Iranian Journal of Radiology
Volume 19 · Issue 4 · e117992

OPEN PEER REVIEW

A Multi-centric Evaluation of Deep Learning Models for Segmentation of COVID-19 Lung Lesions on Chest CT Scans

Saman Sotoudeh-Paima, Navid Hasanzadeh, Ali Bashirgonbadi, Amin Aref, Mehran Naghibi, Mostafa Zoorpaikar, Aryan Arvin, Masoumeh Gity, Hamid Soltanian-Zadeh

SUBMITTED	REVISED	ACCEPTED	DOI
20 Jul 2021	1 Oct 2022	2 Oct 2022	10.5812/iranjradiol-117992

About this report

This open peer review (OPR) report shows the public comments made on this article during peer review, by the reviewers, the associate editor and the editor-in-chief, for every revision round, together with the authors' replies. Comments appear only with the consent of the people who wrote them. Reviewers who chose to sign their report are named; the others are shown as "Referee". Comments are edited for language only, and confidential remarks to the editors are not included.

More about open peer review: brieflands.com/knowledgebase/opr

Contents

1 **Original submission** Referee 1, Hamed Dashti, Referee 2, Associate Editor, Editor-in-Chief 2
2 **Revision 1** Editor-in-Chief 4
Authors' reply to reviewers 5
Reviewer and editor comments 12
3 **Revision 2** Referee 1, Associate Editor, Editor-in-Chief 13
Authors' reply to reviewers 14
Reviewer and editor comments 15
4 **Revision 3** Editor-in-Chief 16
Authors' reply to reviewers 17
Reviewer and editor comments 38

1 Original submission

5 comments · 3 Aug 2021 – 9 Nov 2021

Referee 1 Reviewer

MINOR REVISION 3 Aug 2021

Thanks for submitting the manuscript to the IJR.
Overall, it can be said that this study was designed, conducted, analyzed, and written appropriately.
However, a statistical analysis section needs to be added as the last paragraph of the methods section in the full text.

Hamed Dashti Reviewer

REJECT 20 Oct 2021

This paper presents a study based on a deep neural network to segment lung CT-scan images of COVID-19 patients. The system can highlight the involved regions, calculate the CT severity score, and determine the location of the infected lung volume. However, this manuscript suffers from several main limitations: first, the results are not well presented, and the authors are advised to provide more solid findings, such as ROC analysis and the AUC of the developed AI algorithm. Second, the results may reflect accuracy but not rapid evaluation, as the authors did not report how quickly AI detects and diagnoses these images. Finally, there is no discussion of the findings in relation to other studies, and the literature review is poor. Furthermore, they suggested some approaches that they can apply themselves and for which they can show the results.

Some other comments:
1) The proposed method using CNN and U-Net has been applied in many fields, including COVID-19 detection; the innovation of the paper is too limited;
2) The paper claims that the model has generalization capability, an important aspect of AI models; however, the results show that the model could not achieve generalization;
3) The experimental results should be compared with those of other methods on some public datasets.

Referee 2 Reviewer

MAJOR REVISION 6 Nov 2021

Dear Authors ,
The paper is very well written; however, there are a few technical points that I would like you to take into consideration before publication in IJR.

Section 3.2
Why was the model trained only on a single dataset? This will result in different performance on the two datasets and may provide misleading information about the capability of the model. Ideally, both datasets must be mixed together, and one model must be developed.

4.Results:
Model performance is summarized in table 1. However, the author did not mention the differences between the models that resulted in such variability in performance. Obviously, a more complex NN structure will result in higher accuracy at the cost of longer computation time. It's recommended that the authors report the computation time for training each model to make a fair assessment of the models.

As expected, the model is performing much better on the dataset that was used to train the model. What was the justification for not developing the model on both datasets? It's recommended that the author train the

models by including the second dataset to make a more realistic assessment relative to the radiologist (table 2).

Associate Editor

MAJOR REVISION 8 Nov 2021

Dear authors;
Thank you for considering IJR for submission of your research.
Our reviewers studied your manuscript.
There are some major points raised by the reviewers that must be considered and require reevaluation.
Best regards,

Editor-in-Chief

MAJOR REVISION 9 Nov 2021

Dear authors,
Your article needs Major Revision.
Regards

2

Revision 1

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

Point-by-Point Responses to Reviewers

We would like to thank the reviewers for their insightful comments, which made our study stronger. Based on the comments, we have carefully revised the manuscript and applied the needed changes. The article is re-written carefully to improve readability. The requested changes in the resubmission are highlighted by a light green shade. The comments and our point-by-point responses are as follows.

Reviewers' Comments:

Reviewer 1:

Thanks for submission the manuscript to the IJR.

Totally, it can be said that this study designed, conducted, analyzed, and written appropriately.

But, it needs to be added the statistical analysis section in the last paragraph full-text method section.

Response: Thanks for your suggestion! The statistical analysis section has been added to the revised manuscript based on your feedback. Please refer to page 6 of the revised manuscript.

Reviewer 2:

Dear Authors,

The paper is very well written however there are few technical points that I would like you to take into consideration before publication in IJR,

1. Section 3.2: Why the model was only trained on a single dataset? This will result in different performances over the two datasets and may provide misleading information about capability of the model. Ideally, the both dataset must be mixed together and one model to be developed.

Response: CT scans provided by different vendors (Siemens, GE, etc.) do not have the same characteristics and result in images that have different quantitative and qualitative information. Additionally, the choice of kernel and reconstruction algorithm is specific to each vendor, resulting in acquired CT scans that look unsimilar.

The main reason for training the model on one dataset and testing it on another dataset was to evaluate the generalizability of the trained model on an unseen database (CT scans from another center and a different scanner). The trained model achieved a dice similarity coefficient (DSC) of $74.35\% \pm 1.57\%$ on COVID-19 images in the training set and a DSC of $56.61\% \pm 1.48\%$ on the second dataset. These results suggested the complexity of the second dataset and the limited generalizability of deep learning-based models when being tested on data with different acquisition protocol/from another vendor.

However, based on the reviewer's valuable recommendation, we have conducted two additional experiments to evaluate the incremental value of using data from the second center. The results of the first experimentation have been added to the revised manuscript. The results of the second experimentation are provided only in the 'Response to Reviewers' document to demonstrate the main reason behind training the model on a single database. These results can be added to the main manuscript if needed. The two experiments are explained below.

In the paper, it was shown that manual calibration of the threshold for generating the segmentation binary maps could enhance the performance of the trained model on the second dataset. This is because CT scans acquired from different vendors come with different brightness and contrast levels. Thus, as the first strategy, we have trained the best-performing model on the first private dataset and calibrated the segmentation map (output) obtained by the model using data from the second private dataset. To calibrate, we set the threshold of segmentation map in the range of 0 and 1, with 0.05 increments, and calculate the DSC on the training set. We then use the threshold with highest DSC on the training set for the test set and report the results. To demonstrate this incremental value, we have considered 50% of the cases in the second dataset as testing data. Then, we have trained the model six times, increasing the amount of training data used gradually in 10% increments (0% to 50%). The incremental value is evaluated in terms of DSC on the test set (half of the cases in the second dataset). The result of this experiment is provided below.

Dice Similarity Coefficient (DSC) on the test set of second dataset						
% training data used for calibration (threshold with highest DSC)	0 % (0.5)	10 % (0.10)	20 % (0.05)	30 % (0.05)	40 % (0.10)	50 % (0.05)
DSC on test set	61.16 %	66.29 %	64.88 %	64.88 %	66.29 %	64.88 %

The results demonstrate the possibility of using very limited number of cases to re-calibrate the model when using an unseen dataset.

As the second strategy, we trained the model on a combination of the first and second dataset. To evaluate the effect of training on both datasets, we have conducted five-fold cross validation on the first dataset and used an incremental percentage of cases from the second private dataset in each run. For this analysis, we have considered 50% of the cases in the second private dataset as the testing set and gradually increased the amount of data used for training from 0% to 50% in 12.5% increments. The incremental value of this fine-tuning strategy is evaluated in terms of DSC on both the first and second private datasets. The result of this experiment is provided below.

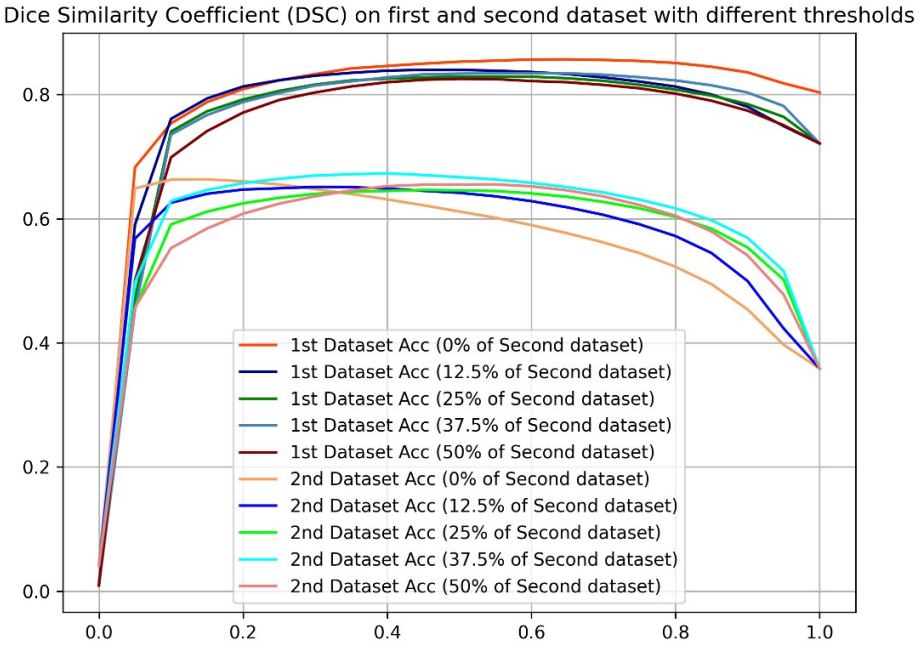


Fig 1. DSC scores on the first and second dataset, trained on 12.5% increments of the second dataset (from 0% to 50%).

The results demonstrate that adding incremental amounts of data from the second private dataset would not help improve the performance on either of the datasets. This result can be explained by the difference between scanners and acquisition protocols. Hence, one solution could be to correct for these differences prior to feeding the data into a CNN. This topic can be the subject of a future study, where the images are first harmonized and then fed to a deep learning-based architecture for lesion segmentation.

2. Results: Model performance is summarized in table 1. However, the author did not mention the differences between the models that resulted in such variability in performance. Obviously, a more complex NN structure will result in higher accuracy with the cost of having larger computational time. It's recommended that the authors report the training computational time of each model to make a fair assessment regarding the models.

Response: Thanks for your attention! The training time and computational complexity (number of parameters) are added to Table II. As can be seen from Tables II and III, the FPN-VGG16 is consistently overperforming other tested models, while having a shorter training time per epoch than most of them (DenseNet and EfficientNet). Additionally, using a UNet-based model increases the training time per epoch by only 2 seconds on a VGG16 backbone (from 60.10 to 62.10 seconds) while increasing the DSC by 2.06%.

3. As expected, the model is performing much better on the dataset that was used to train the model. What was the justification for not developing the model on both datasets? It's recommended the author to train the models by including the second dataset to make more realistic assessment with respect to radiologist (table 2).

Response: Please refer to the answer provided to the 1st comment.

Reviewer 3:

This paper designed a study based on a deep neural network to segment lung CT-scan images of COVID-19 patients. The system can highlight the involved regions and calculate the CT severity score and location of the infected lung volume. However, this manuscript suffers from several main limitations:

1. first, results are not well presented as authors are suggested to provide more sold findings such as ROC analysis, the AUC of the developed AI algorithm.

Response: Thanks for your comment! In the revised manuscript, the ROC curves are added as Fig. 4 and AUC scores are added to Table II.

2. Second, results may reflect the accuracy but not the rapid evaluation as authors did not provide how quick AI to detect and diagnose these images.

Response: Thank you for your attention to this matter! The evaluation time of the model was measured, and the following paragraph was added to the revised manuscript.

The comparable performance of the proposed architecture against radiologists comes with the advantage of time efficiency in AI-based models. While a manual assessment of CT scan volumes would take up to 15 minutes by radiologists, our proposed model segments each slice in 60 milliseconds, resulting in a CT volume segmentation in less than a minute. Thus, AI-based image analysis would have the speed needed to cope with the high demand of image assessment during the COVID-19 era while providing accurate diagnosis.

3. Finally, no discussion of findings concerning other studies as well as a poor literature review.

Response: We have done a major revision of the background section and added related and recent literature to the revised manuscript.

4. Furthermore, they suggested some approaches that can be applied by themselves and show the results.

Response: Based on reviewer's valuable recommendation, we have added an additional experiment to evaluate the incremental value of using data from the second center to calibrate the trained model. The results of the first experimentation have been added to the revised manuscript.

In the paper, it was shown that manual calibration of the threshold for segmentation binary maps could enhance the performance of the trained model on the second dataset. This is because CT scans acquired from different vendors come with different brightness and contrast levels. Thus, as the first strategy, we have trained the best-performing model from the previous analysis on the first dataset and calibrated the segmentation map (output) obtained by the model using data from the second dataset. To calibrate, we set the threshold of segmentation map in the range of 0 and 1, with 0.05 increments, and calculate the DSC on the training set. We then use the threshold with highest DSC on the training set for the test set and report the results. To demonstrate this incremental value, we have considered 50% of the cases in the second dataset as testing data. Then, we have trained the model six times, each time gradually increasing the amount of training data used for calibration by 10% (0% to 50%, in 10% increments). The incremental value is evaluated in terms of DSC on the test set (half of the cases in the second dataset). The result of this experiment is provided below.

Dice Similarity Coefficient (DSC) on the test set of second dataset

% training data used for calibration (threshold with highest DSC)	0 % (0.5)	10 % (0.10)	20 % (0.05)	30 % (0.05)	40 % (0.10)	50 % (0.05)
DSC on test set	61.16 %	66.29 %	64.88 %	64.88 %	66.29 %	64.88 %

The results demonstrate the possibility of using very limited number of cases to re-calibrate the model when using an unseen dataset.

Some other comments:

1. The proposed method using CNN and U-Net has been applied in many fields, including COVID-19 detection, the innovation of the paper is too limited;

Response: The base U-Net model has been implemented in many fields and proved to be a successful architecture in achieving good segmentation performance. Link-Net is another famous architecture, which has been used in several previous studies. One main purpose of this study was to compare the performance of these two models and find the best backbone architecture, which can further enhance their performance while keeping the overall design simple and understandable. In this study, we have demonstrated that combining the famous VGG16 model with the U-Net model gives the best performance on COVID-19 lesion segmentation, which can be used by other researchers in the field as a simple CNN model.

Other contributions of the manuscript are the following:

- Gathering two large-sized datasets from two different cities in Iran, which can be a good representation of difference in CT acquisition protocols, vendors, and patients. Additionally, we have performed training and testing of the proposed model on two public datasets and demonstrated the consistency of the results.
- Evaluating the effect of lung segmentation as a preprocessing step in segmenting COVID-19 lesions, and demonstrating a performance decay in the presence of a lung segmentation preprocessing technique
- Correlation analysis of AI-based model against radiologist in two different datasets for each lung
- Comparing the performance of the AI-based model against an expert radiologist, and providing unbiased comparison in a setting where nothing similar to the test data has been used during the training of the model (by using dataset from another center/city)
- Demonstrating the potential for performance gain by tuning the threshold of the segmentation map on the unseen dataset

2. The paper claims generalization capability of the model as an important part of AI models, by the way, their results show that their model couldn't achieve generalization;

Response: We have added one additional experiment to demonstrate the possibility of improving the segmentation performance through calibrating the threshold value for the segmentation map. Our analysis showed the possibility of calibrating the model using very limited number of cases from the dataset the model was not trained on.

Furthermore, another experiment showed that combining two datasets would not necessarily lead to better performance on each of the two sets. The explanation about this experiment is provided below.

We trained the model on a combination of the first and second dataset. To evaluate the effect of training on both datasets, we have conducted five-fold cross validation on the first dataset and used an incremental percentage of cases from the second dataset in each run. For this analysis, we have considered 50% of the cases in the second dataset as the testing set and gradually increased the amount of data used for training from 0% to 50% in 12.5% increments. The incremental value is evaluated in terms of DSC on both the first and second datasets. The result of this experiment is provided below:

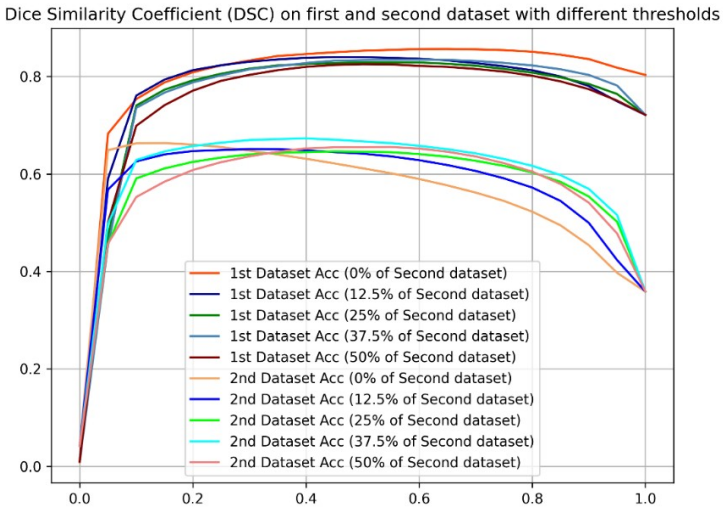


Fig 1. DSC scores on the first and second dataset, trained on 12.5% increments of the second dataset (from 0% to 50%).

The results demonstrate that adding incremental amounts of data from the second dataset would not help improve the performance on either of the datasets. This result can be explained by the difference between scanners and acquisition protocols. Hence, one solution could be to correct for these differences prior to feeding the data into a CNN. This topic can be the subject of a future study, where the images are first harmonized and are then fed to the deep learning-based architecture for lesion segmentation.

3. The experimental results should be compared with other methods on some public datasets.

Response: Thank you for your valuable feedback! The model was tested on two public datasets, once on the COVID19-P20 dataset containing 20 cases (10 cases from Coronacases Initiative and 10 cases from Radiopaedia) and another time on a combination of COVID19-P20 and MosMedData datasets containing 70 cases (20 cases from the COVID19-P20 dataset and 50 cases from the MosMedData dataset).

Associate Editor's Comments:

Associate Editor 1:

Dear authors;

Thank you for considering IJR for submitting your research

Our reviewers studied your manuscript. There are some major points by the reviewers that must be considered and need reevaluation.

Response:

Dear Associate Editor of IJR,

Thank you for your feedback regarding our submission!

The manuscript is carefully revised, and the needed changes have been applied based on reviewers' valuable comments.

2 Revision 1

1 comment · 18 Jun 2022

Editor-in-Chief

MINOR REVISION 18 Jun 2022

Dear authors,
Your article needs Minor Revision.
Regards

3

Revision 2

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

We thank the EIC, the Associate Editor, and the reviewers for their comments and revised the submission accordingly. The comments and our responses are as follows.

Reviewers' Comments:

Reviewer 1: Dear authors, Regarding the importance of “conflict of interest” clarification, the corresponding author should fill out the “ICMJE form” form and upload in the submission process. (The form exists in “conflict of interest” section of the site: https://brief.land/iranjradiol/knowledgebase/category/tree.html#conflicts_of_interest.html).

As some authors are faculty members of Tehran University of medical sciences or editorial board members of the IJR, this should be mentioned as conflict of interest.

Response: We thank the reviewer for careful review of the revised manuscript. We downloaded and completed the form and upload the completed file as “other” files. In addition, we inserted the statement generated by the ICMJE form in the “Conflict of Interests” section of the submission website.

Reviewer 2: Thanks to the authors for their complete edition of the manuscript. Currently, it is ready for publication in the IJR according to methodology.

Response: We thank the reviewer for careful review of the revised manuscript.

Associate Editor's Comments: Dear authors, Thank you for considering IJR for submitting your research. A minor revision has been requested by our reviewer that must be considered and needs reevaluation.

Response: We thank you and the reviewers for careful review of the revised manuscript and have taken care of the minor revision.

EIC Decision: Dear authors, Your article needs Minor Revision.

Response: We thank you and the reviewers for careful review of the revised manuscript and have taken care of the minor revision.

3 Revision 2

3 comments · 25 Jun 2022 – 22 Aug 2022

Referee 1 Reviewer

ACCEPT 25 Jun 2022

At present, this manuscript is ready for publication in IJR from a methodological point of view.

Associate Editor

ACCEPT 20 Aug 2022

Dear authors,
This manuscript is now ready for publication in IJR.
Best regards,

Editor-in-Chief

MINOR REVISION 22 Aug 2022

Dear authors,
Your article needs Minor Revision.
Regards

4

Revision 3

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

Point-by-Point Responses to Reviewers and Associate Editors

We thank the reviewers and the associate editors for their time and consideration in reviewing our manuscript and approving it for publication in IJR. Our responses to the comments are provided in the following. In addition, we applied minor revisions to improve the manuscript readability.

Journal: Iranian Journal of Radiology

- **Manuscript ID:** 117992
- **Manuscript Title:** Multi-Centric Evaluation of Deep Learning Models in Segmentation of COVID-19 Infections from Chest CT
- **Revision:** [2]

> Reviewers' Comments:

Reviewer 1:
At present, this manuscript is ready for publication in IJR, in term of methodological point of view.

Reviewer 2:
In Figure 1, Pls the images with to names A, B, C, D and cite them in the legend.

Response: *Thank you for the suggestion. Figure 1 and the legend are updated.*

> Associate Editor's Comments:

Associate Editor 1:

Dear authors,
This manuscript is now ready for publication in IJR.
Best regards,

Response: *Thank you for your approval of the revised manuscript.*

Associate Editor 2:

1-please add full term of abbreviations in first usage [in abstract and full text separately] for example AI, AUROC, ….

Response: *Thank you for the comment. The abbreviations are added to abstract and full text.*

2-what is AUROC scores? Is it AUROC curve?

Response: *This term is replaced by AUC throughout the text to avoid confusion.*

AUROC (area under receiver operating characteristics) scores are measured by calculating the area under the ROC curve (also known as Area Under Curve or AUC). It is one of the widely used metrics for performance evaluation of any classification model.

3-Please correct this part in line 11 abstract: "deep learning-based models in the automated segmentation of COVID-19 lung lesions and …" [this could be proposed for line 165 and for title too]

Response: *Thank you for the suggestion. COVID-19 infections are replaced by COVID-19 lung lesions throughout the text.*

4-in line 213, please correct: "area under ROC (AUROC) curve,"

Response: *The term is replaced by area under curve (AUC).*

5-it is better the authors add the data of statistical software used for analysis and the data of its manufacturer.

Response: *Statistical analysis was performed using Python (np.corrcoef for calculating Pearson correlation coefficient). This sentence is added in Section 3.2.4 (Statistical Analysis).*

6-it is better the authors add captions of A, B, C, … for different parts of images [for example image 1, A, B, C and D] and address these captions in figure legends.

Response: *Figures are updated based on this suggestion.*

7-please add full term of abbreviations in foot notes of tables. [AUROC, DSC, ….]

Response: *Tables and their captions are updated based on this suggestion.*

Multi-Centric Evaluation of Deep Learning Models in Segmentation of COVID-19 Lung Lesions from Chest CT

Abstract

Background:

Chest computed tomography (CT) scans are among the most common ways of diagnosing patients affected by the coronavirus disease 2019 (COVID-19). While segmenting COVID-19 lung lesions by radiologists could be time-consuming, utilizing advanced deep learning techniques for automated segmentation would be a promising step in managing this disease.

Objectives:

To evaluate the effectiveness of deep learning-based models in the automated segmentation of COVID-19 lung lesions and analyze their generalizability on unseen datasets.

Patients and Methods:

Four separate datasets (two private and two public) were used in this study. The first and second private datasets included 297 subjects (147 healthy, 150 COVID-19 patients) and 82 COVID-19 subjects from different centers, respectively. The public datasets used were the COVID19-P20 dataset, including 20 COVID-19 subjects from two different centers, and the MosMedData dataset containing 50 COVID-19 subjects from a single center. The comparison was made in terms of the Dice similarity coefficient (DSC), receiver operating characteristic (ROC) curves, and area under curve (AUC) scores. The predictions of the deep learning model were compared against radiologists by calculating the Pearson correlation coefficient (PCC) on chest computed tomography (CT) severity scores. Also, the practicality of the model in clinical settings was tested by comparing its performance on an unseen dataset against the DSC between two radiologists. Lastly, the generalizability of the model was discussed, and a simple calibration strategy was proposed.

Results:

The proposed model achieved the best performance across both private datasets with DSC scores of 84.23 % ± 1.73 % on the first private dataset (74.35 % ± 1.57 % on COVID-19 subjects only) and 56.61 % ± 1.48 % on the second private dataset. This superior performance was also validated using two public datasets, achieving DSC scores of 60.10 ± 2.34 on the COVID19-P20 dataset and 66.28 ± 2.80 on the mixed dataset of COVID19-P20 and MosMedData. There was a correlation of 0.89 and 0.85 on the first private dataset and a correlation of 0.77 and 0.74 on the second private dataset for the right and left lungs, respectively. The trained model on the first private dataset was tested on the second private dataset and compared against a radiologist. This comparison revealed a performance gap of 5.74% in terms of DSC. Then, a calibration strategy was utilized to reduce this performance gap to 0.53 %.

Conclusions:
The results indicate that the proposed model performs well in detecting and localizing COVID-19 lesions in CT scans using different datasets, and their accuracy competes with radiologists and can assist them in diagnostic and treatment procedures. The effect of the model calibration was also shown on the performance of an unseen dataset, improving the DSC by more than 5%.

Keywords:
COVID-19, Chest CT, Deep/Machine Learning, Image Segmentation

1. Background

Every day, coronavirus disease 2019 (COVID-19) is infecting hundreds of thousands of people worldwide, resulting in the loss of life in thousands of them. According to the World Health Organization's statistics, by Mar 31, 2022, the number of positive COVID-19 cases is more than 485 million, and the number of deaths is more than 6.1 million (1). Moreover, other aspects of life are also affected in addition to those directly affected by COVID-19. Travels are limited, jobs and businesses are declining, and other illnesses such as depression, obsessive-compulsive disorder, and obesity are becoming more common as a result of staying home.

Recently, several countries have developed vaccines, leading to optimism about the future. However, as vaccines are still not readily available, many countries struggle to fight against this virus. Furthermore, the coronavirus constantly evolves, making it more difficult to detect new variants by previous testing methods. Thus, wearing masks, social distancing, and in cases of infection, early diagnosis, and isolation of the infected person are among the best ways to manage and prevent further spread of the epidemic.

Since the beginning of the epidemic, reverse transcription-polymerase chain reaction (RT-PCR) testing has been proposed as the standard method for COVID-19 screening. However, having a 60-70% sensitivity, not diagnosing the disease in its early stages, being time-consuming, and needing special kits and equipped laboratories to perform tests are among the significant limitations of the RT-PCR test method (1–3).

Medical imaging technology has seen significant advances in recent years and is now a standard method for diagnosing and quantifying various diseases (4–6). Ground-glass opacity (GGO), interlobular septal thickening, and consolidation are among the leading radiological patterns for COVID-19 that differentiate it from healthy individuals and other types of pneumonia (7). A study that included 1014 patients in Wuhan, China, has shown that using chest computed tomography (CT) imaging to diagnose COVID-19 results in a sensitivity of over 97% (3). Therefore, chest CT scans could be considered a more effective, practical, and rapid method of diagnosing and assessing COVID-19 than RT-PCR, particularly in the areas most affected by the epidemic (3). Moreover, the use of chest CT scans to screen for COVID-19 in patients was supported, especially when the results of RT-PCR were negative (2).

The manual evaluation of 3D CT volume data is a tedious and time-consuming process, highly dependent on the expert's clinical experience (8). However, the development of artificial intelligence (AI)-based medical image analysis methods is a way to overcome the challenge

mentioned above. As one of the subsets of AI, deep learning algorithms were found to be effective in fields such as radiology, dermatology, ophthalmology, and pathology (9,10).

Since the beginning of the COVID-19 epidemic, many efforts have been made to automatically analyze chest CT images to expedite and facilitate the proper diagnosis and management of COVID-19 disease. Most proposed methods fall into two general categories: classification and segmentation (11). In classification, the goal is to assign a label to examples from the input domain. For classification of COVID-19 cases, the algorithms are trained on labeled CT images of healthy individuals and COVID-19 patients (and in many cases, other diseases) to learn predictive modeling (5,12–14). Previously, we investigated the performance of various backbone architectures for classifying COVID-19 cases, where the VGG19 architecture demonstrated the best performance (15). In medical image segmentation, the goal is to label each pixel to determine the region of interest (ROI). VB-Net, U-Net, and other variants of U-Net are among the most common algorithms for medical image segmentation (5,16–18). It is shown that classification approaches based on segmentation can obtain higher accuracy compared to the others. This is because segmentation masks in the training phase provide complementary knowledge to the model, leading to more accurate detection of lesions.

Based on the above reasons, many studies used deep learning segmentation techniques for the automated diagnosis of COVID-19 lesions and subsequent interpretation of CT images (19,20). Although most of these studies achieved high accuracy in detecting COVID-19 lesions, the proposed techniques were trained and tested using the CT data acquired by the same scanners. This would result in trained models learning image characteristics of CT scanners which can lead to poor performance on CT data obtained by a new scanner. Thus, analyzing the performance of an unseen dataset and comparing it against a radiologist's performance needs to be a crucial step in assessing the generalizability of AI-based models.

Khan *et al.* (21) proposed a threshold-based segmentation method to quantify COVID-19 lung abnormalities. Their approach obtained a Dice similarity coefficient (DSC) of 46.28% on a combination of two COVID-19 CT datasets. However, they did not assess the generalizability of their method on the datasets separately.

Fan *et al.* (22) introduced a lung infection segmentation network (Inf-Net) to identify infected areas in CT images automatically. They used a semi-supervised approach to make up for the lack of data. However, they had access only to one labeled dataset, making it difficult to investigate the performance of their proposed model on unseen data acquired by different CT devices.

Wang *et al.* (23) proposed a noise-robust learning framework based on a 2D convolutional neural network (CNN) combined with an adaptive self-ensembling framework for slice-by-slice image segmentation. Although they used a dataset consisting of CT images acquired from 10 different hospitals, they randomly split the images into train, validation, and test sets and did not study the generalizability of their proposed method on unseen CT datasets.

A deep learning-based network called VB-Net was developed by Shan *et al.* (24) to segment COVID-19 infection regions in CT scans. The developed VB-Net model was trained using a human-involved-model-iterations (HIMI) strategy on a dataset of 249 COVID-19 patients and validated on another dataset of 300 COVID-19 patients. The model yielded a DSC of 91.6%, while the average DSC between the two radiologists was 96.1%. The closeness of DSC demonstrates the high accuracy of the deep learning-based model in quantifying COVID-19

lesions in CT data. However, the model is validated on a mono-centric dataset, which may not be representative of the generalizability of deep learning systems on different CT scanners.

Müller et al. (1) implemented a standard 3D U-Net architecture through 5-fold cross-validation on 20 CT scan volumes of COVID-19 patients. Their method achieved an average DSC of 76.1%. This study has used a limited dataset size of 20 cases and has been tested on the same dataset it has been trained on; thus, the generalizability of the model was not evaluated on an unseen dataset.

2. Objectives

The aim of this study was to:

- Develop and evaluate a deep learning-based model for automated segmentation of COVID-19 lung lesions from chest CT scans using multi-centric data
- Evaluate the effectiveness of using lung segmentation as a preprocessing technique
- Compare the performance of the AI-based model against radiologists
- Demonstrate the limited generalizability of deep learning-based models through domain shift and propose a calibration strategy to improve the performance

3. Methods

This section describes the image datasets and the proposed method used in this study.

3.1. Image Datasets

Four separate datasets (two private and two public datasets) were used in this study. Table S1 summarizes the COVID-19 CT datasets used in this study. Fig. 1 shows example CT images of each dataset.

3.2. Proposed Method

Fig. 2 provides a schematic presentation of the workflow. The approach used for implementing the proposed algorithm can be explained in three main sections: 1) preprocessing step, 2) automated segmentation of COVID-19 lesions using a VGG16-UNet architecture (25), and 3) experimental setup.

3.2.1. Preprocessing

Data preprocessing is an integral part of medical image segmentation. In this study, the image intensities are normalized to the same scale of zero to one. In addition, all images are resized to the size of 256×256in order to reduce the computational burden. Moreover, we investigated the effect of lung segmentation as a preprocessing stage. To do so, we segmented each lung using (26) as a preprocessing step and trained the model on lung-segmented images. To evaluate the effectiveness of lung segmentation as a preprocessing stage, we evaluated the VGG16-UNet model with and without initial lung segmentation.

3.2.2. Segmentation

To segment COVID-19 lung lesions on CT slices, a deep learning approach based on the U-Net framework (17) was developed. A contracting path (left side) and an expansive path (right side) make up the VGG16-UNet, much like the regular U-Net.

The contracting path is the VGG-16 structure, with five convolutional blocks containing two or three convolutional layers with a small 3×3 receptive field. The resolution is down-sampled at the end of each block using max-pooling layers, and the feature map is up-sampled by a factor of two. The convolutional layers extract higher-level features from the input image as we move from the first block to the last one. On the other hand, the expansive path attempts to restore the images to their original dimension. The resolution is up-sampled by a factor of two at each block, while the number of feature maps is reduced by a factor of two. Two 3×3 convolutional layers then follow each block after the concatenation stage, each having a Batch Normalization (BN) and ReLU layer. Also, via skip connections, a concatenation is performed with the corresponding feature maps from the contracting direction.

To show that our proposed framework is superior to other similar deep learning-based architectures, we performed a comparison among U-Nets (17) and Link-Nets (27) with different encoder structures (25). Fig. S1 provides a comparison of U-Net and LinkNet architectures for semantic segmentation. The number of blocks is subject to change when using different architectures (VGGNets, EfficientNets, ResNets, DenseNets, etc.).

The U-Net and Link-Net architectures are the same in the encoder part. However, two key components differentiate them from one another. First, in the U-Net architecture, the feature maps in the contracting and expansive path are concatenated while they are added in the Link-Net architecture. Second, the decoder block in U-Nets is composed of a 2×2 up-sampling layer and two 3×3 padded convolutional layers, followed by BatchNormalization and ReLU layers. However, the encoder block in Link-Nets consists of a 1×1 convolutional layer reducing the number of channels by a factor of four, a 2×2 up-sampling layer, a 3×3 padded convolutional layer, and a 1×1 convolutional layer that increases the number of channels to the corresponding encoder block's feature map size. A Batch Normalization and a ReLU layer follow all the convolutional layers in the decoder part.

3.2.3. **Experimental Setup**

For quantitative comparison of the tested architectures, we conducted a 5-fold cross-validation on the first private dataset at the patient level, i.e., all the patients were randomly divided into five folds, and each time, four folds were used for training, and one fold was used for testing. Then, within the training dataset, 25% of the patients were selected as the validation set to help the optimization procedure and prevent overfitting. To minimize the segmentation error of the expert in localizing COVID-19 lesions, we excluded CT slices of patients in which no area was specified as the infection region from the training. The main reason for this exclusion was the possibility of missing small lesions from the CT slice by the radiologist during the annotation process. The trained model was tested on the second private dataset to analyze the generalizability in AI-based models. Thus, the mean and standard deviation of five resulting scores are reported.

To further investigate the performance of the trained model, we trained the model on two different datasets. First, the model was trained on the COVID19-P20 dataset containing 20 cases (10 cases from Coronacases Initiative and 10 cases from Radiopaedia) using two-fold cross-validation. In each fold, half of the cases from each group (Coronavirus or Radiopedia) were selected as the

training set, and the rest were selected as validation and test set. Second, the model was trained on a mixture of COVID19-P20 and MosMedData datasets containing 70 cases using five-fold cross-validation.

3.2.4. **Statistical Analysis**

In this work, we employed the DSC, receiver operating characteristic (ROC), area under the curve (AUC), and Pearson's correlation coefficient (PCC) to compare and investigate the obtained results. The Dice coefficient computes the region-based similarity of the model output segmentation result and the ground truth and is defined as:

$$DSC = 2 \vee P \cap G \vee \frac{¿}{|P| + ¿ G \vee ¿ ¿} ¿,$$

where P is the model's output, and G is the mask of infectious areas specified by an expert. The DSC ranges from 0 to 1, with 1 showing the most similarity between the model output and ground truth.

For a binary classification problem, the ROC curve shows the true positive rate (TPR) versus false positive rate (FPR), which are defined as below:

$$TPR = \frac{TP}{TP + FN},$$

$$FPR = \frac{FP}{FP + TN},$$

where TP, FP, TN, TP are the number of true positive, false positive, true negative, and false negative predictions, respectively. The best model would have a TPR of one and an FPR of zero. The model with the highest AUC score was considered the best model since it would have a higher TPR for the same value of FPR.

PCC was used to reveal the performance of the AI model in quantifying chest CT severity scores. We calculated the PCC between the AI model and the radiologist, defined as:

$$PCC = \frac{cov(Z, Z')}{\sigma_Z \sigma_{Z'}},$$

where $cov(Z, Z')$ is the covariance between the AI model and radiologist, σ_Z is the standard deviation of the AI model and $\sigma_{Z'}$ is the standard deviation of the radiologist. The statistical analysis was performed using Python (np.corrcoef for calculating PCC).

4. Results

This section presents the results of all analyses conducted in this study.

First, we compared the proposed architecture against several similar models in terms of DSC, ROC curves, and AUC scores. Table 1 demonstrates the DSC and AUC values, and Fig. 3 shows the corresponding ROC curves attained from testing the model on the first dataset. The results show that the VGG16-UNet architecture achieved superior performance compared to other architectures with a DSC of 84.23 % ± 1.73 % and an AUC of 0.9648. The results were validated using 5-fold cross-validation on the first dataset, including 297 subjects (150 COVID-19 and 147 healthy). Each of the five models trained using five-fold cross-validation was also tested on the second private dataset containing 82 COVID-19 cases. As shown in Table 2, the VGG16-UNet model again demonstrated superior performance with a DSC of 56.61 % ± 1.48 %. The proposed model was also tested on two public datasets, first on the COVID19-P20 dataset and then on a combination of COVID19-P20 and MosMedData datasets. As shown in Table 2, the scores are consistent with prior results, making VGG16-UNet the top-performing model with DSC of 60.10 % ± 2.34 % and 66.28 % ± 2.80 %.

Second, we evaluated the effect of lung segmentation as a preprocessing stage in training the COVID-19 lung lesion segmentation network. Our analysis suggests that having a lung segmentation preprocessing stage does not help the model achieve better performance in both datasets. To that end, we used the model developed in (26), which was also trained on COVID-19 images. Table 3 demonstrates the results of the best-performing model. Considering the model's inferior performance when using this preprocessing stage, no lung segmentation was performed for the following evaluations.

Third, chest CT severity scores were calculated by dividing the overall infection region by the overall lung region segmented using (26) for each lung. Fig. S2 shows the general procedure for calculating the chest CT severity score for a single CT scan. Fig. 4 shows the scatter plot for the AI model's severity prediction against a radiologist for both lungs and datasets.

Fourth, to reach a reasonable understanding of the practicality of using deep learning-based models in real-world settings, we compared the performance of the VGG16-UNet model against the inter-rater agreement between two radiologists in terms of DSC on the unseen dataset (private dataset #2). Table 4 compares the AI model's performance against the average DSC between two radiologists on 67 COVID-19 cases from the second private dataset.

Fifth and last, we investigated the generalizability of the trained model by testing it on an unseen dataset. We then proposed a simple calibration strategy by changing the threshold value of the segmentation map on every new dataset using a very limited number of cases. Table 5 presents the performance of the model with and without calibration on the testing set of the second private dataset.

5. Discussion

5.1. The performance of the encoders

According to Tables 1 and 2, in both U-Net and Link-Net models, using the VGG16 encoder resulted in the best performance. Hence, we can conclude that using the VGG16 encoder causes extracting richer features from the input image and, as a result, achieving a higher DSC. Comparing U-Net and Link-Net structures, we observed superior performance for the U-Net model using all four encoders, which could be explained by the greater number of parameters in the decoder part of the U-Net model. The higher number of parameters could improve the recovery of the lost spatial information in the encoder part.

5.2. Lung segmentation as a preprocessing step

We analyzed the impact of lung segmentation on the overall performance of the model. Table 3 shows that applying lung segmentation as a preprocessing step decreased the performance of the VGG16-UNet model. This could be attributed to inaccurate segmentation of lungs, especially in the presence of a diseased lung. Fig. S3 shows an accurate and inaccurate lung segmentation for a COVID-19 patient. The superior performance of not using lung segmentation can also demonstrate that models can simultaneously learn the location of the lungs and COVID-19 lesion patterns.

5.3. Comparison against radiologists

Validating AI-based models in real-world settings by comparing them against clinical experts has monumental importance. To that end, a correlation analysis was conducted between the AI-based model and radiologists in terms of chest CT severity scores. The results show a high correlation between the AI model and the radiologist for quantifying chest CT severity scores with values of 0.89 and 0.85 on the first private dataset and 0.77 and 0.74 on the second private dataset for the right and left lungs, respectively.

Additionally, the AI model's performance was compared against the average DSC between two radiologists (i.e., inter-rater agreement) using 67 COVID-19 cases from the second private dataset. By comparing the resulting values in Table 4, the performance of the AI-based model was seen to be close to the inter-rater agreement. This demonstrates the accuracy of using deep learning models in quantifying COVID-19 lung lesions in CT scans even when they are not solely trained on that specific dataset.

The comparable performance of the proposed architecture against radiologists comes with the advantage of time efficiency in AI-based models. While a manual assessment of CT scan volumes would take up to 15 minutes by radiologists, our proposed model segments each slice in 60 milliseconds, resulting in COVID-19 segmentation of a volume getting done in less than a minute. Thus, AI-based image analysis would have the speed needed to cope with the high demand for image assessment during the COVID-19 era while providing accurate diagnoses (28).

It is worth mentioning that we used the threshold of 0.5 to classify each pixel in the output segmentation map into either a COVID-19 lesion (labeled as 1) or normal (labeled as 0). This is because the threshold of 0.5 was the best threshold value on the validation part of the first dataset (see Fig. 5A). However, there is no guarantee that this holds true for the unseen dataset (see Fig. 5B). The next section describes a simple calibration idea to improve the performance on unseen datasets simply by changing the threshold of the segmentation map using a very limited number of cases.

5.4. Generalization analysis and calibration

CT images can vary due to the differences in CT scanners and imaging parameters. This variability shows itself in terms of different spatial resolution, image contrast, and noise level, which would negatively impact the accuracy and consistency of deep learning-based models. However, there are several explanations for this performance decay.

First, the performance decay from $84.23\% \pm 1.73\%$ on the first dataset to $56.61\% \pm 1.48\%$ on the second dataset could be partly explained by the presence of healthy subjects in the first (mixed)

private dataset. Eliminating healthy subjects from the testing set would result in a DSC of 74.35 % ± 1.57 % in COVID-19 cases. Second, further qualitative analysis of the second private dataset demonstrated its complexity, especially due to the presence of mild COVID-19 subjects, which leads to significant decay in DSC score when the model misses small lesions. The low inter-rater agreement between two radiologists (DSC of 64.01 %on 67 COVID-19 cases from the second dataset) provides further evidence for the complexity. Third and last, domain shift is the other main reason for the performance decay, happening as a result of differences in scanners and imaging protocols across different centers.

To minimize the performance decay, we investigated the effect of changing the threshold of the output segmentation map on both datasets. Fig. 5 demonstrates that using the same threshold for both datasets would not necessarily lead to the best performance. Thus, one can employ a calibration procedure to determine the best threshold value for the unseen dataset.

To demonstrate this possibility, we have trained the best-performing model on the first private dataset and calibrated the segmentation map (output) obtained by the model using data from the second private dataset. To calibrate, we set the threshold of the segmentation map in the range of 0 and 1, in 0.05 increments, and calculate the DSC on the training set. Then, we use the best threshold for the validation set on the testing set and report the results. To demonstrate this incremental value, we have considered 50% of the cases in the second dataset as testing data. Then, we trained the model six times, increasing the amount of training data used gradually in 10% increments (0% to 50%). The incremental value is evaluated in terms of DSC on the test set (half of the cases in the second private dataset). The results demonstrate the potential for using a very limited number of cases (10%) to calibrate the model when using an unseen dataset.

Other more advanced techniques that can mitigate the variations in CT vendors and acquisition protocols include using CNN-based image normalization and generative adversarial networks (GANs). One idea could be to develop an adversarial neural network similar to (29) but for chest CT scans to be used as a harmonization block to transform the CT data acquired from different scanners and with various imaging protocols into a reference standard. Then, the main network could perform a more effective and uniform assessment of segmenting COVID-19 lesions. This idea could result in a better performance of the main network on the unseen dataset regardless of the scanner specification and without applying any calibration procedure. The investigation of image harmonization methods for mitigating data variations can be the subject of our future study.

6. Conclusion

This work presented a deep learning-based approach to automatically detect and segment COVID-19 lung lesions using chest CT scans. It was shown that a VGG16-UNet model performed better compared to other architectures and achieved a DSC of 84.23 % ± 1.73 % on a mixed dataset of healthy and COVID-19 subjects and a DSC of 56.61 % ± 1.48 %on the unseen dataset of COVID-19 patients. The performance gap was attributed to: 1) the presence of healthy subjects in the first dataset, boosting the DSC from 74.35 % ± 1.57 % to 84.23 % ± 1.73 %, 2) the complexity of the second private dataset due to the presence of mild COVID-19 subjects, 3) the limited generalizability of the model as a result of domain shift happening because of the variations in CT scanners and imaging protocols. The model was further assessed on two public datasets and achieved DSC scores of 60.10 % ± 2.34 % and 66.28 % ± 2.80 % on the COVID19-P20 dataset and the combination of COVID19-P20 and MosMedData datasets. In addition, the

experiments suggested that lung segmentation was an ineffective preprocessing strategy for infection segmentation of COVID-19 cases. Also, the deep learning-based model was shown to have a good agreement with the radiologist. Lastly, the generalizability of the model was discussed, and a simple calibration strategy was proposed, improving the performance by more than 5% in terms of DSC. The future study will focus on image harmonization techniques that can help mitigate the unwanted variations in CT scans.

Figures and Tables

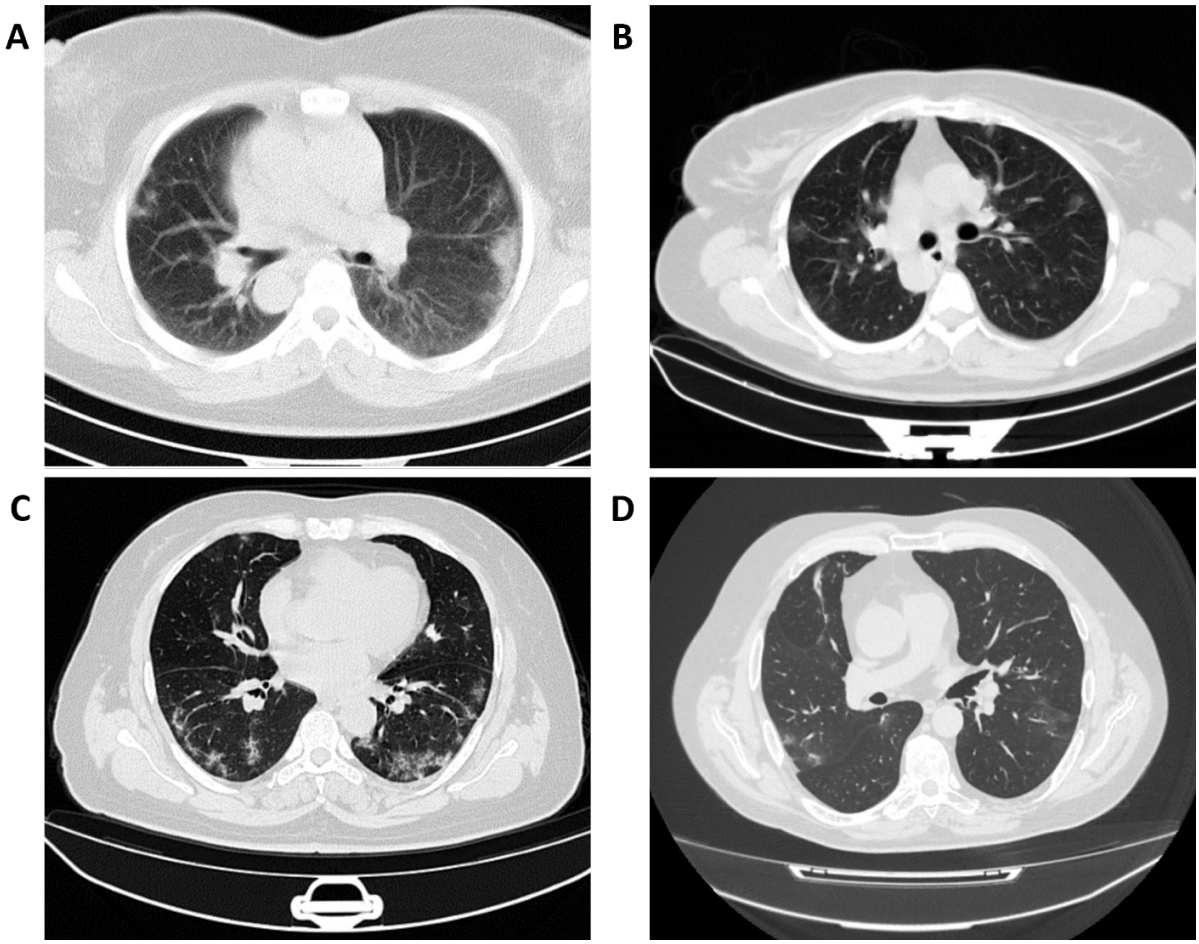
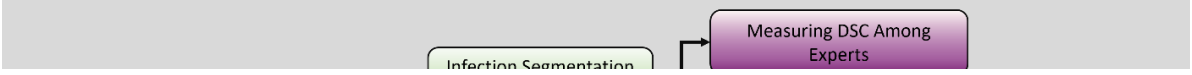


Fig. 1. CT images of patients infected by COVID-19 for the first (A) and second (B) public datasets, the COVID19-P20 (30) dataset (C), and MosMedData (31) dataset (D). The difference in chest CT characteristics (intensity, color, texture, etc.) suggests the importance of generalization analysis of deep learning-based models.



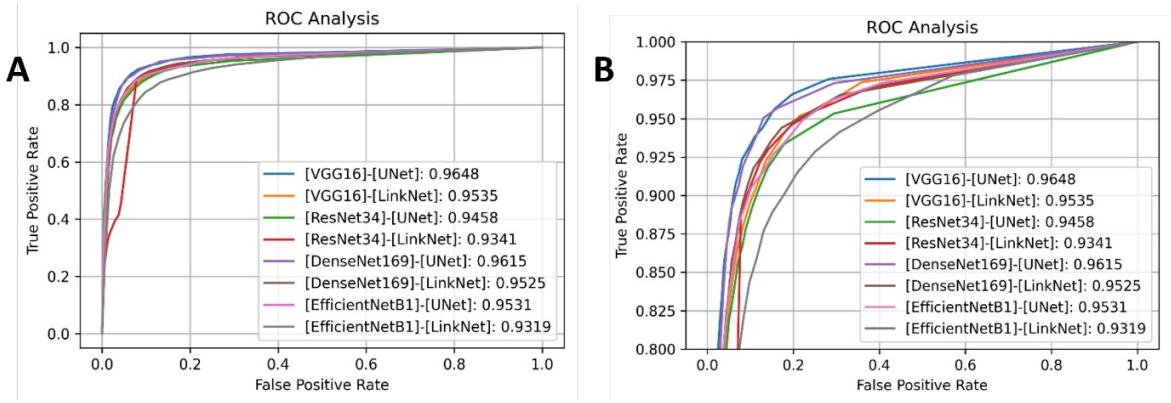


Fig. 3. ROC curve and corresponding AUC values on the first private dataset (A) and its zoomed version (B)

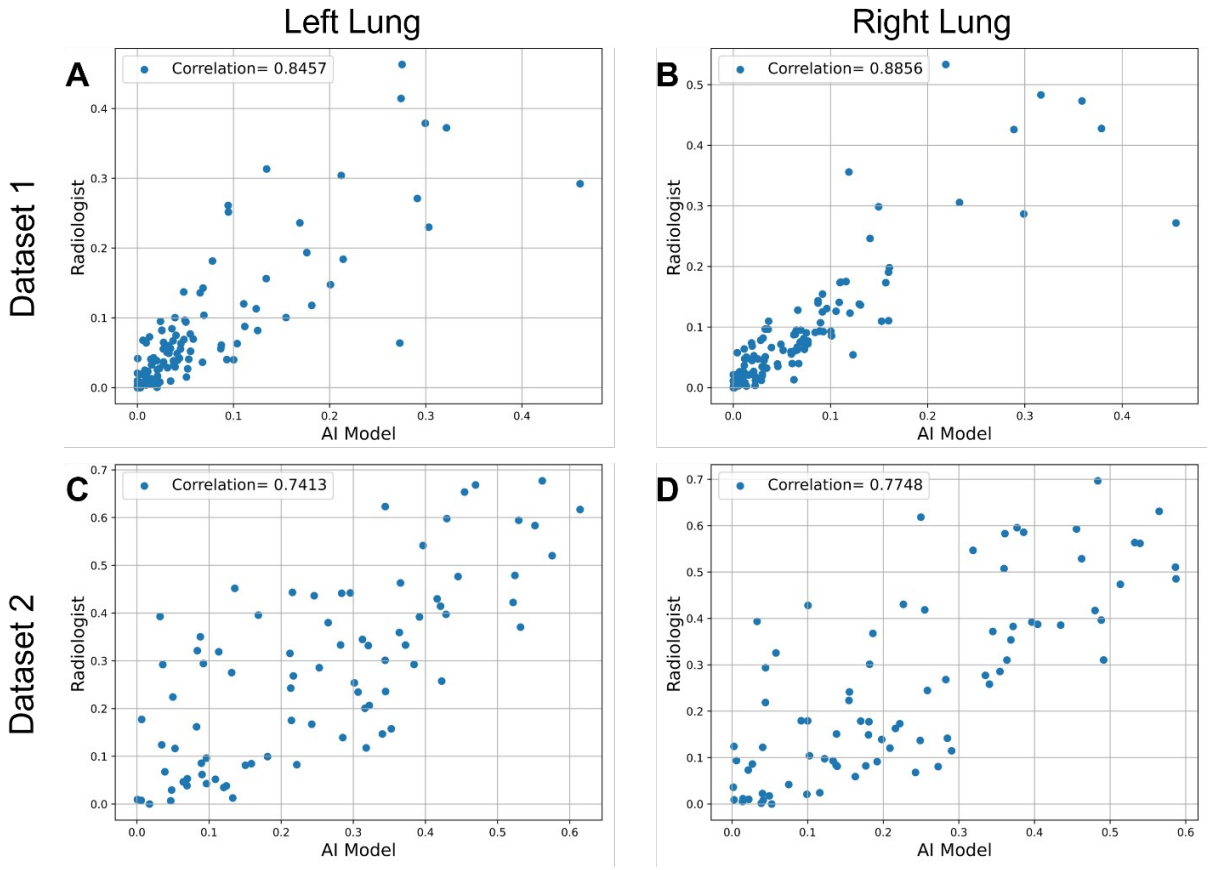


Fig. 4. Scatter plot for the chest CT severity score prediction correlation of AI model vs. radiologist for both private datasets and lungs. (A) First dataset, left lung, (B) First dataset, right lung, (C) Second dataset, left lung, (D) Second dataset, right lung.

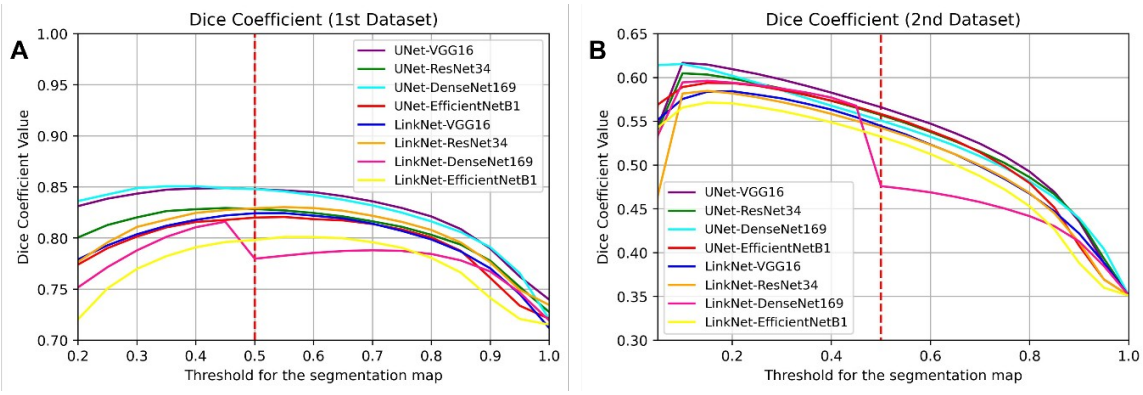


Fig. 5. The effect of changing the threshold of the segmentation map on the model's performance. (A) shows this effect on the validation set of the first private dataset, and (B) shows this effect on the second private dataset (unseen dataset). The plot illustrates that the variability in scanners and imaging protocols can negatively affect the accuracy of deep learning-based models. Thus, employing a calibration procedure to determine the best threshold value could improve the performance.

Table 1. Model specifications, Dice similarity coefficients (DSCs) in % (mean ±standard deviation), and areas under ROC curves (AUCs) for private dataset #1.

Model	Encoder	# of params (M)	Training time (s)	DSC (%)	AUC
U-Net	**VGG16**	23.75	62.10	84.23 ± 1.73	0.9648
	ResNet34	24.45	42.10	82.68 ± 1.55	0.9458
	DenseNet169	19.51	82.13	83.76 ± 0.85	0.9615
	EfficientNetB1	12.64	81.44	82.58 ± 1.25	0.9531
Link-Net	VGG16	20.32	60.10	82.17 ± 1.91	0.9535
	ResNet34	21.63	36.60	82.13 ± 0.95	0.9341
	DenseNet169	15.61	77.13	77.32 ± 5.02	0.9525
	EfficientNetB1	8.55	76.25	80.41 ± 1.07	0.9319

Table 2. Dice similarity coefficients (DSCs) in % (mean ±standard deviation) for the applications of the models on the private dataset #2 and two public datasets (COVID19-P20 and COVID19-P20 + MosMed).

Model	Encoder	Private Dataset #2	COVID19-P20	COVID19-P20 + MosMed
U-Net	**VGG16**	56.61 ± 1.48	60.10 ± 2.34	66.28 ± 2.80
	ResNet34	55.72 ± 1.42	53.35 ± 2.25	57.32 ± 7.56
	DenseNet169	55.09 ± 1.60	57.38 ± 3.04	65.08 ± 3.95
	EfficientNetB1	55.83 ± 1.30	53.16 ± 2.04	62.43 ± 2.47
Link-Net	VGG16	54.46 ± 3.08	54.14 ± 4.05	61.82 ± 1.69
	ResNet34	54.25 ± 1.14	54.54 ± 0.90	57.48 ± 3.03
	DenseNet169	47.60 ± 10.56	54.22 ± 1.04	64.10 ± 5.08
	EfficientNetB1	53.24 ± 1.68	38.94 ± 6.14	49.29 ± 5.05

Table 3. Evaluating the effect of lung segmentation on Dice similarity coefficients (DSCs) in % (mean ±standard

deviation) for private datasets #1 and #2.

Model: VGG16-UNet	DSC (%)	
	Dataset 1	Dataset 2
Without Segmentation	84.23 ± 1.73	56.61 ± 1.48
With Segmentation	83.10 ± 1.39	54.90 ± 1.12

Table 4. Performance comparison of models and radiologist on private dataset #2 using Dice similarity coefficient (DSC) in %.

Type	DSC (%)
VGG16-UNet (Threshold: 0.5)	58.27%
VGG16-UNet (Threshold: 0.15)	63.48%
Radiologist	64.01%

Table 5. Dice similarity coefficients (DSCs) in % for 50% (testing set) of private dataset #2.

% of training data used for calibration (threshold with highest DSC)	0 % (0.5)	10 % (0.10)	20 % (0.05)	30 % (0.05)	40 % (0.10)	50 % (0.05)
DSC (%) on test set	61.16 %	66.29 %	64.88 %	64.88 %	66.29 %	64.88 %

Supplementary Figures and Tables

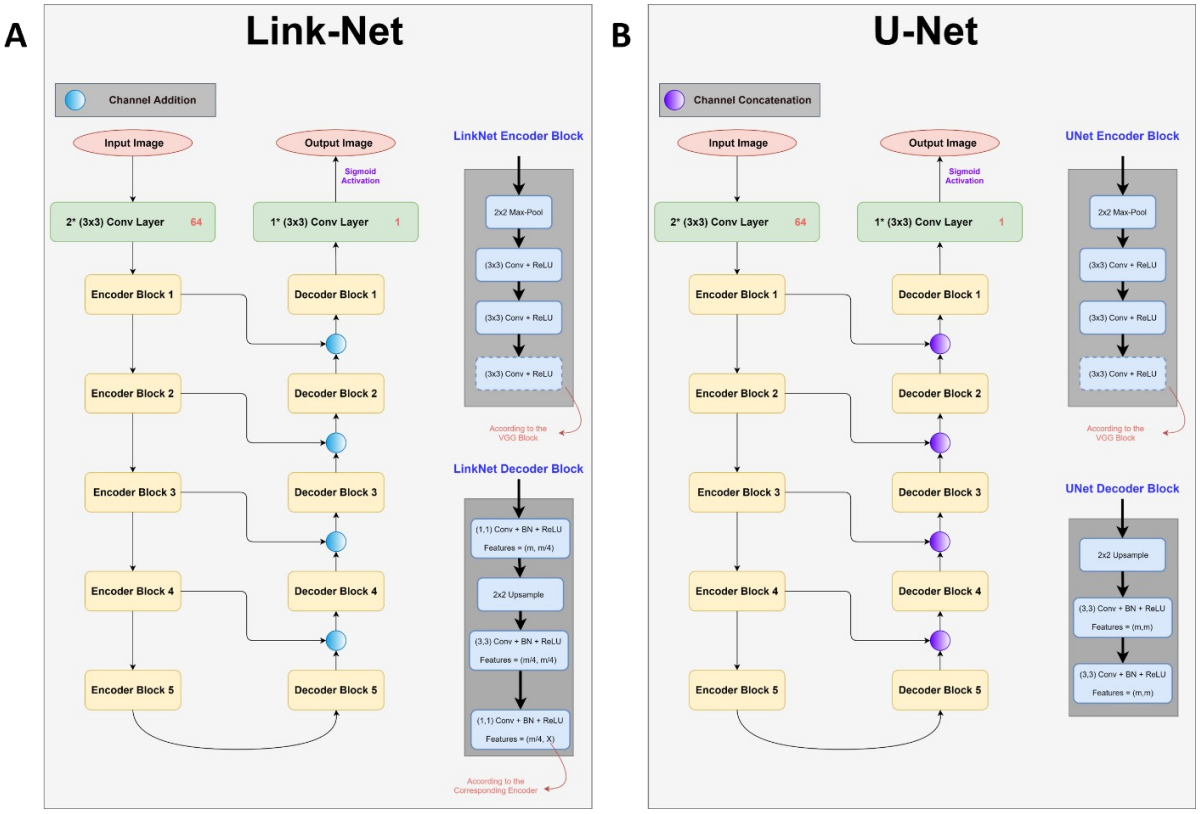


Fig. S1. Comparison of Link-Net (A) and U-Net (B) architectures for semantic segmentation. The block diagrams illustrate the similarity and difference of Link-Net-based and U-Net based models. The encoder parts are the same in both models; however, decoder parts consist of different layer structures.

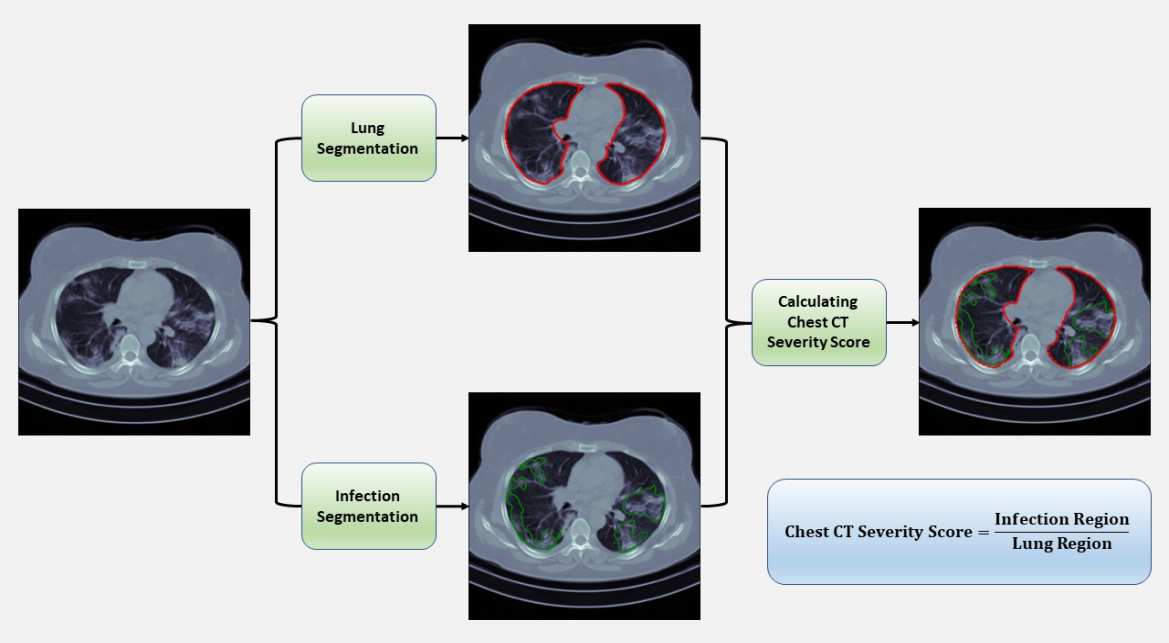


Fig. S2. General procedure for calculating chest CT severity score for a single CT slice. Chest CT severity scores

are calculated for each lung volume.

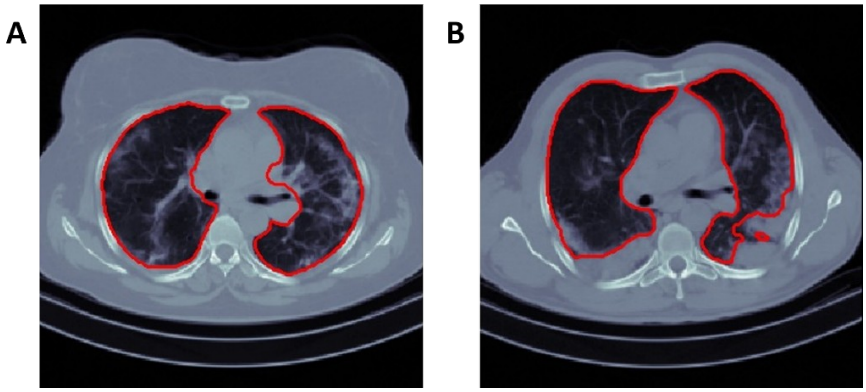


Fig. S3. Accurate (A) and inaccurate (B) lung segmentation. When training the model using lung-segmented CT images, the inaccurate segmentations progress into model's learning stage, reducing the performance in terms of DSC.

Table S1. A summary of used COVID-19 CT datasets

Dataset name	# of subjects	# of slices	Used For
Private #1	297 (147 C, 150 N)*	8739	Training and testing
Private #2	82	3669	Testing of the model trained on the private dataset #1
Public #1 (COVID19-P20)	20	3520	Training and testing
Public #2 (MosMedData)	50	2049	Training and testing
*: C refers to COVID-19 subjects, and N refers to Normal subjects			

References

1. Müller D, Rey IS, Kramer F. Automated Chest CT Image Segmentation of COVID-19 Lung Infection based on 3D U-Net. arXiv. 2020 Jun 24;04774.

2. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. Radiology [Internet]. 2020 Aug;296(2):E115–7. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2020200432

3. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology [Internet]. 2020 Aug;296(2):E32–40. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2020200642

4. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis [Internet]. Vol. 42, Medical Image Analysis. Elsevier BV; 2017. p. 60–88. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1361841517301135

5. Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. IEEE Rev Biomed Eng [Internet]. 2021;14:4–15. Available from: https://ieeexplore.ieee.org/document/9069255/

6. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, et al. The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society. Radiology [Internet]. 2020 Jul 1;296(1):172–80. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2020201365

7. Hasanzadeh N, Paima SS, Bashirgonbadi A, Naghibi M, Soltanian-Zadeh H. Segmentation of COVID-19 Infections on CT: Comparison of Four UNet-Based Networks. In: 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME). 2020. p. 222–5.

8. Ma J, Wang Y, An X, Ge C, Yu Z, Chen J, et al. Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation. Med Phys [Internet]. 2021 Mar 6;48(3):1197–210. Available from: https://onlinelibrary.wiley.com/doi/10.1002/mp.14676

9. Ting DSW, Peng L, Varadarajan A V., Keane PA, Burlina PM, Chiang MF, et al. Deep learning in ophthalmology: The technical and clinical considerations. Prog Retin Eye Res [Internet]. 2019 Sep 1;72:100759. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1350946218300909

10. Sotoudeh-Paima S, Jodeiri A, Hajizadeh F, Soltanian-Zadeh H. Multi-scale convolutional neural network for automated AMD classification using retinal OCT images. Comput Biol Med [Internet]. 2022;144(February):105368. Available from: https://doi.org/10.1016/j.compbiomed.2022.105368

11. Soomro TA, Zheng L, Afifi AJ, Ali A, Yin M, Gao J. Artificial intelligence (AI) for medical imaging to combat coronavirus disease (COVID-19): a detailed review with direction for future research. Artif Intell Rev [Internet]. 2022 Feb 15;55(2):1409–39. Available from: https://link.springer.com/10.1007/s10462-021-09985-z

12. Bullock J, Luccioni A, Hoffman Pham K, Sin Nga Lam C, Luengo-Oroz M. Mapping the landscape of Artificial Intelligence applications against COVID-19. J Artif Intell Res [Internet]. 2020 Nov 19;69:807–45. Available from: https://www.jair.org/index.php/jair/article/view/12162

13. Abdar AK, Sadjadi SM, Soltanian-Zadeh H, Bashirgonbadi A, Naghibi M. Automatic Detection of Coronavirus (COVID-19) from Chest CT Images using VGG16-Based Deep-Learning. In: 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME) [Internet]. IEEE; 2020. p. 212–6. Available from: https://ieeexplore.ieee.org/document/9319326/

14. Tushar FI, Abadi E, Sotoudeh-Paima S, Fricks RB, Mazurowski MA, Segars WP, et al. Virtual versus reality: external validation of COVID-19 classifiers using XCAT phantoms for chest computed tomography. In: Iftekharuddin KM, Drukker K, Mazurowski MA, Lu H, Muramatsu C, Samala RK, editors. Medical Imaging 2022: Computer-Aided Diagnosis [Internet]. SPIE; 2022. p. 9. Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12033/2613010/Virtual-versus-reality--external-validation-of-COVID-19-classifiers/10.1117/12.2613010.full

15. Paima SS, Hasanzadeh N, Jodeiri A, Soltanian-Zadeh H. Detection of COVID-19 from Chest Radiographs: Comparison of Four End-to-End Trained Deep Learning Models. In: 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME) [Internet]. IEEE; 2020. p. 217–21. Available from: https://ieeexplore.ieee.org/document/9319444/

16. Wang B, Jin S, Yan Q, Xu H, Luo C, Wei L, et al. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system. Appl Soft Comput [Internet]. 2021 Jan 1;98:106897. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1568494620308358

17. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. 2015. p. 234–41.

18. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. p. 770–8.

19. Shah FM, Joy SKS, Ahmed F, Hossain T, Humaira M, Ami AS, et al. A Comprehensive Survey of COVID-19 Detection Using Medical Images. SN Comput Sci [Internet]. 2021 Nov 28;2(6):434. Available from: https://link.springer.com/10.1007/s42979-021-00823-1

20. Hassan H, Ren Z, Zhou C, Khan MA, Pan Y, Zhao J, et al. Supervised and weakly

supervised deep learning models for COVID-19 CT diagnosis: A systematic review. Comput Methods Programs Biomed [Internet]. 2022 May;218:106731. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0169260722001171

21. Fan D-P, Zhou T, Ji G-P, Zhou Y, Chen G, Fu H, et al. Inf-Net: Automatic COVID-19 Lung Infection Segmentation From CT Images. IEEE Trans Med Imaging [Internet]. 2020 Aug 28;39(8):2626–37. Available from: https://link.springer.com/10.1007/s11760-022-02183-6

22. Fan D-P, Zhou T, Ji G-P, Zhou Y, Chen G, Fu H, et al. Inf-Net: Automatic COVID-19 Lung Infection Segmentation From CT Images. IEEE Trans Med Imaging [Internet]. 2020 Aug;39(8):2626–37. Available from: https://ieeexplore.ieee.org/document/9098956/

23. Wang G, Liu X, Li C, Xu Z, Ruan J, Zhu H, et al. A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions From CT Images. IEEE Trans Med Imaging [Internet]. 2020 Aug;39(8):2653–63. Available from: https://ieeexplore.ieee.org/document/9109297/

24. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, et al. Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction. Med Phys [Internet]. 2021 Apr 9;48(4):1633–45. Available from: https://onlinelibrary.wiley.com/doi/10.1002/mp.14609

25. Yakubovskiy P. Segmentation Models. GitHub Repos [Internet]. 2019; Available from: https://github.com/qubvel/segmentation_models

26. Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp [Internet]. 2020 Dec 20;4(1):50. Available from: https://eurradiolexp.springeropen.com/articles/10.1186/s41747-020-00173-2

27. Chaurasia A, Culurciello E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP) [Internet]. IEEE; 2017. p. 1–4. Available from: http://ieeexplore.ieee.org/document/8305148/

28. Pham TD. A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks. Sci Rep [Internet]. 2020 Dec 9;10(1):16942. Available from: https://doi.org/10.1038/s41598-020-74164-z

29. Zarei M, Abadi E, Fricks R, Segars WP, Samei E. A probabilistic conditional adversarial neural network to reduce imaging variation in radiography. In: Bosmans H, Zhao W, Yu L, editors. Medical Imaging 2021: Physics of Medical Imaging [Internet]. SPIE; 2021. p. 139. Available from: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11595/2582336/A-probabilistic-conditional-adversarial-neural-network-to-reduce-imaging-variation/10.1117/12.2582336.full

30. Ma J, Wang Y, An X, Ge C, Yu Z, Chen J, et al. Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation. Med Phys [Internet]. 2021

Mar 6;48(3):1197–210. Available from: https://onlinelibrary.wiley.com/doi/10.1002/mp.14676

31. Morozov SP, Andreychenko AE, Pavlov NA, Vladzymyrskyy A V., Ledikhova N V., Gombolevskiy VA, et al. MosMedData: Chest CT Scans With COVID-19 Related Findings Dataset. arXiv Prepr arXiv200506465 [Internet]. 2020 May 13; Available from: http://arxiv.org/abs/2005.06465

4 Revision 3

1 comment · 2 Oct 2022

Editor-in-Chief

ACCEPT 2 Oct 2022

Dear authors,
Your article is acceptable for publication in IJR.
Regards