BİNGÖL UNIVERSITY

Turkish Journal of Nature and Science

e-ISSN: 2149-6366
www.dergipark.gov.tr/tdfd

Research Article

Volume 15, Issue 3, Page 1-11, 2026

Received: 11.03.2026
Accepted: 13.05.2026
Published: 30.09.2026

DOI: https://doi.org/10.46810/tdfd.1907062

# Comparative Evaluation of U-Net-Based Models for Lung Lesion Segmentation in CT Images

**Murat KILIÇ[1], Abdulkadir YELMAN[2*], Hüseyin ÜZEN[3], Hüseyin FIRAT[4], İpek BALIKÇI ÇİÇEK[5], Merve BIYIKLI[6], Abdulkadir ŞENGÜR[7]**

[1] Department of Thoracic Surgery, Faculty of Medicine, Inonu University Turgut Ozal Medical Center, Malatya, Türkiye
[2] Department of Computer Technologies, Genç Vocational School, Bingöl University, Bingöl, Türkiye
[3] Department of Computer Engineering, Faculty of Engineering and Architecture, Bingöl University, Bingöl, Türkiye
[4] Department of Computer Engineering, Faculty of Engineering, Dicle University, Diyarbakır, Türkiye
[5] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University Turgut Ozal Medical Center, Malatya, Türkiye
[6] Department of Thoracic Surgery, Faculty of Medicine, Inonu University Turgut Ozal Medical Center, Malatya, Türkiye
[7] Department of Electrical and Electronics Engineering, Faculty of Technology, Fırat University, Elazığ, Türkiye

**Corresponding author: Abdulkadir YELMAN, ayelman@bingol.edu.tr*

## Abstract

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, and early detection plays a crucial role in improving treatment success and patient survival. In this study, the problem of lesion segmentation in lung computed tomography (CT) images was addressed, and the performance of different U-Net-based deep learning architectures was comparatively evaluated. The dataset used in this study consisted of 22,782 CT slices obtained from 223 patients. Among these images, 3,584 slices contained lesions, while 19,198 slices were healthy images. To investigate the impact of class imbalance in the dataset on model performance, three different experimental scenarios were applied: using only lesion-containing images, using all images together, and using a balanced training dataset. Segmentation performance was analyzed using both image-based and patient-based evaluation approaches. The results indicate that data distribution has a significant impact on segmentation performance. In particular, balancing the training dataset improved the results, especially in patient-based evaluations.

## Keywords

Lung cancer, computed tomography, deep learning, image segmentation, U-Net, medical image analysis

Author(s) publishing with the journal retain(s) the copyright to their work licensed under the CC BY-NC-ND 4.0.
https://creativecommons.org/licenses/by-nc-nd/4.0/

BİNGÖL ÜNİVERSİTESİ

Türk Doğa ve Fen Dergisi

e-ISSN: 2149-6366
www.dergipark.gov.tr/tdfd

Araştırma Makalesi

Cilt 15, Sayı 3, Sayfa 1-11, 2026

Alınış: 11.03.2026
Kabul: 13.05.2026
Basım: 30.09.2026

DOI: https://doi.org/10.46810/tdfd.1907062

# BT Görüntülerinde Akciğer Lezyonu Segmentasyonu için U-Net Tabanlı Modellerin Karşılaştırmalı Değerlendirilmesi

**Murat KILIÇ[1], Abdulkadir YELMAN[2*], Hüseyin ÜZEN[3], Hüseyin FIRAT[4], İpek BALIKÇI ÇİÇEK[5], Merve BIYIKLI[6], Abdulkadir ŞENGÜR[7]**

[1] İnönü Üniversitesi Turgut Özal Tıp Merkezi, Tıp Fakültesi, Göğüs Cerrahisi Anabilim Dalı, Malatya, Türkiye
[2] Bingöl Üniversitesi, Genç Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Bingöl, Türkiye
[3] Bingöl Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Bingöl, Türkiye
[4] Dicle Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Diyarbakır, Türkiye
[5] İnönü Üniversitesi Turgut Özal Tıp Merkezi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Malatya, Türkiye
[6] İnönü Üniversitesi Turgut Özal Tıp Merkezi, Tıp Fakültesi, Göğüs Cerrahisi Anabilim Dalı, Malatya, Türkiye
[7] Fırat Üniversitesi, Teknoloji Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Elazığ, Türkiye

**Sorumlu yazar: Abdulkadir YELMAN, ayelman@bingol.edu.tr*

## Özet

Akciğer kanseri dünya genelinde kanser kaynaklı ölümlerin başlıca nedenlerinden biridir ve hastalığın erken evrede tespit edilmesi tedavi başarısı açısından büyük önem taşımaktadır. Bu çalışmada akciğer bilgisayarlı tomografi (BT) görüntülerinde lezyon segmentasyonu problemi ele alınmış ve farklı U-Net tabanlı derin öğrenme mimarilerinin performansları karşılaştırmalı olarak incelenmiştir. Çalışmada toplam 223 hastaya ait 22.782 BT kesitinden oluşan bir veri kümesi kullanılmıştır. Veri kümesinde 3.584 kesit lezyon içerirken 19.198 kesit sağlıklı görüntülerden oluşmaktadır. Veri kümesindeki sınıf dengesizliğinin model performansı üzerindeki etkisini incelemek amacıyla üç farklı deneysel senaryo uygulanmıştır: yalnızca lezyon içeren görüntülerin kullanılması, tüm görüntülerin birlikte kullanılması ve eğitim veri setinde dengeli veri kullanımı. Segmentasyon performansı görüntü bazlı ve hasta bazlı değerlendirme yaklaşımları kullanılarak analiz edilmiştir. Elde edilen sonuçlar veri dağılımının segmentasyon performansı üzerinde önemli bir etkiye sahip olduğunu göstermiştir. Özellikle eğitim veri setinde dengeli veri kullanımının hasta bazlı değerlendirmelerde performansı artırdığı belirlenmiştir.

## Anahtar kelimeler

Akciğer kanseri, bilgisayarlı tomografi, derin öğrenme, görüntü segmentasyonu, U-Net, medikal görüntü analizi

## 1. INTRODUCTION

Lung cancer remains one of the leading causes of cancer-related mortality worldwide due to its high incidence and mortality rates. According to data reported by the World Health Organization, lung cancer is the primary cause of cancer-related deaths globally. In 2022, approximately 2.5 million new lung cancer cases were reported worldwide, accounting for nearly 12.4% of all cancer cases. In the same year, lung cancer caused approximately 1.8 million deaths, representing 18.7% of all cancer-related fatalities [1]. These statistics clearly demonstrate that lung cancer continues to be a major global health concern.

Lung cancer is broadly classified into two major subtypes based on molecular and histological characteristics: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC accounts for approximately 80-85% of all lung cancer cases and includes several subtypes such as squamous cell carcinoma, adenocarcinoma, large cell carcinoma, adenosquamous carcinoma, and sarcomatoid carcinoma. In contrast, SCLC represents approximately 15-18% of cases and is characterized by an aggressive clinical course. In metastatic cases, the two-year survival rate has been reported to be below 10% [2],[3].

Early detection plays a critical role in improving treatment outcomes and increasing patient survival rates. The effectiveness of treatment largely depends on identifying the disease at an early stage, whereas survival rates significantly decrease in advanced-stage diagnoses. Early detection not only prolongs patient survival but also expands the available treatment options. Consequently, the development of technological approaches aimed at early lung cancer detection has become an important focus of research in the healthcare field [4].

Computed tomography (CT) is one of the most widely used imaging modalities for the early detection of lung cancer. CT images provide detailed information about lung nodules and lesions, facilitating clinical assessment. However, manual examination of CT slices can be a time-consuming and demanding task for radiologists, particularly in clinical environments where a large number of images must be analyzed [5]. This situation may lead to missed findings or variability in interpretation.

To address these challenges, computer-aided diagnosis (CAD) systems have been developed to automatically analyze CT images and support clinical decision-making processes. In recent years, machine learning techniques, particularly convolutional neural networks (CNNs), have been widely applied in lung cancer detection studies. CNN-based CAD systems can automatically detect, classify, and segment lung nodules with high accuracy, enabling faster image analysis and improving diagnostic consistency [6],[7].

Numerous studies have been conducted on the classification and segmentation of lung CT images. Some of these studies are summarized in Table 1. However, several limitations remain in existing literature. The complex background structures in CT images can make accurate lesion detection and segmentation challenging. In addition, datasets often contain a significantly larger number of normal images than lesion-containing images, leading to class imbalance during training and negatively affecting model performance.

**Table 1.** Representative studies on lung lesion detection and classification in CT images

| Reference | Method | Dataset | Result | Potential Limitation |
|---|---|---|---|---|
| Gu et al. (2018) [8] | 3D CNN | LUNA16 | Sensitivity: 92.93% | Some nodules could not be detected because pure ground-glass nodules were underrepresented in the dataset. |
| Lakshmanaprabu et al. (2019) [9] | Modified Gravitational Search Algorithm | ELCAP | Accuracy: 94.56% | Cystic and solid subtypes were not considered and segmentation was not performed. |
| Paing et al. (2020) [10] | Chain Coding + Random Forest | LIDC-IDRI | Accuracy: 93.11% | Nodule detection depends on empirical threshold values. |
| Perez and Arbelaez (2020) [11] | 3D CNN | LIDC-IDRI | AUC: 91.3% | Subtype classification was not performed. |
| Naqi et al. (2020) [12] | Stacked Autoencoder | LIDC-IDRI | Accuracy: 96.9% | Solid and cystic lesions were not evaluated. |
| Khan et al. (2020) [13] | Ensemble Bagging Tree | Kaggle Lung Cancer | Accuracy: 99.4% | Only two classes were used. |
| Suresh and Mohan (2020) [14] | ROI-based CNN | LIDC-IDRI | Accuracy: 93.9% | Segmentation was not performed. |
| Jain et al. (2021) [15] | GAN-based Model | LIDC-IDRI | Dice: 79.86% | Subtype differentiation was not performed. |
| Shimazaki et al. (2022) [16] | CNN | Osaka University | Dice: 52% | Only malignant nodules were used. |
| Swain et al. (2024) [17] | Inception v3 | TCIA-PET/CT | Accuracy: 98.29% | Segmentation was not performed. |

**Table 1 Cont.** Representative studies on lung lesion detection and classification in CT images

| | | | | |
|---|---|---|---|---|
| Xiao et al. (2020) [18] | 3D-Res2UNet | LUNA16 | Dice: 95.3% | High false positive rate. |
| Kumar et al. (2024) [19] | EfficientNet-B3 | LIDC-IDRI | Accuracy: 93% | Only classification was performed. |
| Cifci (2022) [20] | SegChaNet (U-Net + Channel Attention) | TCIA | Accuracy: 98.90%, F1: 98.49% | High performance may be influenced by dataset characteristics and limited external validation. |
| Alshayeji and Abed (2023) [21] | DeepLabv3+ (ResNet-18) | LIDC-IDRI | Accuracy: 99.86%, IoU: 0.715 | High accuracy values may not fully reflect real-world generalization performance. |
| Wu et al. (2024) [22] | S3TU-Net (CNN + Transformer) | LIDC-IDRI | Accuracy: 99.53%, Dice: 89.04% | Model complexity and computational cost are relatively high. |
| Rikhari et al. (2024) [23] | Model-V (Ensemble CNN) | AIIMS + LUNA16 | Dice: 84%, IoU: 74% | Ensemble models increase complexity and require higher computational resources. |
| Akila Agnes et al. (2024) [24] | Wavelet U-Net++ | LIDC-IDRI | Accuracy: 98.9%, Dice: 93.7% | Performance may depend on specific preprocessing techniques. |
| Vijayakumar et al. (2025) [25] | CapsNet + U-Net Hybrid | IQ-OTH/NCCD | Accuracy: 98%, F1: 98%, Dice: 94% | Model performance may be affected by dataset size and class distribution. |

A review of the existing literature indicates that many studies rely on datasets with a limited number of samples and insufficient representation of lesion subtypes. This limitation poses a significant challenge for developing reliable and generalizable models. Studies that do not differentiate between solid and cystic lesions often provide limited clinical insight. In addition, some approaches primarily focus on segmenting large and clearly visible lesions, which may result in smaller or more complex lesions being overlooked. The limitations summarized in Table 1 reflect both the findings reported in the original studies and the interpretations derived from the methodologies and results presented in those works.

Similar limitations are also observed in commonly used datasets. For instance, although the LUNA16 dataset contains 1,186 nodules from 888 CT scans, it does not adequately represent different lesion subtypes such as solid and cystic lesions [26]. Likewise, the LIDC-IDRI dataset includes 7,371 lesions collected from 1,018 patients; however, it provides limited information regarding detailed subtype differentiation [27]. Other widely used datasets, including ELCAP [28], IQ-OTH/NCCD and TCIA-PET/CT, also present several limitations, such as relatively small sample sizes or the absence of pixel-level segmentation annotations. Therefore, more comprehensive datasets and advanced deep learning approaches are required to enable a more detailed analysis of different lesion types in lung CT images. In particular, accurate pixel-level segmentation of lesions and reliable classification of lesion subtypes play a crucial role in supporting diagnosis and treatment processes in clinical practice.

In this context, the present study investigates the lesion segmentation problem in lung CT images by considering different data usage strategies and deep learning-based models. To this end, the performance of several U-Net-based segmentation architectures was comparatively analyzed. The impact of the pronounced class imbalance in the dataset on model performance was evaluated through three experimental scenarios: training using only lesion-containing images, training using all available images, and training with balanced data in the training set. Model performance was evaluated using both image-based and patient-based assessment approaches. The findings demonstrate that data distribution and data usage strategies have a significant influence on model performance and generalization capability in lung lesion segmentation tasks.

## 2. MATERIAL AND METHOD

### 2.1. Dataset

In this study, a dataset was used to investigate the problem of lesion segmentation in lung computed tomography (CT) images. The dataset used in this study is an original dataset obtained from CT images acquired at Inonu University Turgut Ozal Medical Center. The dataset was obtained under the approval of the ethics committee dated July 30, 2024, and numbered 2024-6289. The dataset includes lung CT images obtained from a total of 223 patients. The CT scans of each patient were divided into slices covering the lung parenchyma. Since the slice ranges containing the lung region vary among patients, the number of slices per patient also differs. Accordingly, the number of slices per patient ranges between approximately 50 and 250. In total, the dataset consists of 22,782 CT slices. Among these images, 3,584 slices contain lung lesions, while 19,198 slices do not contain any lesions. The overall distribution of the dataset is presented in Table 2.

**Table 2.** Overall distribution of the dataset

| Feature | Value |
|---|---|
| Total number of patients | 223 |
| Total CT slices | 22,782 |
| Number of slices containing lesions | 3,584 |
| Number of slices without lesions | 19,198 |
| Number of slices per patient | 50-250 |

To ensure a reliable evaluation of model generalization and to prevent data leakage, the dataset was divided at the patient level. All CT slices belonging to a single patient were assigned exclusively to either the training or the test set, and no overlap was allowed between these sets. The dataset was split using an 80:20 ratio at the patient level, where approximately 80% of the patients were used for training and the remaining 20% were reserved for testing. This strategy ensures that the model is evaluated on entirely unseen patient data and provides a more realistic assessment of generalization performance in medical image analysis.

The lesion regions in the acquired images were manually annotated at the pixel level. During the annotation process, Label Studio, an open-source data annotation tool, was used, and the outputs were generated in COCO format. CT slices that did not contain lesions were included in the dataset without annotation. Examples of annotated CT slice images are presented in Figure 1, which illustrates different lesion types, including malignant, benign, and cystic cases. The dataset includes lesions with varying sizes, shapes, and anatomical locations. Some lesions appear as small, well-defined nodules, while others present as larger and more irregular mass-like structures. In addition, certain lesions are located in anatomically complex regions, such as near the hilum or pleural boundaries, which further increases the difficulty of accurate segmentation. However, detailed subtype-level annotations (such as nodule vs. mass classification) were not available. All annotations were performed by a single annotator with domain knowledge in medical image analysis. Although inter-annotator agreement was not evaluated in this study, careful attention was paid to ensure consistency during the annotation process. This may represent a limitation in terms of ground truth reliability and will be addressed in future work by incorporating multiple annotators and evaluating annotation consistency.

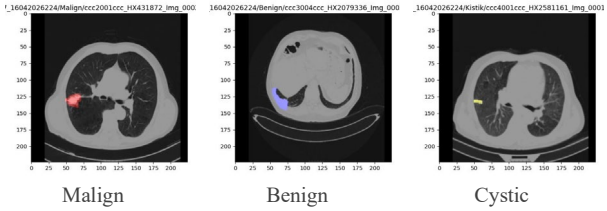



**Figure 1.** Example annotated CT slice images showing different lesion types. From left to right: malignant, benign, and cystic lesions. The lesion regions are highlighted with colored annotations.

## 2.2. Data Preprocessing

Several preprocessing steps were applied to the dataset to enable effective training of deep learning-based segmentation models. First, all CT slices were resized to match the input dimensions required by the model. All CT slices were resized to a fixed spatial resolution of 224 x 224 pixels to match the input requirements of the employed encoder backbones (ResNet50, VGG16, MobileNetV2, InceptionV3, and EfficientNet-B0). This resolution was selected to ensure compatibility with pre-trained ImageNet weights and to maintain a balance between computational efficiency and the preservation of anatomical details. This step ensures that images obtained at different resolutions can be processed consistently. During the resizing process, the aspect ratio of the images was preserved to minimize geometric distortion and to reduce the risk of losing small lesion information. In addition, the pixel values were normalized to a predefined range, which improves numerical stability during the training process.

Since the dataset contains significantly fewer lesion-containing images than healthy images, a class imbalance problem arises. To address this issue, different data usage strategies were developed, and their effects on model performance were evaluated experimentally.

## 2.3. Segmentation Models

U-Net-based encoder-decoder architectures were employed for the segmentation of lung lesions. The U-Net architecture is a widely used deep learning approach in medical image segmentation and generates a segmentation map by producing a class prediction for each pixel. In the encoder part of the architecture, multi-scale features are extracted using successive convolution and pooling layers. During this process, the spatial dimensions of the image decrease while the depth of the feature maps increases. In the decoder part, the extracted feature maps are expanded through up-sampling operations to obtain pixel-level segmentation outputs [29].

One of the key characteristics of the U-Net architecture is the use of skip connections. These connections transfer low-level spatial information obtained during the encoder stage to the decoder stage, thereby improving the segmentation of small or irregularly shaped lesions.

To enhance segmentation performance, different deep convolutional neural network (CNN) architectures were employed in the encoder part of the U-Net model. In this context, ResNet50, VGG16, MobileNetV2, InceptionV3, and EfficientNet-B0 architectures were used as feature extractors in the encoder component of the U-Net framework. These architectures were integrated into the model using a transfer learning approach with weights pre-trained on the ImageNet dataset.

All encoder backbones were initialized with ImageNet pre-trained weights. During training, the entire network, including both encoder and decoder components, was trained in an end-to-end manner without freezing any layers. This fine-tuning strategy enables the model to adapt pre-trained features to the domain-specific characteristics of lung CT images and improves segmentation performance.

The ResNet-based encoder structure enables more stable training of deep networks through residual connections. The VGG16 architecture provides strong and stable feature representations through sequential convolutional layers. The MobileNet architecture offers efficient feature

extraction with lower computational cost due to its depthwise separable convolution structure. The Inception architecture facilitates the simultaneous learning of structural features at multiple scales using multi-scale filter structures. The EfficientNet architecture achieves high accuracy and computational efficiency by balancing the scaling of network depth, width, and resolution.

For comparison purposes, the classical U-Net architecture was also included in the experimental study. All models were trained using the same dataset and training parameters, allowing a comparative analysis of the effects of different encoder architectures on segmentation performance. The overall architecture of the U-Net-based encoder-decoder structure used in this study is illustrated in Figure 2.

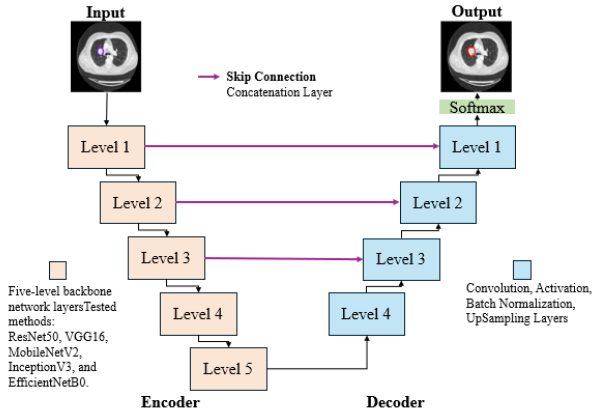



**Figure 2.** U-Net-based encoder-decoder segmentation architecture used in this study.

## 2.4. Training Parameters

All models were trained using the same training parameters to ensure a fair comparison among different architectures. The Adam optimization algorithm was used during training, with a learning rate of 0.0001, a batch size of 16, and a total of 100 epochs.

In addition, a combined loss function consisting of Binary Cross-Entropy (BCE) and Dice Loss was used during the training process. The BCE component enables pixel-wise classification by penalizing incorrect predictions at the individual pixel level, while Dice Loss focuses on maximizing the overlap between the predicted segmentation and the ground truth regions. This combination was specifically preferred to address the class imbalance problem inherent in the dataset, where lesion regions represent a relatively small portion of the overall image. While BCE contributes to stable convergence during training, Dice Loss improves the model's sensitivity to small lesion areas and enhances segmentation performance under imbalanced conditions. The same loss function was consistently applied across all experimental scenarios to ensure a fair and controlled comparison.

All experiments were conducted using a single training run for each model under fixed training conditions. The primary objective of this study was to perform a controlled comparative analysis of different encoder architectures and data usage strategies under identical settings. Therefore, all models were trained and evaluated using the same data split and training parameters to ensure fairness in comparison. No cross-validation or multiple-run averaging was performed in this study. This limitation may affect the assessment of result variability and robustness. Future work will include repeated experiments with different random seeds and cross-validation strategies to provide a more comprehensive statistical evaluation of model performance.

The experimental studies were conducted on a computer equipped with an RTX 3080 Ti GPU, an Intel i5 processor, and 32 GB of RAM. The implementations were carried out using the Python programming language with the TensorFlow deep learning library.

## 2.5. Experimental Scenarios

Due to the imbalance between lesion-containing and healthy images in the dataset, three different experimental scenarios were designed to evaluate segmentation performance. The dataset contains a total of 22,782 CT slices, of which only 3,584 include lesions. For this reason, three scenarios were defined: segmentation using only lesion-containing images, segmentation using all available images, and segmentation with balanced data in the training set.

In the balanced data scenario, the class imbalance problem was addressed by applying an undersampling strategy on the training dataset. Specifically, the number of non-lesion (healthy) images was reduced to match the number of lesion-containing images. This approach ensured that both classes were equally represented during the training process.

No data augmentation or oversampling techniques were applied in this study in order to avoid introducing synthetic bias into the dataset. The test dataset was kept unchanged to reflect the real clinical distribution and to ensure a fair evaluation of model generalization performance.

In each scenario, model performance was evaluated using both image-based and patient-based evaluation approaches.

## 2.6. Performance Metrics

To evaluate the performance of the segmentation models, the F1-score and the Jaccard index were used. These metrics are commonly employed evaluation measures in medical image segmentation studies [29].

The F1-score represents the harmonic mean of precision and recall and is calculated as given in Equation (1).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1)$$

The Jaccard index measures the overlap between the predicted segmentation area and the ground truth annotation. This metric evaluates the similarity between the predicted region and the actual lesion region, as defined in Equation (2).

$$Jaccard = \frac{|Prediction \cap GroundTruth|}{|Prediction \cup GroundTruth|} \quad (2)$$

In the patient-based evaluation, predictions obtained from individual CT slices belonging to the same patient were aggregated to generate a single patient-level result. Specifically, the segmentation performance metrics (F1-score and Jaccard index) were first computed for each slice and then averaged across all slices of the corresponding patient. This approach provides a more comprehensive assessment of model performance by reflecting the overall segmentation quality at the patient level rather than relying solely on individual slice-based results.

Using these metrics, the segmentation performance results obtained in this study are presented in the following section in tabular form.

## 3. RESULTS

This section presents the results of the experimental studies conducted for lesion segmentation in lung CT images. The experimental analyses were conducted under three different data usage scenarios. Model performances were evaluated using two different evaluation approaches: image-based and patient-based. Segmentation performance was measured using the F1-score and Jaccard index metrics. Example CT slices from the dataset and the corresponding pixel-level annotation masks are shown in Figure 3.

### 3.1. Results Obtained Using Only Lesion-Containing Images

In the first experimental scenario, only CT slices containing lesions were used during both the training and testing stages. This approach was adopted to directly evaluate the model's ability to distinguish lesion regions from the surrounding lung parenchyma and other anatomical structures. Excluding non-lesion images from this scenario prevents the background class from becoming dominant during training, thereby enabling the model to focus more effectively on lesion regions. As a result, the model's capacity to learn the morphological characteristics of lesion tissues can be assessed more clearly.

This scenario is also important for examining the fundamental feature extraction capability of segmentation models. Since only lesion-containing images are used, the effect of class imbalance is largely eliminated, allowing a clearer evaluation of how accurately the models can segment lesion regions at the pixel level. The segmentation performance results obtained under this experimental scenario are presented in Table 3.

**Table 3.** Segmentation performance results obtained using only lesion-containing images

| Model | Image-Based F1 (%) | Image-Based Jaccard (%) | Patient-Based F1 (%) | Patient-Based Jaccard (%) |
|---|---|---|---|---|
| ResNet-UNet | 81.77 | 69.17 | 59.94 | 42.79 |
| VGG16-UNet | 80.44 | 67.28 | 56.47 | 39.34 |
| MobileNet-UNet | 79.55 | 66.04 | 52.00 | 35.13 |
| Inception-UNet | 82.25 | 69.85 | 60.57 | 43.45 |
| EfficientNet-UNet | 83.70 | 71.97 | 67.39 | 50.82 |
| U-Net | 77.12 | 62.76 | 62.08 | 45.01 |

The results presented in Table 3 indicate that all models achieved high segmentation performance in the image-based evaluation. This result can be explained by the use of only lesion-containing images, which allows the models to focus more effectively on lesion regions during the training process. In this scenario, the EfficientNet-UNet model achieved the highest performance with an F1-score of 83.70% and a Jaccard index of 71.97%. The EfficientNet architecture, which scales network depth, width, and input resolution in a balanced manner, may enhance segmentation performance by enabling the model to learn features at different scales more effectively.

Similarly, the Inception-UNet and ResNet-UNet models also demonstrated strong segmentation performance. The Inception architecture employs multi-scale filter structures that facilitate the simultaneous learning of structural features at different sizes, while the ResNet architecture utilizes residual connections that allow deeper networks to be trained more stably. In contrast, the MobileNet-UNet and the classical U-Net architectures achieved relatively lower performance compared to the other models. This difference may be attributed to their comparatively limited feature extraction capacity.

Compared with the image-based evaluation, a noticeable decrease in performance was observed for all models in the patient-based evaluation. The primary reason for this decline is that CT images from different patients may vary considerably in terms of anatomical structures, lesion size, location, and intensity distribution. Even if a model can achieve high segmentation accuracy on specific images, its generalization performance may decrease in patient-based evaluations due to the increased variability in the data. Nevertheless, the highest performance in the patient-based evaluation was still achieved by the EfficientNet-UNet model, with an F1-score of 67.39% and a Jaccard index of 50.82%.

Overall, this experimental scenario demonstrates that U-Net-based architectures with different encoder backbones are capable of accurately detecting lesion regions at the pixel level. However, the difference between image-based and patient-based evaluation results highlights the importance of assessing model performance not only at the image level but also at the patient level in medical image analysis.

### 3.2. Results Obtained Using All Images

In the second experimental scenario, segmentation experiments were conducted using all CT slices available in the dataset. This approach more closely reflects the real data distribution encountered in clinical practice. In clinical CT examinations, a substantial portion of the acquired slices typically do not contain lesions, while lesion-containing slices represent only a smaller fraction of the overall dataset. Therefore, this scenario is important for evaluating how the models perform under data distributions that closely resemble real clinical conditions. The segmentation performance results obtained in this scenario are presented in Table 4.

**Table 4.** Segmentation performance results obtained using all images

| Model | Image-Based F1 (%) | Image-Based Jaccard (%) | Patient-Based F1 (%) | Patient-Based Jaccard (%) |
|---|---|---|---|---|
| ResNet-UNet | 75.17 | 60.21 | 29.80 | 17.51 |
| VGG16-UNet | 65.06 | 48.21 | 25.41 | 14.55 |
| MobileNet-UNet | 65.88 | 49.12 | 33.77 | 20.31 |
| Inception-UNet | 77.94 | 63.86 | 42.93 | 27.33 |
| EfficientNet-UNet | 74.76 | 59.70 | 38.46 | 23.81 |
| U-Net | 46.99 | 30.71 | 38.31 | 23.69 |

As shown in Table 4, the segmentation performance in this scenario-where all images are used-is generally lower compared to the previous scenario. One of the main reasons for this decrease is the class distribution in the dataset. Since the number of slices without lesions is considerably higher than the number of lesion-containing slices, the model may tend to focus more on background tissue during the training process. As a result, learning the characteristics of lesion regions becomes more challenging for the model, which may lead to a decrease in segmentation performance.

In the image-based evaluation, the highest performance was achieved by the Inception-UNet model. This model demonstrated superior segmentation performance compared to the others, with an F1-score of 77.94% and a Jaccard index of 63.86%. The ResNet-UNet and EfficientNet-UNet models also achieved relatively high-performance values. In contrast, the classical U-Net architecture produced lower results compared to the other models.

In the patient-based evaluation, however, performance values were generally lower than those observed in the image-based results. The main reason for this decline is that CT images from different patients may exhibit significant variations in anatomical structure, lesion location, and lesion size. Even if a model performs well on specific images, the increased variability in patient-based evaluations can lead to lower overall performance. In this scenario, the highest patient-based F1-score was again achieved by the Inception-UNet model, with a value of 42.93%.

Overall, this scenario demonstrates that class imbalance within the dataset has a considerable impact on segmentation performance. These findings highlight that data distribution and data usage strategies play a critical role in determining model success in medical image segmentation studies.

### 3.3. Results Obtained Using Balanced Data

In the third experimental scenario, a data balancing approach was applied to reduce the effect of class imbalance present in the training dataset. In this context, the training dataset was arranged so that the number of lesion-containing and non-lesion images was equal. The test dataset was kept unchanged and used during the evaluation process to represent the real data distribution. This approach aims to enable the model to learn lesion and background classes more evenly during the training process. The segmentation performance results obtained in this scenario are presented in Table 5.

**Table 5.** Segmentation performance results obtained using balanced data

| Model | Image-Based F1 (%) | Image-Based Jaccard (%) | Patient-Based F1 (%) | Patient-Based Jaccard (%) |
|---|---|---|---|---|
| ResNet-UNet | 72.66 | 57.06 | 39.13 | 24.32 |
| VGG16-UNet | 70.94 | 54.97 | 44.12 | 28.31 |
| MobileNet-UNet | 69.14 | 52.84 | 41.85 | 26.46 |
| Inception-UNet | 76.23 | 61.60 | 45.38 | 29.35 |
| EfficientNet-UNet | 74.38 | 59.22 | 51.68 | 34.85 |
| U-Net | 70.69 | 54.67 | 41.70 | 26.34 |

Table 5 demonstrates that the data balancing strategy has a positive impact, particularly on patient-based evaluation results. While the performance values obtained in the image-based evaluation remain at levels similar to those obtained in the previous scenario, a noticeable improvement is observed in the patient-based evaluation. This finding suggests that balancing the class distribution in the training dataset may help the model produce more consistent predictions across data from different patients.

When the models are compared, the highest patient-based performance was achieved by the EfficientNet-UNet model, with an F1-score of 51.68% and a Jaccard index of 34.85%. The EfficientNet-based encoder structure, which can effectively learn multi-scale features, may contribute to more accurate segmentation of lesions with varying sizes and locations. The Inception-UNet and VGG16-UNet models also demonstrated relatively strong performance.

Overall, the results indicate that the data balancing approach has a positive effect on the generalization performance of segmentation models. The improvement observed particularly in patient-based evaluations highlights the importance of maintaining a balanced class distribution in the training dataset in medical image

segmentation studies. These findings demonstrate that the class distribution within a data set can directly influence model performance and emphasize the importance of developing appropriate data usage strategies.

### 3.4. Visual Segmentation Results

To better evaluate the performance of the segmentation models, visual analyses were conducted on several example CT slices. In order to obtain clinically more meaningful results, the evaluations were performed on a patient-based basis. The visual comparisons were carried out using the results obtained from the experimental scenario in which balanced data were used in the training dataset. Example segmentation outputs are presented in Figure 3.

In Figure 3, the first column shows the input CT images, where the ground truth lesion regions are indicated with red polygons. The remaining columns present the prediction outputs of different segmentation models, with the detected lesion regions illustrated using colored polygons.

In the first-row example, the nodule located in the posterior region of the upper lobe of the right lung was correctly detected by all models. The clear boundaries of the lesion and its distinguishable characteristics from the surrounding tissue contributed to this successful detection.

In the second-row example, a mass measuring approximately 4.3 cm located in the anterior segment of the upper lobe of the right lung is observed. However, some models were found to have difficulty detecting this lesion. This may be attributed to the lesion's proximity to the hilar region, where the dense presence of vascular and bronchial structures increases image complexity.

In the third-row example, there is a lesion located in the upper lobe of the left lung showing invasion toward the chest wall. In this case, only the ResNet-UNet and Inception-UNet models were able to correctly detect the lesion. The failure of the other models may be explained by the continuity of the lesion with the chest wall, which may have caused it to be confused with normal anatomical structures.

In the fourth-row image, a true nodule located in the upper lobe of the left lung is observed together with bone tissue extending from the anterior part of the first right rib toward the lung parenchyma. During the manual annotation process, the bone tissue was mistakenly labeled as a lesion. Nevertheless, all models except the ResNet-UNet correctly classified this region as normal tissue. In addition, the Inception-UNet model successfully detected the actual lesion.

In the fifth-row example, a consolidation area caused by an infectious process is observed. Although this structure does not represent a true tumor, some models partially marked it as a lesion. Other models, however, correctly interpreted this region as normal parenchymal tissue.

The sixth and seventh rows contain healthy CT slices without any lesions. The fact that none of the models detected lesions in these images indicates that the models were able to correctly learn and distinguish background tissue.

In the eighth-row example, a CT image obtained during a biopsy procedure and containing no lesions is presented. However, some models identified lesion-like structures in the hilar region. This may be attributed to the complex anatomical structures in the hilar region being interpreted by the model as pathological features.

Overall, the visual analyses demonstrate that deep learning-based segmentation models are generally capable of accurately detecting lesions in lung CT images. Nevertheless, complex anatomical structures in certain regions and infectious infiltrations may affect model performance. These findings suggest that artificial intelligence-based systems can serve as valuable tools to assist radiologists in the analysis of lung CT images.

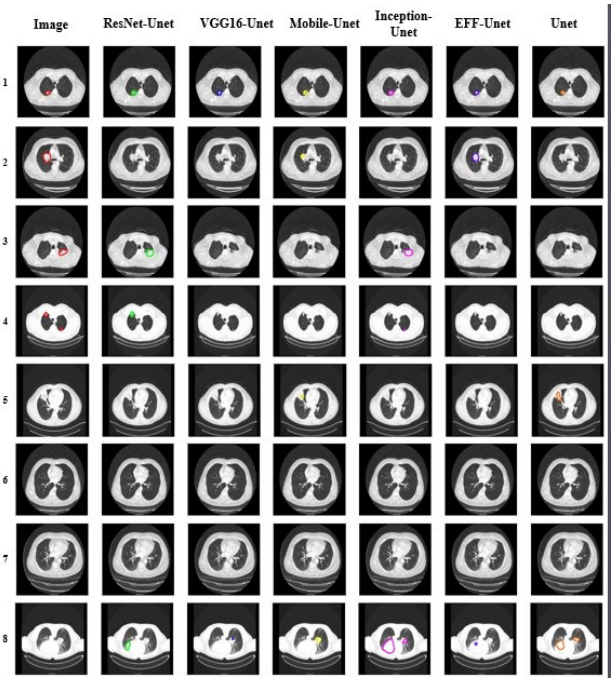



**Figure 3.** Example segmentation outputs obtained under the balanced data scenario.

## 4. DISCUSSION AND CONCLUSION

In this study, the performance of different deep learning-based models for lesion segmentation in lung CT images was investigated. The results indicate that segmentation performance is significantly influenced by the characteristics of the dataset. The dataset used in this study consists of 22,782 CT slices, of which only 3,584 contain lesions. This distribution reveals a pronounced class imbalance within the dataset. Such an imbalance led to a decrease in model performance, particularly in the scenario where all images were used together. During the training process, the models tended to focus more on background tissue, which may limit their ability to effectively distinguish lesion regions.

The experimental results reveal noticeable performance differences between image-based and patient-based evaluations. In the scenario where only lesion-containing images were used, image-based performance values were found to be relatively high, whereas lower results were obtained in patient-based evaluations. This finding suggests that although models can produce accurate predictions on specific images, their generalization capability may be limited when confronted with data distributions from different patients.

Balancing the number of lesion-containing and healthy images in the training dataset contributed to improved performance in patient-based evaluations. This result indicates that the class distribution within the dataset is a critical factor affecting the performance of segmentation models. In addition, among the compared architectures, the EfficientNet-UNet and Inception-UNet-based approaches achieved higher performance values. Nevertheless, segmentation was observed to be more challenging in certain anatomical regions-such as areas close to the pleura and around the hilum-as illustrated in the visual examples presented in Figure 3.

In addition, although the dataset includes lesions with varying sizes and characteristics, the absence of detailed subtype-level annotations (e.g., differentiation between nodules and masses) limits a more comprehensive analysis of model performance across different lesion types. Furthermore, although care was taken during the preprocessing stage, resizing operations may still have a minor impact on the visibility of very small lesions. These factors should be considered when interpreting the results.

In this study, the problem of lesion segmentation in lung computed tomography (CT) images was investigated, and the performance of different U-Net-based deep learning architectures was comparatively analyzed. A dataset consisting of 22,782 CT slices from 223 patients was used, and the impact of class imbalance in the dataset on segmentation performance was examined through three different experimental scenarios.

The results showed that image-based segmentation performance was high in the scenario where only lesion-containing images were used. However, a noticeable decrease in performance was observed in patient-based evaluations. In the scenario where all images were used together, segmentation performance declined due to class imbalance in the dataset. In contrast, when lesion-containing and healthy images were balanced in the training dataset, an improvement in performance-particularly in patient-based evaluations-was observed.

Overall, this study provides important findings regarding the evaluation of lesion segmentation in lung CT images from the perspectives of data distribution and model architecture. The obtained results are expected to contribute to determining appropriate data usage strategies in future studies on the automatic detection and segmentation of lung lesions.

In future work, it is planned to expand the dataset by including a larger number of patients and to investigate the effects of different deep learning architectures and attention mechanisms on segmentation performance. In addition, developing segmentation approaches that operate on three-dimensional (3D) CT data and exploring the integration of such models into clinical decision support systems are considered important future research directions.

**Acknowledgement**

This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the 1001 Scientific and Technological Research Projects Support Program (Project No: 125E062). This study was approved by the Scientific Research and Publication Ethics Committee of İnönü University (Date: 30.07.2024, Session No: 13, Decision No: 2024/6289).

## REFERENCES

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-63. doi:10.3322/caac.21834

[2] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 770-8. doi:10.1109/CVPR.2016.90

[3] Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances since the 2004 Classification. Journal of Thoracic Oncology. 2015;10(9):1243-60. doi:10.1097/JTO.0000000000000630

[4] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49. doi:10.3322/caac.21660

[5] Riquelme D, Akhloufi MA. Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans. AI. 2020;1(1):28-67. doi:10.3390/ai1010003

[6] Wang X, Wang L, Zheng P. SC-Dynamic R-CNN: A Self-Calibrated Dynamic R-CNN Model for Lung Cancer Lesion Detection. Comput Math Methods Med. 2022;2022:9452157. doi:10.1155/2022/9452157

[7] Hanbay K, Çalışan M, Özdemir TB. SAR Ship Detection Using Image Histograms and Machine Learning Approach. Turkish Journal of Nature and Science. 2024;13(3):171-175. doi:10.46810/tdfd.1528267

[8] Gu Y, Lu X, Yang L, Zhang B, Yu D, Zhao Y, et al. Automatic lung nodule detection using a 3D deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs. Comput

Biol Med. 2018;103:220-31. doi:10.1016/j.compbiomed.2018.10.011

[9] Lakshmanaprabu SK, Mohanty SN, Shankar K, Arunkumar N, Ramirez G. Optimal deep learning model for classification of lung cancer on CT images. Future Generation Computer Systems. 2019;92:374-82. doi:10.1016/j.future.2018.10.009

[10] Paing MP, Hamamoto K, Tungjitkusolmun S, Visitsattapongse S, Pintavirooj C. Automatic detection of pulmonary nodules using three-dimensional chain coding and optimized random forest. Applied Sciences. 2020;10(7). doi:10.3390/app10072346

[11] Perez G, Arbelaez P. Automated lung cancer diagnosis using three-dimensional convolutional neural networks. Med Biol Eng Comput. 2020;58(8):1803-15. doi:10.1007/s11517-020-02197-7

[12] Naqi SM, Sharif M, Jaffar A. Lung nodule detection and classification based on geometric fit in parametric form and deep learning. Neural Comput Appl. 2020;32(9):4629-47. doi:10.1007/s00521-018-3773-x

[13] Khan MA, Rubab S, Kashif A, Sharif MI, Muhammad N, Shah JH, et al. Lungs cancer classification from CT images: An integrated design of contrast based classical features fusion and selection. Pattern Recognit Lett. 2020 Jan 1;129:77-85. doi:10.1016/j.patrec.2019.11.014

[14] Suresh S, Mohan S. ROI-based feature learning for efficient true positive prediction using convolutional neural network for lung cancer diagnosis. Neural Comput Appl. 2020;32(20):15989-6009. doi:10.1007/s00521-020-04787-w

[15] Jain S, Indora S, Atal DK. Lung nodule segmentation using Salp Shuffled Shepherd Optimization Algorithm-based Generative Adversarial Network. Comput Biol Med. 2021;137:104811. doi:10.1016/j.compbiomed.2021.104811

[16] Shimazaki A, Ueda D, Choppin A, Yamamoto A, Honjo T, Shimahara Y, et al. Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. Sci Rep. 2022 Dec 1;12(1). doi:10.1038/s41598-021-04667-w PubMed PMID: 35031654.

[17] Swain AK, Swetapadma A, Rout JK, Balabantaray BK. Classification of non-small cell lung cancer types using sparse deep neural network features. Biomed Signal Process Control. 2024 Jan 1;87. doi:10.1016/j.bspc.2023.105485

[18] Xiao Z, Liu B, Geng L, Zhang F, Liu Y. Segmentation of lung nodules using improved 3D-UNet neural network. Symmetry (Basel). 2020;12(11):1-15. doi:10.3390/sym12111787

[19] Kumar V, Prabha C, Sharma P, Mittal N, Askar SS, Abouhawwash M. Unified deep learning models for enhanced lung cancer prediction with ResNet-50-101 and EfficientNet-B3 using DICOM images. BMC Med Imaging. 2024 Dec 1;24(1). doi:10.1186/s12880-024-01241-4 PubMed PMID: 38500083.

[20] Cifci MA. SegChaNet: A Novel Model for Lung Cancer Segmentation in CT Scans. Appl Bionics Biomech. 2022;2022. doi:10.1155/2022/1139587

[21] Alshayeji MH, Sa'ed Abed. Lung cancer classification and identification framework with automatic nodule segmentation screening using machine learning. Applied Intelligence. 2023;53:19724-41. doi:10.1007/s10489-023-04552-1

[22] Wu Y, Liu X, Shi Y, Chen X, Wang Z, Xu Y, et al. S3TU-Net: Structured Convolution and Superpixel Transformer for Lung Nodule Segmentation [Internet]. 2024 Nov 19. Available from: http://arxiv.org/abs/2411.12547

[23] Rikhari H, Baidya Kayal E, Ganguly S, Sasi A, Sharma S, Antony A, et al. Improving lung nodule segmentation in thoracic CT scans through the ensemble of 3D U-Net models. Int J Comput Assist Radiol Surg. 2024 Oct 1;19(10):2089-99. doi:10.1007/s11548-024-03222-y PubMed PMID: 39044036.

[24] Akila Agnes S, Arun Solomon A, Karthick K. Wavelet U-Net++ for accurate lung nodule segmentation in CT scans: Improving early detection and diagnosis of lung cancer. Biomed Signal Process Control. 2024 Jan 1;87. doi:10.1016/j.bspc.2023.105509

[25] Vijayakumar SR, Aarthy S, Deepa D, Suresh P. Sustainable framework for automated segmentation and prediction of lung cancer in CT image using CapsNet with U-net segmentation. Biomed Signal Process Control. 2025 Jan 1;99. doi:10.1016/j.bspc.2024.106873

[26] Setio AAA, Traverso A, de Bel T, Berens MSN, Bogaard C van den, Cerello P, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal. 2017 Dec 1;42:1-13. doi:10.1016/j.media.2017.06.015 PubMed PMID: 28732268.

[27] Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. Med Phys. 2011 Feb 24;38(2):915-31. doi:10.1118/1.3528204

[28] Survival of Patients with Stage I Lung Cancer Detected on CT Screening. New England Journal of Medicine. 2006 Oct 26;355(17):1763-71. doi:10.1056/NEJMoa060476

[29] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging. 2015 Aug 12;15(1). doi:10.1186/s12880-015-0068-x PubMed PMID: 26263899.