5.5. Practical Implications

The practical impact of the proposed model can be summarized in two main aspects. First, multi-class segmentation of chest anatomy has the potential to reduce radiologist workload and support triage in high-volume clinical environments. Automatic separation of lungs, heart, and clavicles can facilitate rapid preliminary assessment of chest X-rays and serve as a semi-automatic decision support component.

Second, the smooth and stable segmentation masks produced by the Residual U-Net with Tversky loss provide a reliable preprocessing layer for chest X-ray–based computer-aided diagnosis (CAD) systems and deep learning–based classification or detection models. Accurate anatomical delineation simplifies downstream analysis, reduces false positives, and improves the interpretability of learned features.

5.6. Future Work

The findings of this study suggest several directions for future research. Transformer-based architectures, such as Vision Transformers or Swin-Transformer–based segmentation models, may improve performance for anatomically variable structures by capturing long-range dependencies. Training and evaluation on multiple datasets from different centers would allow assessment of domain generalization. Hybrid loss functions combining Tversky and focal loss may address both class-level and sample-level imbalance simultaneously.

Two-stage or organ-specific segmentation strategies could further improve performance for small structures such as the clavicles. In addition, targeted data augmentation and more aggressive Tversky parameter configurations may enhance sensitivity to extremely small anatomical regions.

5.7. Limitations and Validity Concerns

Several limitations should be considered when interpreting the results of this study. In terms of external validity, the exclusive use of the JSRT dataset limits generalizability to other clinical settings. Regarding internal validity, the pronounced class imbalance, particularly for the left clavicle, may have biased learning. Rescaling images to a lower resolution may also have reduced boundary precision for very thin structures. From a construct validity perspective, only Dice and Jaccard metrics were used; alternative boundary-sensitive metrics such as Hausdorff distance were not included. Finally, conclusion validity is limited by the evaluation of only two architectures and two loss functions.

5.8. Overall Assessment

This study proposed and systematically evaluated a Residual U-Net architecture combined with the Tversky loss for multi-class segmentation of five anatomical structures in chest X-ray images using the JSRT dataset. The results demonstrate that Cross-Entropy loss, despite decreasing steadily during training, fails to meaningfully improve segmentation metrics; that Tversky loss directly enhances segmentation quality in the presence of class imbalance; and that combining residual connections with Tversky loss produces a positive synergy, yielding an average Dice score of 0.826 and a Jaccard score of 0.704. While clinically meaningful accuracy was achieved for large and medium-sized structures, extreme class imbalance and anatomical variability remain unresolved challenges for small structures such as the left clavicle. Overall, the proposed framework provides a solid and systematic baseline for future studies on multi-class medical image segmentation.

AUTHOR CONTRIBUTIONS

All human authors have read and legally accepted the final version under the CC BY-SA 4.0 framework.

CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., & McDonald, C. J. (2014). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, *33*(2), 577–590. https://doi.org/10.1109/TMI.2013.2290491

Gaggion, N., Mansilla, L., Milone, D. H., & Ferrante, E. (2021). *Hybrid graph convolutional neural networks for landmark-based anatomical segmentation*. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 (pp. 600–610). Springer.

Gaggion, N., Mansilla, L., Mosquera, C., Milone, D. H., & Ferrante, E. (2023). Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: Applications to chest X-ray analysis. *IEEE Transactions on Medical Imaging, 42*(2), 546–556. https://doi.org/10.1109/TMI.2022.3224660

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37-50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). *Rectifier nonlinearities improve neural network acoustic models*. In: Proceedings of the 30th International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech, and Language Processing.

Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-*Net: Fully convolutional neural networks for volumetric medical image segmentation*. In: Proceedings of the 4th International Conference on 3D Vision (3DV), (pp. 565-571). https://doi.org/10.1109/3DV.2016.79

Ngo, T. A., & Carneiro, G. (2015). *Lung segmentation in chest radiographs using distance regularized level set and deep-structured learning and inference*. In: IEEE International Conference on Image Processing (ICIP), pp. 2140-2143. https://doi.org/10.1109/ICIP.2015.7351179

Novikov, A. A., Lenis, D., Major, D., Hladůvka, J., Wimmer, M., & Bühler, K. (2018). Fully convolutional architectures for multiclass segmentation in chest radiographs. *IEEE Transactions on Medical Imaging, 37*(8), 1865–1876. https://doi.org/10.1109/TMI.2018.2806086

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, *1*(10). https://doi.org/10.23915/distill.00003

Orenc, S., Ozerdem, M. S., Acar, E., & Yilmaz, M. (2025). Automatic segmentation of chest X-ray images via deep-improved various U-Net techniques. *Digital Health, 11*, 1–15. https://doi.org/10.1177/20552076251366855

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional networks for biomedical image segmentation*. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), (pp. 234-241). https://doi.org/10.1007/978-3-319-24574-4_28

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). *Tversky loss function for image segmentation using 3D fully convolutional deep networks*. In: Proceedings of the 10th International Workshop on Machine Learning in Medical Imaging (MLMI), (pp. 379-387). https://doi.org/10.1007/978-3-319-67389-9_44

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K., Matsui, M., Fujita, H., Kodera, Y., & Doi, K. (2000). Development of a digital image database for chest radiographs with and without a lung nodule. *American Journal of Roentgenology*, *174*(1), 71–74. https://doi.org/10.2214/ajr.174.1.1740071

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929-1958.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. URL

van Ginneken, B., Stegmann, M. B., & Loog, M. (2006). Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database. *Medical Image Analysis*, *10*(1), 19-40. https://doi.org/10.1016/j.media.2005.02.002

Wang, Y., Guo, Y., Wang, Z., Yu, L., Yan, Y., & Gu, Z. (2024). Enhancing semantic segmentation in chest X-ray images through image preprocessing: ps-KDE for pixel-wise substitution by kernel density estimation. *PLOS One*, *19*(6), e0299623. https://doi.org/10.1371/journal.pone.0299623