ORIGINAL ARTICLE

Automatic assessment of lung involvement in systemic sclerosis using deep learning

Matin Esnaashari[1], Roya Arian[2], Ali Hajihashemi[3], Narges Saeedizadeh[4], Somayeh Hajiahmadi[3], Somayeh Sadeghi[5], Azin Shayganfar[3], Rahele Kafieh[2]

[1]Al-Zahra Research Institute, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran, [2]Department of Engineering, Durham University, Durham, UK, [3]Department of Radiology, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran, [4]Institute for Intelligent Systems Research and Innovation, Deakin University, Geelong, Australia, [5]Department of Internal Medicine, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

Background: Systemic sclerosis (SSc) is a relatively uncommon connective tissue disorder, commonly manifesting as interstitial lung disease (ILD) and affecting both the lung parenchyma and the modification of the space between endothelium and epithelium. Imaging modalities like computed tomography (CT) scans are essential for diagnosing and revealing specific abnormal findings (ILD patterns) in SSc, such as reticulation and Ground-glass opacity (GGO). To enhance diagnostic precision and minimize human error, we leverage deep learning (DL) techniques. **Materials and Methods:** In our study, we collected and annotated a new public dataset from 22 individuals, encompassing 2190 lung CT scan slices. After preprocessing and exclusion of slices without abnormalities, 1777 slices from 17 patients were used for model training and validation, and 413 slices from five patients were reserved for independent testing. We use a specialized U-net model to segment these patterns, categorizing them into reticulation or GGO, and employ an automated algorithm to outline lung areas in each CT slice. The model's objective is to quantify the patient's lung involvement in SSc by calculating the total identified GGO and reticulation areas across all slices and normalizing this by the total lung surface area. **Results:** The U-net model shows promising results in segmenting both reticulation and a combination of GGO and reticulation, as indicated by Dice coefficients of 87.22% and 86.20%, respectively. Furthermore, the automated algorithm effectively outlines the lung region in each slice, enabling accurate measurement of lung involvement in SSc patients. **Conclusion:** In conclusion, using DL using the U-Net model and an automated algorithm has shown promising results in accurately segmenting and quantifying lung involvement in Scleroderma patients using CT scans.

Key words: Deep learning, ground-glass opacity, interstitial lung disease, lung involvement, reticulation, systemic sclerosis

How to cite this article: Esnaashari M, Arian R, Hajihashemi A, Saeedizadeh N, Hajiahmadi S, Sadeghi S, *et al*. Automatic assessment of lung involvement in systemic sclerosis using deep learning. J Res Med Sci 2026;31:10.

INTRODUCTION

Systemic Sclerosis (SSc) is a rheumatic disease resulting from the activation of the immune system, which targets multiple organs in the body.[1-3] This condition leads to fibrosis and damage to various organs, including the lungs.[4] In SSc, 55% of fatalities can be directly linked to damage caused by the immune system, with pulmonary fibrosis (35%), pulmonary hypertension (26%), and cardiac issues such as heart failure and arrhythmias (26%) being the primary causes. Additionally, non-SSc factors account for 45% of total deaths, including infections (33%) and malignancies (31%).[5] Advanced pulmonary fibrosis and pulmonary hypertension are significant risk factors for mortality.[6,7]

SSc is a rare connective tissue disease that can manifest as interstitial lung disease (ILD), among other symptoms.[5] ILD is a common manifestation of SSc and is a leading cause of death among patients with this condition.[8,9] ILD encompasses a range of lung conditions, spanning from rare lethal abnormalities to more common ones, all characterized by involvement and restructuring of the lung parenchyma between

Access this article online

Quick Response Code:

Website: https://journals.lww.com/jrms

DOI: 10.4103/jrms.jrms_994_25

Address for correspondence: Dr. Somayeh Hajiahmadi, Isfahan University of Medical Sciences, Hezar Jerib Avenue, Isfahan, Iran.
E-mail: som.hajiahmadi1372@gmail.com
Submitted: 08-Sep-2025; **Revised:** 07-Dec-2025; **Accepted:** 14-Dec-2025; **Published:** 26-Feb-2026

the endothelium and epithelium.[1] Despite comprising over 200 distinct abnormalities, the majority of patients typically exhibit symptoms like exertional dyspnea or persistent dry cough. Similar to radiomics-based machine learning (ML) frameworks used to classify lesion activity in Multiple Sclerosis from multi-sequence magnetic resonance imaging (MRI), our approach leverages high-dimensional imaging information to identify subtle parenchymal abnormalities in SSc-ILD.[10,11] However, some individuals with ILD may present non-respiratory systemic symptoms when seeking medical attention. Diagnostic methods for ILD often rely on imaging techniques such as chest X-rays (CXRs) and computed tomography (CT). Both imaging modalities aid in examining the lung parenchyma to evaluate the extent of lung involvement, facilitating accurate diagnosis and the development of appropriate treatment plans.[12]

In recent years, computer-aided diagnosis algorithms have significantly advanced the analysis of High-resolution CT (HRCT) patterns in diffuse lung diseases like SSc-ILD,[13] starting with machine-learning texture features introduced in 2010 to detect and quantify ground glass, quantify lung fibrosis (QLF), and honeycomb cysts.[14] Subsequent developments include Computer-aided Lung Informatics for Pathology Evaluation and Rating, utilizing computer-vision techniques for quantitative assessment of pulmonary parenchymal disease.[14-16] Radiomics-driven ML has also shown strong predictive value in oncology, such as in multiparametric MRI-based prediction of neoadjuvant chemotherapy response, reinforcing the clinical utility of quantitative imaging analytics.[17]

While AI models demonstrate proficiency in categorizing images of ILD, mere classification is insufficient. Pinpointing and annotating critical patterns like reticulation and ground-glass opacity (GGO) in ILD CT scans remain complex tasks for radiologists. The accuracy in identifying GGO varies among radiologists and largely depends on individual expertise.[18] Prompt and precise identification of these abnormal tissue alterations is crucial for early detection and reliable severity assessment. Manual marking by experienced radiologists, although thorough, is labor-intensive and subject to variations between and within observers, resulting in delays in CT scan evaluations. There is a recognized demand for automated segmentation of involvement indicators, and deep learning (DL) techniques present an excellent option for determining the extent of GGO and reticulation involvement in clinical-quality images.[19]

During the coronavirus disease 2019 (COVID-19) pandemic, significant efforts were made to apply DL tools for segmenting lung involvement, aiming to provide doctors with faster and more accurate diagnostic tools. A study by Antar *et al*.[20] employed a U-net model trained on COVID-19 CT scans from larxel, achieving impressive results with a Dice coefficient of 0.85 and accuracy rates between 99.2% and 99.5%. In[21] Osadebey and colleagues used a U-net model for lung region segmentation, achieving Dice scores ranging from 0.76 to 0.95. Recognizing that COVID-19 tissue regions tend to aggregate into cohesive units identifiable on individual CT slices, authors in[22] introduced a regularization term into the training loss function specifically designed to promote connectivity. This addition aims to encourage the model to prioritize identifying adequately large connected segmentation regions with desired characteristics.[23]

Automated and objective quantification of ILD patterns has the potential to support clinical decision-making by enabling earlier detection of disease progression, reducing inter-observer variability in severity scoring, and informing treatment decisions such as the timing of immunosuppressive or antifibrotic therapy. Integrating such automated tools into routine imaging workflows may also facilitate more consistent monitoring across follow-up visits, ultimately improving patient management in SSc-ILD.

In this study, we use a DL based framework designed for the automatic segmentation of GGO and reticulation areas from a clinical CT dataset, which we have made publicly available for other researchers. Utilizing the popular U-Net model, a DL medical image segmentation architecture, our approach focuses on the segmentation task, demonstrating good performance through quantitative validation. This research marks the application of the U-net model for segmenting such data and determining the percentage of lung involvement. This model is adept at not only segmenting ILD patterns in each slice but also calculating the percentage of lung involvement. This is achieved by initially employing a U-net to segment various patterns in each CT image, followed by the use of an automated algorithm to semantically segment the lung region in each slice. Finally, for each patient, we determine the extent of involvement by dividing the total area of all segmented patterns in all slices by the total area of all segmented lungs, providing a comprehensive assessment for each patient. Figure 1 illustrates an overview of our proposed method.

MATERIALS AND METHODS

Dataset

This study includes 2190 CT scan images collected from 22 individuals at Al-Zahra Hospital, Isfahan Medical University, Isfahan, Iran. Ethical approval for the study was obtained from the Ethics Committee of Isfahan University

of Medicine and the Iranian National Committee for Ethics and Biomedical Research, identified by the code IR.MUI.MED.REC.1400.476. Two radiologists participated in annotating the images: (1) a senior radiology resident and (2) a radiologist with more than 7 years of experience in lung imaging. Each radiologist independently labeled and annotated all the images, identifying primary patterns such as GGO and reticulations. If there were any discrepancies in their annotations, they resolved them through discussions to achieve consensus and ensure consistency in their interpretations, potentially involving a third senior radiologist to finalize decisions. Although the total number of slices is high, patient-level diversity is limited (22 patients in total), which may affect the generalizability and stability of the lung involvement estimates.

All demographic characteristics of patients' data are shown in Table 1.

Preprocessing

To standardize the data, all the original CT scans were normalized, adjusting each RGB color channel to a scale from 0 to 1. We eliminated images without any abnormalities to avoid false improvements in our analysis, as these would have empty masks. Each mask contains three pixel types: Background (representing the absence of abnormality), reticulation, and GGO. To speed up the training phase, the original CT scan images with their respective masks are resized to a uniform size of 128 × 128.

A total of 1130 images belong to the reticulation class, while 93 images fall under the GGO pattern class. Due to a class imbalance between reticulation and GGO images, an additional preprocessing technique is implemented. To address this issue, initially, a data augmentation method is leveraged on training images following the data splitting into training and validation sets. Specifically, all training

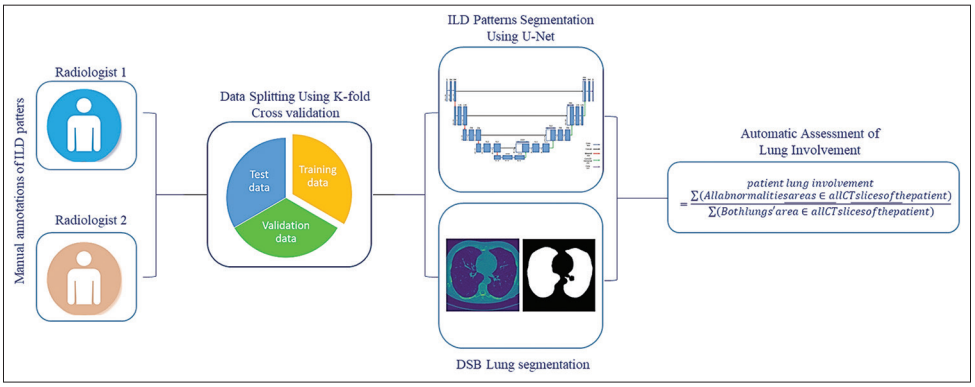

Figure 1: An overview of the proposed method. ILD: Interstitial lung disease

Table 1: Demographic characteristics of the data population

Parameter	Patients with reticulation	Patients with GGO	Explanation
Gender (male/female)	2/20	1/9	Higher concentration of females in both classes Significantly greater proportion of females among patients with reticulation
FEV_1 (median)	1.89	1.875	FEV_1 is nearly identical between the two classes Suggesting similar levels of airway obstruction
FVC (median)	2.21	2.275	FVC is slightly higher in the GGO class A slightly better overall lung capacity in this class
FEV_1 (%) (median)	77.5	63	Higher percentage of predicted FEV_1 in patients with reticulation Better preserved lung function compared to those with GGO
FVC (%) (median)	75.5	63.5	Higher FVC (%) in patients with reticulation. Less restrictive lung disease compared to those with GGO
FEV_1/FVC (median)	85.5	80.17	Higher FEV_1/FVC ratio in the reticulation class Less obstructive airway disease compared to those with GGO
DLCO (%) (median)	72	67.5	Higher DLCO in the reticulation class Better gas exchange efficiency compared to those with GGO
KCO (%) (median)	106	105	Similar K_{CO} in both classes Comparable efficiency in gas transfer per unit lung volume
TLC (median)	4.3	4.275	Similar TLC between the two classes Similar overall lung volume
TLC (%) (median)	85	67	Higher percentage of predicted TLC in the reticulation class Less restrictive lung disease compared to the GGO class

FEV_1: Forced expiratory volume in 1 s, FVC: Forced vital capacity, DLCO: Diffusing capacity of the lungs for carbon monoxide, K_{CO}: Transfer coefficient of the lung for carbon monoxide, TLC: Total lung capacity, GGO: Ground-glass opacity

GGO images undergo augmentation using two approaches to triple the image count: "Random rotation" to simulate the rotation and tilt of patients, and "Width and height shift" to mimic the shifting of images from center alignment along the horizontal and vertical axes. Augmentation was applied only to the training subset after the train–validation split, and never to validation or test images, to avoid data leakage. Only GGO slices were augmented in this way, which increased their effective number in the training set and improved the stability of GGO segmentation performance compared with training on the non-augmented data.

The final preprocessing step involves One-hot encoding. One-hot encoding entails breaking down the n values of a nominal variable into n distinct binary discrete variables, where 1 signifies a positive value, and 0 signifies a negative value for each of the n items.[24]

Nevertheless, since our goal is to conduct a 2-class segmentation for each abnormality, each scenario only involves two classes: Either it is background/reticulation or background/GGO. This results in our pixels being identified by two channels.

- Every pixel in the first channel equals one when it corresponds to the background class; otherwise, it is zero
- Each pixel in the second channel is set to one if it represents a reticulation/GGO class; otherwise, it is zero.

Figure 2 illustrates some samples from our dataset alongside their corresponding masks. These samples are randomly selected to provide enhanced visualization and comprehension of the dataset. When observed side by side, the GGO regions in the lungs are depicted as darker gray areas, while the reticulation regions, previously marked by the radiologist, are represented by white areas.

Train, validation, and test split and metrics of the segmentation model

In this research, the test dataset comprised CT slices from five patients (413 images), all of which were held out as unseen data. Splitting for model development was performed at the slice level, but patient assignment ensured that these 5 test patients did not appear in the training or validation folds. The remaining 17 patients were used for 5-fold cross-validation (CV). For each fold, approximately 13–14 patients were included in the training set and 3–4 patients in the validation set. The test data combines both GGO and reticulation patterns to accurately evaluate the real-world performance of our model. The remaining data undergoes random splitting into train and validation sets using k-fold CV with k set to 5. Employing k-fold CV is favored over random splitting for its completeness and generalization benefits. This approach ensures that

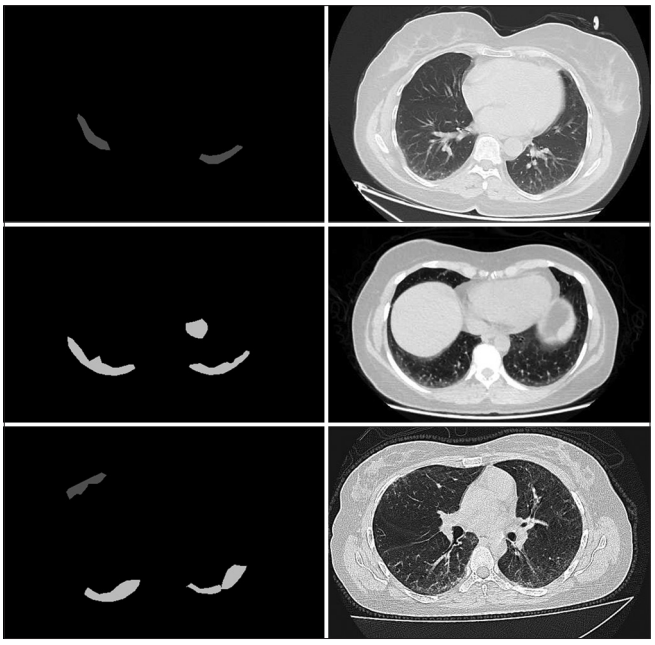

Figure 2: Three sample images from our datasets. The masks for each respective image are presented in the left column including ground-glass opacity (dark gray) and reticulation (white), while the right column showcases the original images

the system encounters the entire dataset for training, and both the training and test sets include each observation an equal number of times ($k - 1$ and 1 times, respectively). In contrast, a random split, with re-sampling at each iteration, may result in selecting duplicate members of the test set more than once. Consequently, *k*-fold CV is a preferable choice as it guarantees the comprehensive use of the entire dataset during model training and maintains an equal representation of each observation in both training and testing phases. In contrast, random splitting may lead to duplications in the test set due to resampling in each iteration.

Interstitial lung disease patterns segmentation

The U-Net model, initially introduced by Ronneberger *et al*., stands out as one of the widely adopted segmentation models in medical imaging. It relies on an encoder-decoder architecture and incorporates skip connections. Skip connections in the U-Net model are utilized to overcome the issue of information loss that occurs during the processes of down-sampling and up-sampling in tasks related to image segmentation.[25]

Our proposed U-net structure consists of three main parts:

- Encoding: A set of four blocks, each consisting of two CNN layers and batch normalization (BM). As we progress through each layer, the number of filters doubles incrementally. The initial filter size utilized in this study is 32. For each CNN, a rectified linear unit (ReLU) and a 2 × 2 max pooling operation with a stride of 2 are applied for down-sampling

- Bottleneck: Comprising the initial filters multiplied by 128, the total filter size is 512
- Decoding: A set of four blocks where up-sampling operators replace pooling operations. Each step in the expansive path involves up-sampling the feature map, followed by a 2 × 2 convolution (up-convolution) that reduces the number of feature channels by half, concatenation with the correspondingly cropped feature map from the contracting path, and two 3 × 3 convolutions, each succeeded by a ReLU. Cropping is essential due to the loss of border pixels in each convolution. In the final layer, a 1 × 1 convolution is utilized to map each 64-component feature vector to the desired number of classes.

An overview of the U-net model is represented in Figure 3.

Reticulation segmentation

In this study, for segmentation of the reticulation dataset, a learning rate of 10^{-2} was utilized along with an Adam optimizer. Regarding callbacks, the Early Stopping callback (with a patience of 20),[26] the Reduce Learning Rate on Plateau callback (with a minimum learning rate of 1.0–8, a factor of 0.1, and a patience of 10),[27] and the Model Checkpoint are utilized. The Model Checkpoint serves the purpose of preventing overfitting and saving the best possible model based on validation subset losses. Furthermore, the batch size and the number of epochs are set to 8 and 300, respectively. These hyperparameters are determined empirically to achieve optimal loss and minimize errors during the training phase. However, the high learning rate proved unsuitable for effective training due to the limited and diverse nature of the dataset. Hyperparameters (learning rate 10^{-2}, batch size 8, 300 training epochs) were selected after a sensitivity analysis on a held-out validation subset, in which we varied the learning rate, batch size, and number of epochs. Lower learning rates led to slower convergence without clear gains in segmentation accuracy, whereas higher learning rates increased instability in training loss. Similarly, larger batch sizes slightly reduced performance, likely due to less frequent weight updates on this relatively small dataset. The chosen configuration provided a balance between convergence speed, stability, and overfitting control. In addition, Dice loss has been selected as the loss function for the proposed model. When the predefined mask and the predicted mask from our model exhibit greater similarity, the Dice score increases and approaches 1, indicating total similarity between both masks. Conversely, in the context of the loss function, we aim to assess the dissimilarity between these two images. Therefore, weights and biases can be optimized based on the Dice loss, calculated as (1 - Dice score). Because the dataset contained substantially more reticulation images than GGO images, class imbalance was mitigated through two mechanisms: (1) augmentation of GGO slices to expand their representation in the training set, and (2) use of Dice loss, which is inherently robust to imbalance by emphasizing overlap of minority-class pixels rather than raw pixel frequency.

Ground-glass opacity segmentation

As previously mentioned, the number of GGO images is lower than the number of reticulation images. Consequently, we exclusively utilize the reticulation dataset for the training phase and later fine-tune the model for the GGO images.

Given the satisfactory performance of our model trained on the reticulation dataset, we opt to reuse the weights of

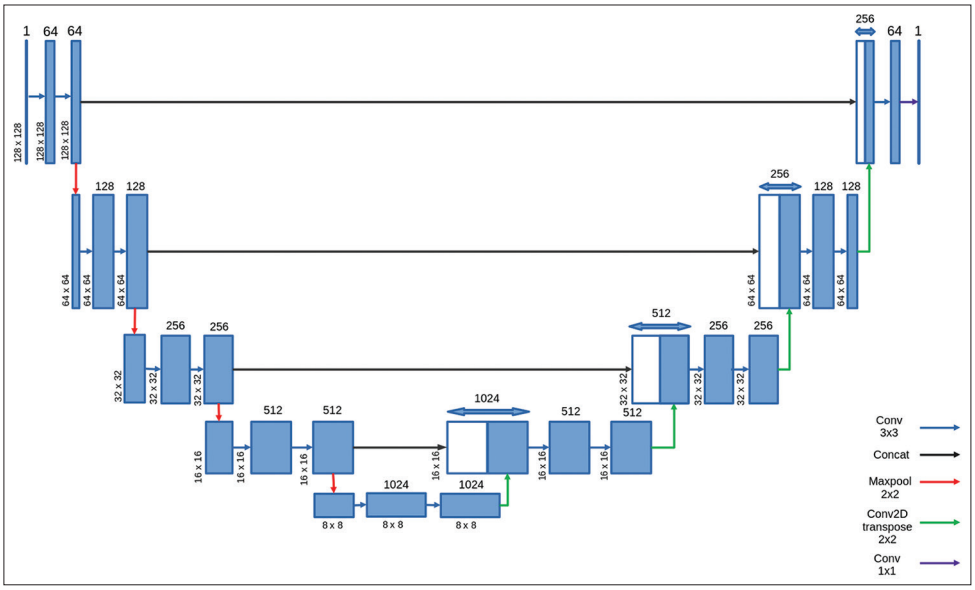

Figure 3: An overview of U-net model structure. Number of each filter and filter sizes are represented in the diagram[21]

the encoded pathway in this model for our GGO dataset. All encoding layers are frozen, rendering the weights of these layers immutable and preventing them from being trained during the training phase on the GGO dataset. Freezing all encoder layers was empirically found to be the most effective strategy for the limited GGO dataset, as it allowed us to leverage the reticulation-trained feature extractor while adapting only the decoder to the GGO appearance. We also experimented with unfreezing subsets of encoder layers (e.g., only the deepest or only the last two encoding blocks), but these alternatives resulted in lower Dice scores and clear signs of overfitting on the small GGO subset, confirming that full encoder freezing offered the best trade-off between flexibility and generalization. Only the decoder pathway undergoes training during this phase (Note that we have explored various numbers of frozen layers for fine-tuning, and the mentioned strategy achieves the optimal performance during the training phase).

The model structure and compilation remain consistent, employing the same optimization and loss functions as mentioned earlier for the reticulation dataset. With the first four layers locked, the total number of parameters is outlined in Table 2.

Segmentation of merged patterns

Another approach is to merge both ILD patterns into a single class for segmentation, regardless of whether they present as GGO or reticulation, and then perform segmentation on this aggregated class. This approach allows us to determine the extent of SSc impact on both lungs. The model is trained from scratch, employing the same architecture without using any transfer learning techniques. Moreover, the chosen hyperparameters are identical to those selected for the segmentation of reticulation.

Post processing

Before generating the masks, there is a crucial step to remove artifacts and noise. For this purpose, a threshold is applied to classify pixels with values higher than the threshold as one and others as zero. To determine the optimal threshold value, a function is developed to iterate through different cutoffs, seeking the best value. The iterative process focuses on finding the optimal threshold for each fold, not each individual image. Initially, we set the threshold at 0.1. Subsequently, in each iteration, the value increases by 0.02 until the threshold reaches 1. This threshold optimization was performed separately for each fold on the validation data and then applied to the corresponding test predictions for that fold. The goal of this post-processing step was to suppress low-confidence activations and small, noisy regions, thereby improving the precision of the predicted masks while preserving high-confidence lesion regions. If the Dice score increases in any iteration, the new threshold value is saved. This process is repeated for every fold until optimal thresholds are identified for each one.

Table 2: A summary of parameters after freezing encoding layers

Parameters type	Numbers
Total	4,708,354
Trainable	184,770
Non-trainable	4,523,584

Lung segmentation

Subsequent to that, we employed an algorithm proposed by K Scott Mader[28] to segment the lung area within each CT scan image. The primary segmentation algorithm encompasses eight steps:

a. Conversion of the image into binary values
b. Removal of all blobs connected to the image border
c. Labeling of the image using the label function from the scikit-image measure function
d. Retention of the largest and second-largest areas of the labeled image, representing two areas of the lung surface
e. Separation of lung nodules, arteries, and vessels, which may be attached in a CT scan image, using scikit's erosion function with a radius of 2
f. Preservation of nodules attached to the lungs by employing the closure function with a radius of 10
g. Filling of small holes with lung true values
h. The superimposition of the previously created binary mask.

Lung involvement

Or precise computation of the overall percentage of lung involvement, it is necessary to aggregate the areas of both pattern (merged pattern) regions across all CT slices for each patient. Subsequently, this summation should be divided by the total area of all segmented lungs across slices for that particular patient to determine the degree of involvement. The following formula can be utilized to calculate the lung involvement for each patient:

$$\text{patinent lung involvement} = \frac{\sum\left(\begin{array}{c}\text{All abnormalities areas}\\ \in \text{all CT slices of the patient}\end{array}\right)}{\sum\left(\begin{array}{c}\text{Both lungs' area}\\ \in \text{all CT slices of the patient}\end{array}\right)} \quad (1)$$

All implementations were performed using Python 3.11 with libraries including TensorFlow/Keras for the U-Net model, scikit-image for lung segmentation, and NumPy/SciPy for preprocessing and metrics.

RESULTS

The assessment of the proposed U-Net for segmenting either individual patterns or merged patterns is conducted using Dice and Jaccard coefficients, F1-score, and Recall metrics.

First, the predicted masks of each pattern are compared against manually generated masks, and the Dice similarity coefficient (DSC) is computed using the following formula:

$$Dice = \frac{2(A \cap B)}{|A| + |B|} \quad (2)$$

Here, A and B represent the predicted and ground truth masks, respectively. To provide an accurate Dice coefficient focused on the area of interest, considering that the area of the considering pattern is significantly smaller than the background, the DSC value is specifically reported for the segmented areas rather than the background.

Subsequently, we employ the Jaccard Similarity Index,[29,31] F1-score, and recall metrics for model evaluation. The Jaccard Similarity Index serves as a measure of similarity between two sets, determined by dividing the number of elements in the intersection of the sets by the number of elements in their union. In simpler terms, it represents the ratio of the intersection size to the union size of the two sets. The formula for the coefficient is as follows:

$$J(A,B) = \frac{A \cap B}{A \cup B} \quad (3)$$

Where A and B represent two sets. The Jaccard Similarity Index between them equals 1 when the two datasets share identical members. Conversely, if they have no common members, the result would be zero.

F1-score and recall can also be calculated as follows:

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (4)$$

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

Where TP represents true positive, FN and FP represent false negative and false positive, respectively.

All evaluation metrics were computed at the pixel level. For each image, pixel-wise Dice, Jaccard, F1-score, and Recall were calculated, then averaged across all slices within a fold. Final reported values represent the mean of all five folds.

The average results across all five folds on validation data are presented in Table 3. Only cross-validated mean values were stored from the original experiments, so standard deviations across folds could not be recovered retrospectively.

In addition, Figure 4 presents the confusion matrix of the proposed U-Net for segmenting Reticulation + GGO, demonstrating promising results for both the background (0) and Reticulation + GGO (1) classes. The confusion matrix was computed at the pixel level, where each pixel was assigned to either the background or lesion class. Values represent aggregated pixel predictions across all slices in the test set.

Moreover, the original CT images are annotated with both the actual and predicted masks for comprehensive analysis and comparison. Figures 5 and 6 provide a visual comparison of the true and predicted masks for GGO and reticulation, respectively. In addition to the representative successful examples shown in Figures 5 and 6, during model development, we observed several recurrent failure modes, particularly for GGO segmentation. The model tended to under-segment very faint or low-contrast GGO regions, especially at the lung periphery, and occasionally over-segmented noisy areas where vascular or artifact structures partially resembled ILD patterns. These qualitative observations are consistent with the lower Dice scores obtained for GGO compared with reticulation and highlight the inherent difficulty of reliably capturing very subtle parenchymal changes. We also compared our results to several state-of-the-art (SOTA) models, as shown in Table 4. The evaluation was conducted using the GGO + reticulation dataset to generate mask predictions for the test data. All SOTA models in Table 4 were trained under identical conditions to our proposed U-Net, including the same preprocessing pipeline, data splits, augmentation strategy, loss function, and optimization settings. This design ensures that performance differences primarily reflect architectural characteristics rather than changes in training protocol.

Subsequently, the automatic lung segmentation algorithm is employed to calculate the area of the lung in every CT image. Figure 7 illustrates an example of this segmentation algorithm being applied to an image and compares it

Table 3: The evaluation metrics uses to assess the segmentations obtained from the proposed U-Net

Different classes	Dice	Jaccard	F1-score	Recall
Reticulation	0.87	0.93	0.72	0.74
GGO	0.54	0.93	0.77	0.78
Reticulation + GGO	0.87	0.93	0.93	0.94
Reticulation + background	0.86	0.92	0.93	0.93
GGO + background	0.54	0.93	0.75	0.75

The average of all folds on validation data has been reported. GGO: Ground-glass opacity

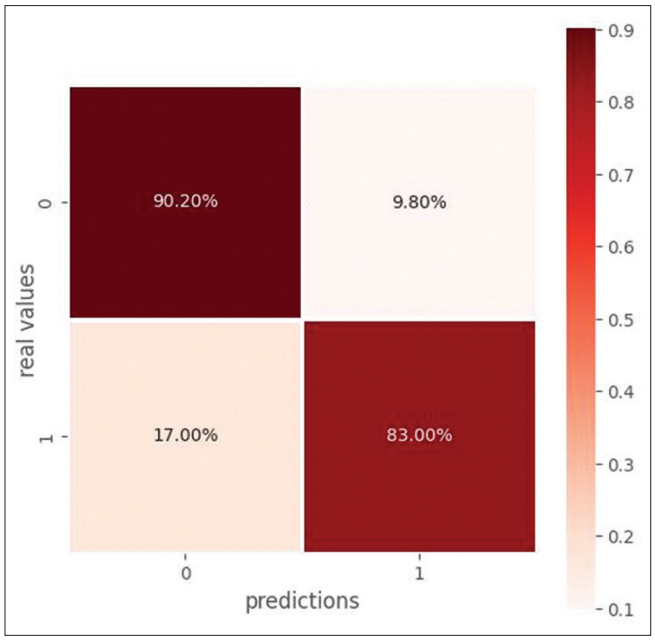

Figure 4: The confusion matrix of the proposed U-Net for segmenting both the background (0) and reticulation + ground-glass opacity (1) classes

with the true lung mask marked by an expert radiologist. Moreover, Table 5 shows precision, recall, and F1-score for true and false pixels.

It is evident that the lung surfaces generated by the computer and the radiologist are nearly identical. The shortcomings of human marking include reduced accuracy and a tendency to over- or underestimate, whereas algorithms are unaffected by these issues. Furthermore, the automated algorithm significantly expedites and simplifies the process, particularly when dealing with a large number of CT slices.

Ultimately, lung involvement is determined for each case using the method described in section 2.6. These results are then compared to the estimates provided by our radiologist in Table 6. Furthermore, Figure 8 displays the Bland–Altman plot, which offers a clear representation of the findings.[30] Bland–Altman plots, widely used in medical research, serve as robust visual tools for comparing two measurement methods and evaluating the concordance between two datasets. This plot is useful for assessing the level of agreement between two

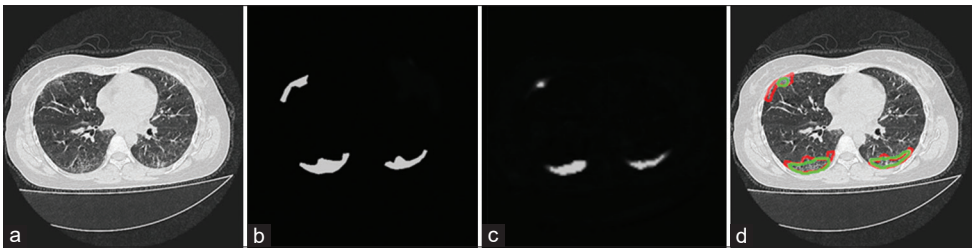

Figure 5: Visual comparisons between the ground-glass opacity true masks and the predicted ones throughout the proposed U-Net. Where (a-d) represent the original computed tomography (CT)-scan, the true mask labeled by the radiologist, the predicted mask generated by the model, and the original CT-scan image annotated with borders of both the true and predicted masks. It is important to highlight that the borders of the true mask are displayed in red, while the borders of the predicted mask are displayed in green

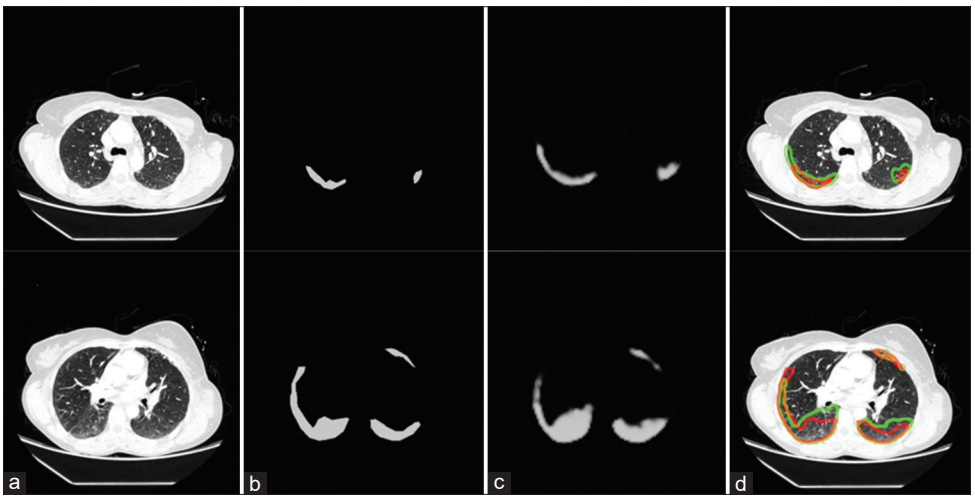

Figure 6: Comparison between the reticulation true masks and the predicted ones throughout the proposed U-Net for two samples. Where Columns (a-d) represent the original computed tomography (CT) scan images, the true masks labeled by the radiologist, the predicted masks generated by the model, and the original CT-scan image annotated with borders of both the true and predicted masks. It is important to highlight that the borders of the true mask are displayed in red, while the borders of the predicted mask are displayed in green

assessors and detecting any consistent patterns of bias. It depicts the disparity between two measurements along the y-axis and the average of the two measurements along the x-axis. Clearly, our automated approach demonstrates strong performance, yielding results closely aligned with radiologist predictions. Quantitatively, the Bland–Altman analysis demonstrated a small mean bias in lung involvement estimates (model minus radiologist) with relatively narrow limits of agreement, indicating good concordance between the automated method and the expert readings. The observed bias was modest compared with the overall clinical range of lung involvement values, suggesting that systematic differences between the two approaches are limited and unlikely to be clinically critical.

We also evaluate the effectiveness of our proposed model in calculating lung involvement, employing our U-Net architecture and other SOTA models for segmenting lung patterns. Table 7 illustrates the extent of lung involvement for every set of five test patients using either U-Net or other SOTA models.

DISCUSSION

Our proposed DL approach, based on the U-net architecture for segmenting abnormalities, demonstrates promising results, particularly excelling in the segmentation of the combination of reticulation and GGO. The precision is comparable to that of experienced human experts in identifying SSc findings. Although the predicted segmented areas exhibit a slight offset compared to true masks, they precisely encapsulate the abnormal areas with minimal false positives. Dice scores are consistently high, ranging from 0.86 to 0.87, except for segmenting only GGO regions, where we have limited labeled images. This suggests substantial similarity between masks marked by radiologists and masks predicted by the proposed U-net. While the model does not surpass human expert performance, it demonstrates radiologist-like proficiency in the segmentation task. Although it may not excel in GGO segmentation, it's worth noting that GGO patterns pose a challenge even for experienced radiologists or pulmonologists. The average Dice score for the GGO model is 0.539, indicating that the model performed less accurately than the reticulation model but still outperforms the baseline. Increasing the number of images with GGO labels in our dataset, following standardized procedures and thorough preprocessing, will significantly enhance the accuracy of our data-hungry model. This improvement could potentially surpass human ability to recognize this finding.

In a comparable study conducted by Guillaume Chassagnon *et al.*, the U-net model demonstrated an average DSC of 0.72 with respect to expert radiologist number one and 0.73 with respect to radiologist number two.[33] In our study, which specifically concentrates on the joint evaluation of reticulation and GGO labels, our model attains a mean dice score of 0.86. This result suggests that our model accurately ILD, segmenting it with a high level of precision, aligning closely with the outcomes observed in the aforementioned study.

In a notable study by Pang *et al.*, the final modified model achieved a Dice coefficient of 89.42% for lung segmentation on HRCT images, while its U-net model obtained a Dice coefficient score of 85.67%.[32] Our model not only achieves Dice scores ranging from 86% to 87% for reticulation and GGO but also attains an average precision score of 98% for segmenting lungs using solely algorithmic approaches.

Table 4: Comparison between different encoders used within U-Net model on test data

Model name	Dice	Jaccard	F1-score	Recall
Our model	0.804963	0.926977	0.931622	0.900550
InceptionV3	0.745759	0.887742	0.799114	0.746555
Resnet34	0.992402	0.985032	0.502760	0.503353
VGG16	0.862648	0.905859	0.927085	0.900514
VGG19	0.771556	0.844055	0.87055	0.832278

Table 5: Lungs segmentation results based on 3 different evaluation methods for True or false pixel values

Evaluation methods	Precision	Recall	F1-score
False	0.99	0.99	0.99
True	0.96	0.96	0.96
Average/total	0.98	0.98	0.98

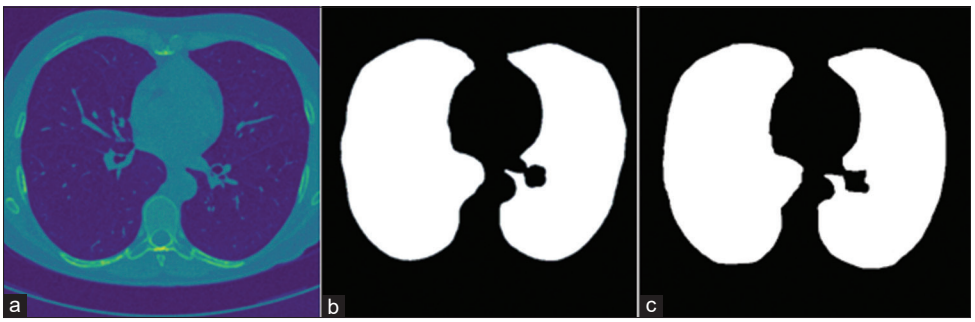

Figure 7: Comparison between human and computer-based findings for lung regions. Where (a-c) represent the original, the lung mask generated by the algorithm, and the lung mask marked by the radiologist

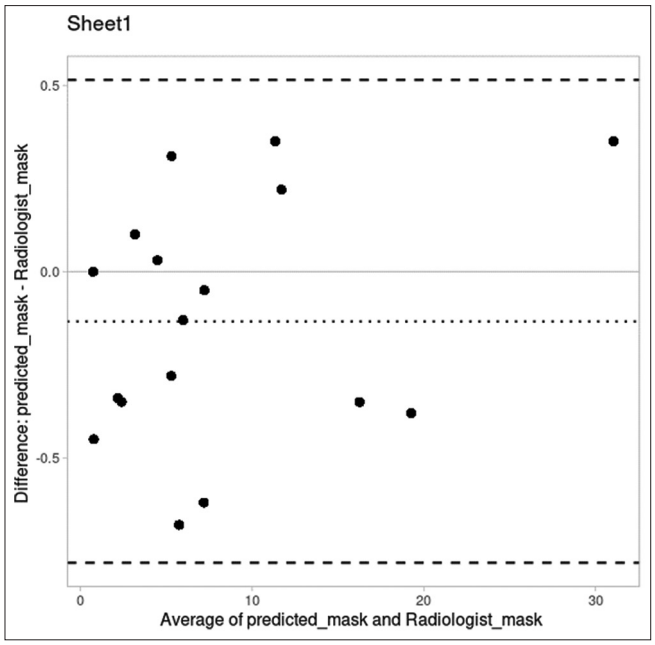

Figure 8: Bland–Altman analysis for comparing outcomes produced by the deep learning model and scores reported by the radiologist

Table 6: Extent of lung involvement in every training patient with systemic sclerosis

Patient number	Involvement calculated by computer (%)	Involvement based on radiologist segmentation (%)
1	5.47	5.16
2	16.09	16.44
3	3.22	3.12
4	2.01	2.35
5	11.52	11.17
6	19.07	19.45
7	31.24	30.89
8	7.19	7.24
9	5.40	6.08
10	0.74	0.74
11	4.50	4.47
12	5.90	6.03
13	6.88	7.50
14	2.22	2.57
15	11.82	11.60
16	5.15	5.43
17	0.55	1.00

Another relevant study by Gang Li *et al*. demonstrated that employing a multi-scale cross-layer attention fusion network for honeycombing segmentation could achieve an Intersection over Union (IoU) of 89.5%, mean IoU of 92.1%, and mean Dice coefficient of 94.9%.[34] Although the targeted patterns were different, their model achieved higher average Dice coefficient and IoU. However, it could only detect single-entity pattern. In contrast, our model could identify two entities and distinguish them in a single CT scan image.

Our U-net outcomes, in comparison to other research, were already satisfactory and skillfully executed; however, we went a step further by employing an algorithmic method to delineate lung regions and subsequently calculating the lung involvement. While segmentation is a crucial task, assessing lung involvement in an objective manner, immune to person-to-person perception variability, will help physicians accurately evaluate a patient's status. This is how our model thrived among similar models.

In addition, we compared our U-net with other SOTA models using our dataset, as shown in Table 4, which highlights the performance and accuracy of our model relative to these methods. Although our model outperforms others in terms of F1-score and Recall metrics, ResNet34 shows better results in Jaccard and Dice coefficients. However, its F1 score and Recall are notably low. The high Dice coefficient but relatively low F1-score observed in models such as ResNet34 can be explained by over-prediction of abnormal regions. Dice captures spatial overlap, so even a broad mask with excessive predicted area may achieve a high Dice if it overlaps the ground truth. In contrast, the F1-score is strongly penalized by false positives, so over-segmentation lowers the F1-score despite a seemingly good Dice score. When analyzing the predicted pattern borders using ResNet34, it becomes evident that while it detects patterns effectively, it also misidentifies many regions as patterns that are not, thus reducing the reliability of ResNet34 in accurate prediction. Furthermore, Table 7 illustrates that our proposed U-Net model achieves the lowest mean absolute error in calculating lung involvement compared to other models. These extensive comparisons illustrate

Table 7: Extent of lung involvement in every test patient with systemic sclerosis using our novel method for automatic calculating lung involvement

Model	Patient 1 (%)	Patient 2 (%)	Patient 3 (%)	Patient 4 (%)	Patient 5 (%)	Mean absolute error (%)
Ground truth	2.09	1.1	6.48	1.37	5.72	-
Our model	2.52	1.19	6.92	1.41	5.98	0.252
Resnet34	2.48	1.32	6.98	1.59	6.65	0.452
InceptionV3	2.89	1.56	7.56	1.98	6.89	0.824
VGG16	2.91	1.48	7.1	1.72	6.21	0.532
VGG19	2.61	1.21	6.88	1.48	5.99	0.282

This assessment utilizes our proposed U-Net model or other SOTA models for segmentation of patterns within the lung. The final column displays the mean absolute errors between the method employed and the ground truth values provided by radiologists. SOTA: State-of-the-art

the robustness and effectiveness of our model compared to established SOTA techniques, thus enhancing the validity of our findings.

In the context of lung segmentation in CT scan images, the utilization of non-DL algorithms, such as the K Madet Scott's algorithm, has demonstrated remarkable efficacy.[32] This algorithm has proven to be highly accurate, capable of identifying and delineating the surfaces of both lungs with nearly 100% accuracy across nearly all CT scan images. Importantly, it achieved this without the need for time-consuming training sessions and operated efficiently using only CPU resources, ensuring rapid execution. With an impressive F1-score of 98%, the model is suitable for implementation in mobile applications, running seamlessly on resource-limited devices without experiencing any lag or sluggish performance.

Both results, namely the segmentation of ILD patterns and the identification of lung regions, offer a precise estimation of the extent of involvement in both lungs. Clinicians can utilize this information to obtain an objective overview of the lungs, aiding in the assessment of disease progression and prognosis. Clinically, such automated and objective quantification of SSc-ILD involvement has the potential to reduce inter-observer variability in visual scoring, accelerate the estimation of lung involvement for treatment decisions (for example, initiation or escalation of antifibrotic or immunosuppressive therapy), and facilitate more consistent serial monitoring of disease progression during follow-up. As demonstrated in Tables 6, 7, and Figure 8, our automated method for lung involvement calculation performs effectively, approaching optimal results. Therefore, radiologists can depend on these automated algorithms to expedite their diagnostic and analytical processes. A recent systematic review also emphasizes that SSc is strongly associated with cardiovascular and atherosclerotic risk, and highlights the value of AI-based imaging for detecting subtle vascular abnormalities – supporting our findings regarding the broader role of automated image analysis in SSc management.[23] Our model demonstrates promising results on unseen data, outperforming SOTA models despite the data-intensive nature of DL methods, showcasing its robustness and potential impact. However, the study faced several limitations. Although the favorable results of the proposed U-net model were validated with an independent dataset, the training, validation, and testing images were sourced from a single center and derived from a limited number of patients, primarily those with GGO patterns. This constraint prevented us from obtaining and processing a more extensive set of images for our DL U-net model. While the model performed well overall, additional data would have further optimized its capabilities. Moreover, we faced a shortage of publicly accessible datasets relevant to our research, limiting our ability to access additional data and forcing us to rely solely on our collected data. In addition, our center specializes mainly in GGO and reticulation patterns, limiting our access to data on other ILD patterns and restricting our ability to include a wider range of abnormalities in our study.

These limitations underscore the critical need for future research to address the challenges associated with dataset size and diversity in medical image analysis. Collaborative initiatives across multiple centers and enhanced access to comprehensive, publicly available datasets are essential for advancing the development and validation of segmentation algorithms in clinical practice.

CONCLUSION

Our DL approach, centered on the U-net architecture for segmentation, has shown promising outcomes, particularly excelling in the segmentation of combined reticulation and GGO. The precision demonstrated is comparable to experienced human experts in identifying SSc findings. In addition, our study contributes to the field by comparing favorably with prior research and introducing an automatic algorithmic method for the calculation of lung involvement. Automated segmentation of lung abnormalities and the objective determination of the percentage of lung involvement prove to be advantageous tools for the radiologists and pulmonologists. By employing U-net and algorithmic approaches, we can avoid the subjectivity inherent in methods that may vary widely and carry a substantial margin of error. This approach enables a more precise estimation of lung involvement with reduced variation among physicians, leading to a more accurate assessment and an increased likelihood of identifying ILD in its early stages. Ultimately, this contributes to an improved patient prognosis, reducing the occurrence of complications.

In our future endeavors, we can explore various segmentation methods beyond the initially proposed U-net model. Moreover, we aim to enhance the robustness and reliability of our findings by expanding our dataset, particularly by including data from a larger cohort of patients, with a specific emphasis on those presenting GGO and other ILD patterns. This comprehensive approach is expected to contribute to a more nuanced and accurate assessment of our methodologies, ultimately improving the overall quality and validity of our results. These findings are encouraging, although they should be interpreted with the awareness that our dataset was relatively small and derived from a single center, which may limit generalizability. In addition, only cross-validated mean segmentation metrics were retained from the original experiments, so we were unable to provide fold-wise standard deviations

or confidence intervals, which we acknowledge as a limitation of the current analysis. Future work could further expand this framework toward prognostic prediction by incorporating longitudinal imaging and clinical data, allowing automated models to estimate future disease progression and patient outcomes.

Ethics approval and consent to participate

Ethical approval for the study was obtained from the Ethics Committee of Isfahan University of Medicine and the Iranian National Committee for Ethics and Biomedical Research, identified by the code IR.MUI.MED.REC.1400.476.

Consent for publication

Written informed consent was obtained from the patients for publication of this article and any accompanying images. A copy of the written consent is available for review by the Editor-in-Chief of this journal.

Availability of data and materials

All data and materials are available from the corresponding author upon request.

Authors' contributions

M.E. and R.A. implemented the final method. R.A., M.E., and N.S. wrote the main manuscript. A.H., S.H., and A.S. reviewed and annotated all images. A.H., N.S., S.H., S.S., and R.K. contributed significantly on data pre-processing and design of the paper. R.K., R.A., N.S., and A.H. improved the initial method and discussed about the experimental results.

All authors contributed to the conception of the study and reviewed the paper.

Acknowledgments

This research did not receive any specific grant from any funding agency in the public, commercial, or not-for-profit sector.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Antoniou KM, Margaritopoulos GA, Tomassetti S, Bonella F, Costabel U, Poletti V. Interstitial lung disease. Eur Respir Rev 2014;23:40-54.
2. Volkmann ER, Andréasson K, Smith V. Systemic sclerosis. Lancet 2023;401:304-18.
3. Pope JE, Denton CP, Johnson SR, Fernandez-Codina A, Hudson M, Nevskaya T. State-of-the-art evidence in the treatment of systemic sclerosis. Nat Rev Rheumatol 2023;19:212-26.
4. Hinchcliff M, Varga J. Systemic sclerosis/scleroderma: A treatable multisystem disease. Am Fam Physician 2008;78:961-8.
5. Tyndall AJ, Bannert B, Vonk M, Airò P, Cozzi F, Carreira PE, *et al*. Causes and risk factors for death in systemic sclerosis: A study from the EULAR scleroderma trials and research (EUSTAR) database. Ann Rheum Dis 2010;69:1809-15.
6. Foocharoen C, Nanagara R, Kiatchoosakun S, Suwannaroj S, Mahakkanukrauh A. Prognostic factors of mortality and 2-year survival analysis of systemic sclerosis with pulmonary arterial hypertension in Thailand. Int J Rheum Dis 2011;14:282-9.
7. Poormoghim H, Andalib E, Jalali A, Ghaderi A, Ghorbannia A, Mojtabavi N. Survival and causes of death in systemic sclerosis patients: A single center registry report from Iran. Rheumatol Int 2016;36:925-34.
8. Muangchan C, Canadian Scleroderma Research Group, Baron M, Pope J. The 15% rule in scleroderma: The frequency of severe organ complications in systemic sclerosis. A systematic review. J Rheumatol 2013;40:1545-56.
9. Milletari F, Navab N, Ahmadi SA, editors. V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV). IEEE; 2016.
10. Elhaie M, Koozari A, Shahbazi-Gahrouei D. Machine learning and neural network approaches for enhanced measuring and prediction of radiation doses. J Radiat Res Appl Sci 2025;18:101252.
11. Elhaie M, Etemadifar M, Adariani AR, Khorasani A, Shahbazi-Gahrouei D. Machine learning-based classification of multiple sclerosis lesion activity using multi-sequence MRI radiomics: A complete analysis of T1, T2, FLAIR, DWI, and SWI features. Pol J Radiol 2025;90:e394-403.
12. Jameson JL, Fauci AS, Kasper DL, Hauser SL, Longo DL, Loscalzo J. Medicina Interna de Harrison-2. Vol. 20. McGraw Hill Brasil; 2020.
13. Kim HG, Tashkin DP, Clements PJ, Li G, Brown MS, Elashoff R, *et al*. A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients. Clin Exp Rheumatol 2010;28:S26-35.
14. Kim HJ, Brown MS, Elashoff R, Li G, Gjertson DW, Lynch DA, *et al*. Quantitative texture-based assessment of one-year changes in fibrotic reticular patterns on HRCT in scleroderma lung disease treated with oral cyclophosphamide. Eur Radiol 2011;21:2455-65.
15. Zavaletta VA, Bartholmai BJ, Robb RA. High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis. Acad Radiol 2007;14:772-87.
16. Walsh SL, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: A case-cohort study. Lancet Respir Med 2018;6:837-45.
17. Elhaie M, Koozari A, Sheikh M, Abedi I. Predicting neoadjuvant chemotherapy response in breast cancer using multiparametric MRI-Based radiomics and machine learning: A systematic review and meta-analysis of 1,469 patients. Indian J Surg Oncol 2025. p. 1-17.
18. Ram D, Egan H, Ramanathan T. Ground glass opacity: Can we correlate radiological and histological features to plan clinical decision making? Gen Thorac Cardiovasc Surg 2022;70:971-6.
19. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, *et al*. Lung infection quantification of COVID-19 in CT images with deep learning. arXiv 2020. p. 200304655.
20. Antar S, Abd El-Sattar HK, Abdel-Rahman MH, Ghaleb FF. COVID-19 infection segmentation using hybrid deep learning and image processing techniques. Sci Rep 2023;13:22737.
21. Osadebey M, Andersen HK, Waaler D, Fossaa K, Martinsen AC,

Pedersen M. Three-stage segmentation of lung region from CT images using deep neural networks. BMC Med Imaging 2021;21:112.
22. Saeedizadeh N, Minaee S, Kafieh R, Yazdani S, Sonka M. COVID TV-Unet: Segmenting COVID-19 chest CT images using connectivity imposed Unet. Comput Methods Programs Biomed Update 2021;1:100007.
23. Ganieva N, Djuraeva E, Dwivedi K. Systemic sclerosis and cardiovascular risk: A systematic review of its association with atherosclerosis. Web Discov J Anal Inventions 2025;3:16-23.
24. Wang Y, Ji Y, Xiao H. A data augmentation method for fully automatic brain tumor segmentation. Comput Biol Med 2022;149:106039.
25. Arian R, Mahmoudi T, Riazi-Esfahani H, Faghihi H, Mirshahi A, Ghassemi F, *et al*., editors. Automatic Choroid Vascularity Index Calculation in Optical Coherence Tomography Images with low-Contrast Sclerochoroidal Junction Using Deep Learning. Photonics. MDPI; 2023.
26. Ronneberger O, Fischer P, Brox T, editors. U-net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2015.
27. Gulli A, Pal S. Deep learning with Keras. Packt Publishing Ltd.; 2017.
28. Ray A. Term Project II: Image Processing Using Deep Learning; 2023.
29. Khouy M, Jabrane Y, Ameur M, Hajjam El Hassani A. Medical image segmentation using automatic optimized U-Net architecture based on genetic algorithm. J Pers Med 2023;13:1298.
30. Suárez-García JG, So PW, Hernández-López JM, Hidalgo-Tobón SS, Dies-Suárez P, de Celis-Alonso B. Automatic quantification of abdominal subcutaneous and visceral adipose tissue in children, through MRI study, using total intensity maps and Convolutional Neural Networks. arXiv 2023. p. 230906535.
31. Raff E, Nicholas C. Lempel-Ziv Jaccard distance, an effective alternative to ssdeep and sdhash. Digit Investigation 2018;24:34-49.
32. Pang T, Guo S, Zhang X, Zhao L. Automatic lung segmentation based on texture and deep features of HRCT images with interstitial lung disease. Biomed Res Int 2019;2019:2045432.
33. scikit-learn developers. sklearn.model_selection.KFold [Internet]. Available from: https://scikit-learn/stable/modules/generated/sklearn.model_selection.KFold.html. [Last accessed on 2024 Jan 13].
34. Li G, Xie J, Zhang L, Sun M, Li Z, Sun Y. MCAFNet: multiscale cross-layer attention fusion network for honeycomb lung lesion segmentation. Med Biol Eng Comput 2023. pp. 1–17.