Fung *et al. J Transl Med* (2021) 19:318
https://doi.org/10.1186/s12967-021-02992-2

Journal of
Translational Medicine

RESEARCH Open Access

Self-supervised deep learning model for COVID-19 lung CT image segmentation highlighting putative causal relationship among age, underlying disease and COVID-19

Daryl L. X. Fung[1†], Qian Liu[1,2†], Judah Zammit[1], Carson Kai-Sang Leung[1] and Pingzhao Hu[1,2,3*]

Abstract

Background: Coronavirus disease 2019 (COVID-19) is very contagious. Cases appear faster than the available Polymerase Chain Reaction test kits in many countries. Recently, lung computerized tomography (CT) has been used as an auxiliary COVID-19 testing approach. Automatic analysis of the lung CT images is needed to increase the diagnostic efficiency and release the human participant. Deep learning is successful in automatically solving computer vision problems. Thus, it can be introduced to the automatic and rapid COVID-19 CT diagnosis. Many advanced deep learning-based computer vison techniques were developed to increase the model performance but have not been introduced to medical image analysis.

Methods: In this study, we propose a self-supervised two-stage deep learning model to segment COVID-19 lesions (ground-glass opacity and consolidation) from chest CT images to support rapid COVID-19 diagnosis. The proposed deep learning model integrates several advanced computer vision techniques such as generative adversarial image inpainting, focal loss, and lookahead optimizer. Two real-life datasets were used to evaluate the model's performance compared to the previous related works. To explore the clinical and biological mechanism of the predicted lesion segments, we extract some engineered features from the predicted lung lesions. We evaluate their mediation effects on the relationship of age with COVID-19 severity, as well as the relationship of underlying diseases with COVID-19 severity using statistic mediation analysis.

Results: The best overall F1 score is observed in the proposed self-supervised two-stage segmentation model (0.63) compared to the two related baseline models (0.55, 0.49). We also identified several CT image phenotypes that mediate the potential causal relationship between underlying diseases with COVID-19 severity as well as the potential causal relationship between age with COVID-19 severity.

Conclusions: This work contributes a promising COVID-19 lung CT image segmentation model and provides predicted lesion segments with potential clinical interpretability. The model could automatically segment the COVID-19 lesions from the raw CT images with higher accuracy than related works. The features of these lesions are associated with COVID-19 severity through mediating the known causal of the COVID-19 severity (age and underlying diseases).

Keywords: COVID-19, Self-supervised learning, Lung CT images, Image segmentation, Mediation analysis

*Correspondence: pingzhao.hu@umanitoba.ca
†Dary L.X. Fung and Qian Liu contributed equally to the manuscript
[1] Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
Full list of author information is available at the end of the article

Background

Coronavirus disease 2019 (COVID-19) is a newly identified infectious disease, which was first reported in December 2019 [1]. According to an interactive COVID-19 dashboard created by Johns Hopkins University, COVID-19 has spread to more than 190 counties and caused 3,957,898 global deaths out of more than 182 million diagnosed cases by July 2nd, 2021 [2]. Several interventions have been applied worldwide to control the COVID-19 pandemic, such as case isolation, close contact quarantine, population lockdown, face covering, sanitization, and vaccination. Although these preventative measures have successfully reduced the number of deaths and confirmed cases, we will likely experience more waves of COVID-19 as restrictions are loosened and the new variants appear [3].

Recently, many efforts have been made to develop artificial intelligence (AI) models to support medical imaging-based COVID-19 rapid diagnosis [4]. Compared to Polymerase Chain Reaction (PCR) which is the current gold standard COVID-19 diagnosis test, medical imaging such as computerized tomography (CT) scans of the lungs does not waste consumables. Therefore, CT imaging-based COVID-19 diagnosis is more efficient as it would not be limited by the delay of available testing kits, especially when AI is introduced to release the need for human involvement in image reading [5]. However, current AI COVID-19 diagnostic models based on medical imaging generally lack transparency and clinical interpretability [6]. The complexity of AI models and their low reproducibility have weakened their applications in clinical practice [7]. Hence, it is critical to develop AI-based COVID-19 diagnosis models with clinical interpretability. Age, underlying diseases, and sometimes gender are observed to be related to the risk of COVID-19 [8, 9]. Lung CT image is a good predictor of COVID-19 status. This is likely due to its associations with age, gender, and underlying diseases [10, 11]. If this is the case, mediation analysis [12] between the age, gender, underlying disease and the risk of COVID-19 through lung CT image phenotypes can potentially be used to reason on the model predictions both biologically and statistically. This may have the potential to improve the cost-effectiveness, diagnosis efficacy, and clinical utility of AI-based COVID-19 CT imaging diagnosis [6].

There are several related works that used self-supervised learning approaches for predicting if the CT lung images is COVID-19 positive or COVID-19 negative. Chen et al. [13] proposed to use contrastive self-supervised learning with 3 major components—data augmentation, representation learning, and few-shot classification. The data augmentation that they used involved cropping two parts of the CT lung images, one part undergone random cropping followed by random flipping, the other part undergone random cropping followed by colour distortion. Then, representation learning was trained to improve on the similarity score where cropped images from the same CT lung image achieved higher similarity score and cropped images from different CT lung images achieved a lower similarity score. After training representation learning on the model, the pre-trained model was used to encode the query image and the support set of CT lung images. The encoded features were passed into prototypical networks to conduct the few-shot classification. The limitation of the work in this approach is that support set images are required for the classification. It can also be hard for the classification to get a good performance if the encoded features of the query image are very different from the support set images. Li et al. [14] used self-supervised dual-track learning to rank. Since there are more available COVID-19 negative samples than COVID-19 positive samples, their method selected a subset of the negative samples to train on the network so that a more balanced data was trained. The way the subset of the negative samples was selected is that they generated two soft labels ("difficulty" and "diversity") for the negative samples by computing the earth mover's distance between the COVID-19 negative samples and the COVID-19 positive samples and selected the soft labels generated accordingly.

As it has been more than one year since the occurrence of COVID-19, many lung CT image datasets are now available online. Though effective, deep learning requires data sets with a large sample size to achieve better performance [15]. However, only few COVID-19 CT image dataset has segmentation label information. Therefore, it is important to effectively utilize them. Voulodimos et al. proposed two deep learning models to do COVID-19 infected area segmentation from CT image patches [16, 17]. These patches acted as augmentations of the raw images. Ma et al. further proposed the data-efficient learning which involves few-shot learning, domain generalization for COVID-19 segmentation with limited training data [18]. Transfer learning is also widely used to complement the sample size issue in COVID-19 infection detection and segmentation from medical images [19–21]. However, transfer learning usually involves models pretrained on non-medical images, which may not perform well in the medical image scenario. Incorporating unlabeled data into model training strategies is also an approach to improve the prediction performance when the labeled data have limited size [22]. Yao et al. even proposed a label-free deep learning-based segmentation model which took advantage of unsupervised anomaly detection techniques [23]. However, their model only outperformed other unsupervised approaches and maybe not comparable with supervised methods. Self-supervised learning is another way to involve unlabeled data. It aims at creating tasks to generate auto-achievable labels without additional human annotations [24]. In the context of self-supervision, image inpainting [25] refers to the creation of a task for the model to generate the content of missing or damaged regions based on the surrounding information [26–28]. The images are damaged on purpose by making some missing regions. Then the model is trained to recover the damaged images to their raw versions. Image inpainting was reported to have excellent pre-training ability for convolutional neural network (CNN) based image segmentation, because it can improve network feature learning [26, 29]. By controlling the complexity of missing regions in the images, we can manage the difficulty of the inpainting task. However, it is hard to create proper missing regions for network to learn, because the missing regions can either be too complex for the network to start learning or too simple to be able to learn good representations [30]. A coach network with generative adversarial mechanism [31] can be used to create the missing region masks with proper complexity. The created mask can initially be simple. Once the network can predict the inpainting of the CT images with good performance, the coach network increases the complexity of the masking to reduce the performance of the network, similar to how the generative adversarial network (GAN) [31] works.

Automatic lung CT image lesion segmentation is not easy due to its variation [32]. To distinguish different kinds of lesions is even harder. To exhaust the information in COVID-19 lung CT images and help us to understand the disease thoroughly, it is important to segment and understand COVID-19 lung CT lesions at a pixel level, including ground-glass opacity (GGO) and consolidation. The recently developed COVID-19 Lung Infection Segmentation Network (InfNet) [22] uses a two-stage strategy to solve the multi segmentation problem. That is, the overall lesion is first segmented, and then passed to the second stage for further distinguishing into the GGO or consolidation lesion [22].

In diagnostic radiology, consolidation could be either pus, edema, blood, or a tumor replacing the airspace in the lung, while GGO is either the filling of pus, edema, hemorrhage, inflammation, or tumor cells in the alveolar space [33, 34. These two lung CT patterns often present together, but GGO is more commonly observed in COVID-19 lung CT images than consolidation [35]. This is the case in the segmented lung CT image dataset we found online [36]. This imbalanced label problem might overwhelm the default binary cross entropy loss function in training the binary classifier. When distinguishing the GGO and consolidation from overall lesion segment, the negative samples— which are the consolidations—are easier to classify than the positive samples (GGO) because the large number of negative samples contribute to the majority of the loss and have a huge influence on the gradient. Focal loss is an improved loss function that could reduce the weight of easy samples and focus more on samples in minority class. It can improve the performance of the classification network when the dataset has class imbalance [37].

In addition to choosing the loss function wisely, another important technical point for training the network is to configure the most advanced iterative method to optimize the loss function. Currently, many successful networks are trained using the stochastic gradient descent (SGD) algorithm [38], and its variants. To improve SGD, and other optimizers, a novel algorithm called Lookahead was proposed [39. It uses two nested loops to update two sets of network weights. The fast weights of the network are trained several times in a small inner loop using an optimizer such as SGD, then the direction of the gradient is used to update the slow weights using the outer loop [39]. It is almost guaranteed to achieve fast convergence with minimal computational overhead [39].

In the current study, we propose an advanced deep learning model called self-supervised InfNet (SSInfNet) which uses InfNet as a backbone, and integrates generative adversarial image inpainting, focal loss, and lookahead optimizer techniques (Fig. 1) to improve lung lesion segmentation performance compared to benchmark models. Furthermore, the clinical mechanisms of the predicted multi lung lesion segments (GGO segment and consolidation segment) on COVID-19 are evaluated in this study using statistic mediation analysis. The identified mediation effects could significantly increase the interpretability of the network and support certain image features as potential diagnostic image biomarkers for COVID-19.

Methods

Data split

Two COVID-19 datasets were involved in this study. One is the Integrative Resource of Lung CT Images and Clinical Features (ICTCF) [40] which contains the clinical severity for each patient. There are 6654 lung CT images from 1338 patients with their clinical severity in ICTCF. The other is Med-Seg (medical segmentation) COVID-19 dataset [36] which contains 932 CT lung images with the ground truth labels of their GGO and consolidation segments. The segments and data splits are shown in Fig. 2. We split the dataset into training set and test set. To prevent data leakage, we split the dataset based on the patients rather than the CT lung images.

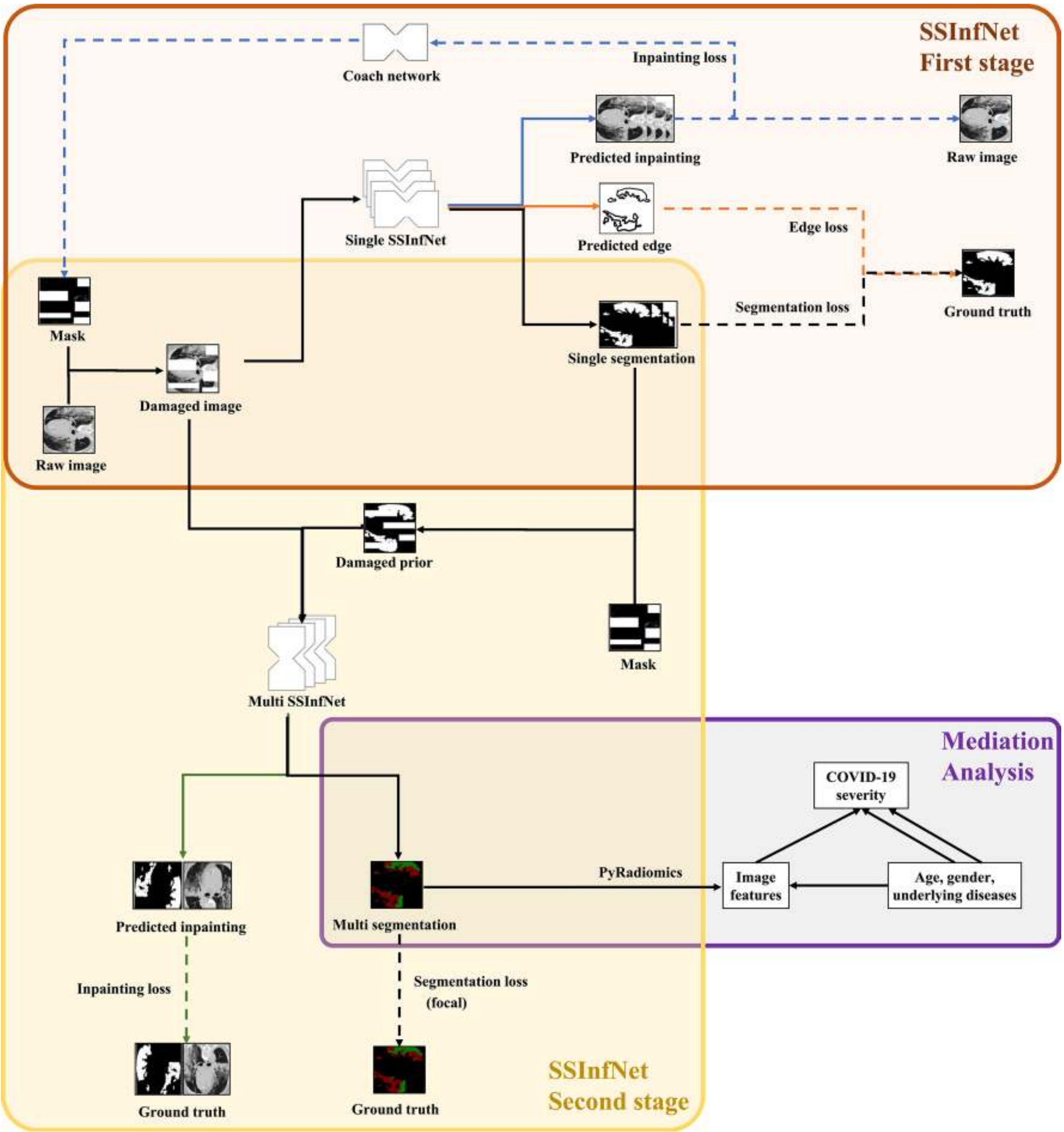
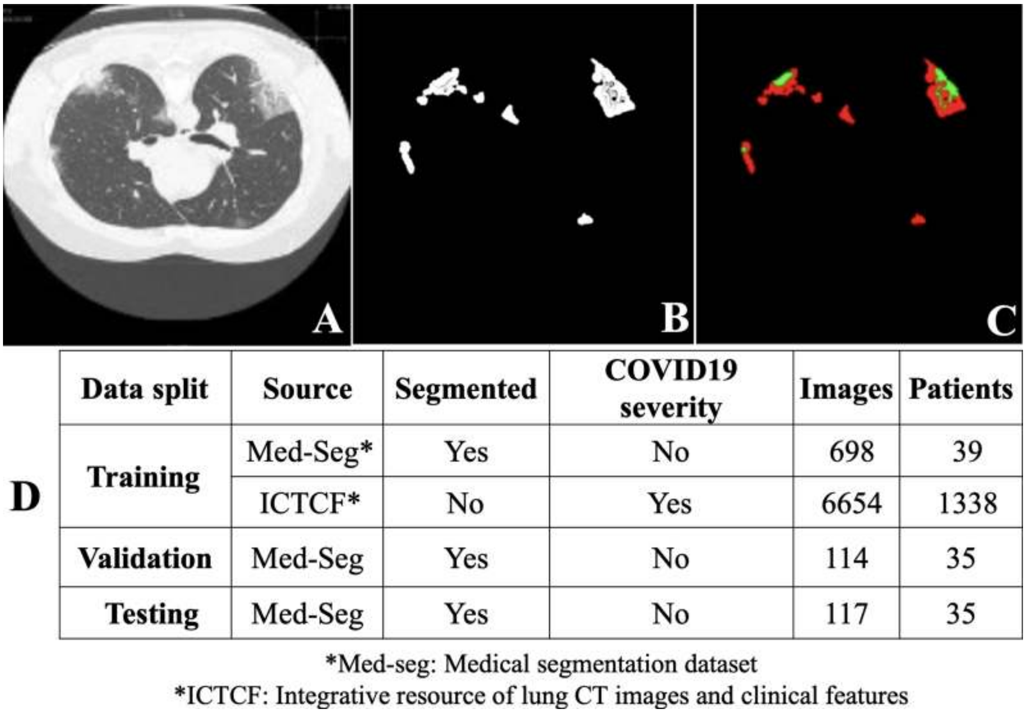


Fig. 1 Overview of the proposed self-supervised COVID-19 lung infection segmentation (SSInfNet) model and statistic causal mediation analysis of the predicted segments. The black path shows the main workflow of the proposed two-stage SSInfNet model and the follow-up statistical mediation analysis. The first stage is a single SSInfNet which takes the damaged CT image as input, and outputs the reconstructed image (blue path), the edges of overall lesion segment (orange path), and the single segment itself. The inpainting loss and edge loss are intended to increase the complexity of the single SSInfNet to improve its segmentation ability. The coach network (presented in the blue path) forms a generative adversarial mechanism with single SSInfNet to further improve the later model's performance. Continuing to proceed along the black path, the raw CT image and the predicted overall lesion segment (as prior) are used as input for the multi SSInfNet to further divide the overall lesion segments into ground-glass opacity and consolidation segments. Image inpainting is also involved in this stage (green path). For the multi segmentation, we use the focal technique as its loss function and lookahead optimizer as its training strategy. At the end, the predicted multi segments are used to extract several images features with Python's PyRadiomics [41] package. The image features act as mediators in the mediation analysis model between the independent variables (age, gender, and underlying diseases) and the dependent variable (COVID-19 severity)

Supervised InfNet

The supervised InfNet (SInfNet) is a recently developed CNN model for COVID-19 lung CT segmentation [22], which was used as both our backbone and one of the baseline models. We did not change the overall structure and default hyperparameters of the original SInfNet



Data split	Source	Segmented	COVID19 severity	Images	Patients
Training	Med-Seg*	Yes	No	698	39
	ICTCF*	No	Yes	6654	1338
Validation	Med-Seg	Yes	No	114	35
Testing	Med-Seg	Yes	No	117	35

*Med-seg: Medical segmentation dataset
*ICTCF: Integrative resource of lung CT images and clinical features

Fig. 2 Segment visualization and data split. **A** Examples of raw lung CT images in both Med-seg dataset and ICTCF dataset. Images are all in the axial view which looks down through the body. **B** The overall lesion segment. This is the label for the proposed single self-supervised COVID-19 network (SSInfNet) model for lung infection segmentation, and it exists only in Med-seg dataset. **C** The ground-glass opacity segment (red) and consolidation segment (green). This is the label for multi SSInfNet and it is also only available for the Med-seg dataset. **D** The table shows the data utilization in the development of the proposed SSInfNet models. As ICTCF does not contain segment labels, it was used only for the self-supervised image inpainting in the training stage. The Med-seg image data was split into training, validation, and testing sets, approximately under the ratio of 6:1:1. After the model was well developed, it was applied to the ICTCF dataset for further statistic mediation analysis because only ICTCF contains COVID-19 clinical severity information, which means Med-seg data was not used in the mediation analysis

model (Additional file 1: Figure S1). A complete SInfNet consists of two parts: a single SInfNet (Additional file 1: Figure 2A) and a multi SInfNet (Additional file 1: Figure 3A). The single SInfNet only predicts the infected region without classifying them more specifically. The input of the single SInfNet is a raw CT lung image and the output includes the edge contour of the overall lesion regions and four overall lesion region segmentations with different sizes as shown in Additional file 1: Figure S1. A CT lung image is first passed into the initial convolutional layers of the single InfNet to extract image features. Then, the features generated from the convolutional layer are fed into the partial decoder module, reverse attention module, and the edge detection module. The edge detection module is meant to help the network with the detection of the boundaries of the segmentation. The reverse attention and the partial decoder generate the segmentation of the infection regions of the CT lung images.

The prediction from the single SInfNet represents the overall infected regions and acts as a prior to be fed, concatenated with the original CT images, into the multi SInfNet. The multi SInfNet is used to predict multiple labeled segmentations. The segmentations include the predicted background, GGO, and consolidation.

Self-supervised InfNet

The self-supervised InfNet (SSInfNet) is our proposed COVID-19 segmentation CNN model, which, like the SInfNet, includes two parts, a single SSInfNet (Additional file 1: Figure S2B) and a multi SSInfNet (Additional file 1: Figure S3B). It is obtained by integrating generative adversarial image inpainting, focal loss, and Lookahead optimizer to SInfNet. The original SInfNet model generates 5 different predictions: an edge segmentation prediction and the 4 segmentations of the infected regions. To utilize the ability of a self-supervised method for the SInfNet's segmentation, we generate masks fed into the SInfNet model. The last convolution layer that outputs the prediction is not used for the self-supervised case. However, the last convolutional layer is replaced with a different convolutional layer to reconstruct the image and the edge appropriately. Everything else is kept the same as the SInfNet architecture (Additional file 1: Figure S2A). This process allows the network learns meaningful representations of the CT images. We can use these meaningful representations to segment the infected regions of CT lung images. After learning the self-supervised features for InfNet, the training continues as normal, similar to the SInfNet algorithm. The training starts with the weights trained using the self-supervised inpainting method. The last layer is changed to its original layer instead of the replaced convolutional layer.

By learning features from image inpainting, the model can learn features that are closer related to image segmentation. As creating masks can be a complex task for the network to learn to inpaint, the mask can either be too complex for the network to start learning or too simple to be able to learn good representations. We use a coach network that increases the complexity of the masking of the CT images throughout the training of the network. The mask created is initially simple, once the network can predict the inpainting of the CT images with good performance, the coach increases the complexity of the masking to reduce the performance of the network. The loss for the coach network is constructed from the loss of the image inpainting from the SSInfNet. The coach network and the SSInfNet both works together as a MinMax algorithm. The SSInfNet tries to minimize the loss to generate better image inpainting while the coach network tries to increase the loss of the image inpainting through generating more complex masks. In the beginning, the masks generated by the coach network are quite simple. Through the training of the coach network, as the SSInfNet gets better at predicting image inpainting, the coach network starts to generate more complex masks. The loss function for the coach network is:

$$L_{coach}(x) = 1 - L_{rec}(x \odot M) \tag{1}$$

where.

- M is the mask created by the coach network
- x is the CT lung image
- L_{coach} is the loss for the coach network
- L_{rec} is the loss for the reconstruction loss

A constraint is applied to this loss function because the coach network would just create a mask that masks all regions. After all, no context information would be present for the network to learn and a maximum loss is achieved. The constraint is:

$$\hat{B}(x) = B(x) - SORT(B(x))^{k|B(x)} \tag{2}$$

$$M = C(x) = \sigma(\alpha\hat{B}(x)) \tag{3}$$

The backbone, B, of the coach network has a similar network architecture as the InfNet models. SORT*(B(x))* sorts the features in descending order over the activation map. k represents the k^{th} elements in the sorted list and k helps to control the fraction of the image to be erased. The regions that have scores smaller than the k^{th} element are erased from the images. If k is 0.75, then 0.75 percent of the image is not erased. The score is scaled into a range of [0,1] using a sigmoid, σ, activation function. *C(x)* is the coach network that is fed with the CT lung images. The illustration of the coach network can be seen in Fig. 1.

After the self-supervision training is finished, the single SSInfNet is reused to train normally, using the segmentation of the CT lung images. Likewise, the multi SSInfNet network reuses the weights that are trained during the self-supervised multi SSInfNet to train normally, using the multi segmentations of the CT lung images.

The proposed single SSInfNet architecture can be seen in Fig. 1 and Additional file 1: Figure S2B. Additional file 1: Figure S2A shows the original single SInfNet architecture. The difference is that the last layer for each output prediction is replaced with a different linear activation layer. The linear activation layer recreates the original image that is covered by the masks. The proposed multi SSInfNet architecture is shown in Fig. 1 and Additional file 1: Figure S3B. The changes in the architecture for the multi SSInfNet are similar to the single SSInfNet where the last convolutional layer is replaced with a different linear activation layer to output the inpainting of the original image.

A loss is calculated for each of the outputs of the single SInfNet model. The first loss function is the edge loss, L_{edge}, which guides the model in representing better segmentation boundaries. The other loss function is the segmentation loss, L_{seg}. The segmentation loss combines both the loss of Intersection over Union (IoU) and the binary cross entropy loss (*LBCE*). The segmentation loss equation for the single SInfNet is as follow:

$$L_{seg} = L_{IoU} + \lambda L_{BCE} \tag{4}$$

λ is a hyperparameter that controls how much weight we want to put on the binary cross entropy loss. The segmentation loss is adapted to all S_i predicted output where S_i are created from f_i such that $i = 3, 4, 5$. As low-level features use more computational resources due to larger spatial resolutions but achieves lesser performance. We use the features in the higher level ($i = 3, 4, 5$) instead. The total loss function for the single SInfNet model is then:

$$L_{total} = L_{seg}(G_t, S_g) + L_{edge} + \sum_{i=3}^{5} L_{seg}(G_t, S_i) \quad (5)$$

The summation of the segmentation loss functions is calculated from the output of the parallel partial decoder and the three convolutional layers ($i=3$, 4, 5)). G_t refers to the ground truth labels. S_g is the output from the parallel partial decoder to match with the ground truth label. S_i is the different sizes of the segmentation of infected regions output by the InfNet. The different sizes of the segmentation of infected regions outputted by the SSInfNet are resized to the same shape as the ground truth segmentation image.

As for the multi SSInfNet, we use the default model and hyperparameters from the multi SInfNet. However, we train the multi SSInfNet without using any unlabeled images during self-supervision because the multi SSInfNet requires the prior (infected region) as input. The CT lung images and prior (infected region) for the CT lung images are concatenated together before being fed into the multi SSInfNet. The prior is generated from the single SInfNet. The multi SSInfNet labels the prior with background, ground-glass opacities, and consolidations. The loss function for the multi SSInfNet is as follow:

$$L_{bce} = \frac{1}{N} \sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \quad (6)$$

Where,

- y_i is the ground truth value for the segmentation—background, ground-glass opacities, or consolidation
- $\widehat{y_i}$ is the network's predicted value for the segmentation
- N is the total number of the current training batch of data samples

The loss function for the multi SInfNet uses the binary cross-entropy loss between the predicted and the ground truth segmentation. In order to improve the performance of the model and to aid in its generalization, we chose to use self-supervised learning to learn good representations of the CT lung images.

Additionally, we use the focal loss instead of the binary cross-entropy loss function for the Multi SSInfNet model to provide more weight on the smaller data label samples (consolidation). The focal loss function is:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \quad (7)$$

where.

- FL is the focal loss
- p_t is the Multi SSInfNet's predicted output
- α_t is a hyperparameter that controls the weight of positive and negative samples
- γ is the term that controls the rate of the down-weighed examples

We also wrap the Lookahead optimizer around the SGD optimizer with $k=5$ and alpha $=0.5$. k is the number of inner-loops the SGD will optimize before the Lookahead optimizer starts optimizing. In our case, after the SGD optimizes the network weights for 5 iterations, the Lookahead optimizer will optimize using alpha multiplied by the difference between the network weights after the 5 iterations of SGD optimizer and the network weights before the 5 iterations of the SGD optimizer. The alpha is used to control the intensity of the difference. The pseudo code for our single/multi SSInfNet can be found in Additional file 1: Algorithm 1.

Experimental settings

For the Single SInfNet, we train the network for 500 epochs. We use Adam as the optimizer with a learning rate of 0.0001. For the Multi SInfNet, we train the network for 500 epochs. We use SGD as the optimizer. The momentum is set as 0.7 and the learning rate is set as 0.01. As for the Multi SSInfNet with added focal loss and lookahead optimizer, we train the network for 500 epochs, use lookahead optimizer with $k=5$ and alpha $=0.5$, and wrap the Lookahead optimizer around the SGD optimizer where the momentum is set as 0.7 and the learning rate is set as 0.01.

For the self-supervised version of both the Single SInfNet and Multi SInfNet, the self-supervised image inpainting is first trained. Then the weights from the trained networks, except for the last layer, are transferred and be used to train on the segmentation of the CT lung images. During the self-supervised image inpainting stage, we train the network for 2000 epochs. The network is trained for the first 200 epochs before we train the coach network for 200 epochs which increases the complexity of the masks generated. After that, we alternate in between training the self-supervised image inpainting for 100 epochs and the coach network for 100 epochs for 1800 epochs in total. For every alternating between the training of the self- supervised image inpainting and the coach network, we set the learning rate to 0.1 at the start of the epoch, 0.01 at the 40th epoch, 0.001 at the 80th epoch, and 0.0001 at the 90th epoch to speed up convergence. We use SGD as the optimizer for the self-supervised image inpainting, set the momentum to 0.9 and the weight decay to 0.0005. As for the optimizer for the coach network, we use the Adam optimizer with a learning rate of 0.00001.

We compare our self-supervised method against some supervised models trained on the COVID-19 datasets. We train and follow the same network structure, but change from supervised learning to self-supervised learning, and compare the performance between the supervised and the self-supervised approaches. We want to determine if self-supervised learning is a useful way to help the SInfNet improve its performance in segmenting the ground-glass opacities or consolidation around the infected region of the CT lung images.

Performance evaluation metrics

Five metrics are used to measure the models' performance: F1, intersection over union (IoU), Recall, and Precision and the area under the curve (AUC) of a receiver operating characteristic (ROC):

The F1-Score is also called the Dice Coefficient. It is used to measure the overlap between the ground-truth infected region and the predicted infected region. The F1-Score equation is defined as:

$$F1 = \frac{2 * |T \cap P|}{|T| + |P|} \tag{8}$$

where T is the ground truth infected region and P is the predicted infected region.

The Intersection over Union (IoU) is a different method to measure the overlap between the ground truth infected region and the predicted infected region. The IoU equation is defined as:

$$IoU = \frac{T \cap P}{T \cup P} \tag{9}$$

where T is the ground truth infected region and P is the predicted infected region.

The Recall is used to measure how much of the ground truth infected region is present in the predicted infected region. The equation is as follow:

$$\mathrm{Re}call = \frac{T \cap P}{T} \tag{10}$$

where T is the ground truth infected region and P is the predicted infected region.

The Precision is used to measure how much of the predicted infected region is present in the ground truth infected region. The equation is as follow:

$$\mathrm{Pr}\,ecision = \frac{T \cap P}{P} \tag{11}$$

where T is the ground truth infected region and P is the predicted infected region.

For each of above performance metrics, we first perform the calculation within each test sample, separately. We compute the mean and the error interval of the estimated mean for each of the metrics in the entire test set. The mean is defined as:

$$mean = \frac{\sum_{i=1}^{N} Metric(\hat{y}_i, y_i)}{N} \tag{12}$$

where Metric refers to F1, IoU, Recall, Precision or AUC. N refers to the number of test samples. The error is defined as:

$$error = SE \times 1.96 \tag{13}$$

where SE is the standard error of the test samples for the given metric. The error interval of the estimated mean is defined as $-error$ and $+error$.

Generation of image phenotypes

The well-trained multi SSInfNet outputs three kinds of image-level segments: the overall lesion segments, the GGO segments, and the consolidation segments. These image-level segments act as masks in the Python radiomic package PyRadiomics [41] for extracting image phenotypes, separately. Three runs of phenotype extraction are executed with the inputs of overall lesion segments plus the original images, the GGO segments plus the original images, and the consolidation segments plus the original images, respectively. We select first order measurements, such as Gray Level Co-occurrence Matrix (GLCM) measurement, Gray Level Dependence Matrix (GLDM) measurement, and Neighboring Gray Tone Difference Matrix (NGTDM) measurement, as our image phenotypes. The definition and formulas of these image phenotypes can be found in Additional file 1: Table S1. After the segments-based image-level phenotypes are generated, we take the average of them to make the image phenotypes at patient-level.

Mediation analysis

Univariate mediation analyses are performed to determine the potential causal mechanism in which age, gender, or underlying diseases is associated with COVID-19 severity through an intermediate image phenotype. Let y be the dependent variable which is the binarized COVID-19 severity. In the original ICTCF dataset, severity is measured with 9 levels: Control (Healthy), Control, Control (Community-acquired pneumonia), Suspected, Suspected (COVID-19-confirmed later), Mild, Regular, Severe, and Critically ill. We code these 9 levels of the severity into 2 levels by grouping Control (Healthy), Control, Control (Community-acquired pneumonia), Suspected into one group (coded as 0) and Suspected (COVID-19-confirmed later), Mild, Regular, Severe, and Critically ill into another group (coded as 1). Let m be a mediator (patient-level image phenotype), x be an independent variable (age, gender, or underlying diseases). Hence, we can fit the below regression models [12]:

$$\begin{aligned} y &= \beta_{10} + \beta_{11}x + \in_1 \\ m &= \beta_{20} + \beta_{21}x + \in_2 \\ y &= \beta_{30} + \beta_{31}x + \beta_{32}m + \in_3 \end{aligned} \tag{14}$$

Here, β and ϵ are the parameters of the models to be estimated and tested. βs are the coefficients of variables, while ϵ are the residuals. If the abovementioned three regressions are significant (adjusted p-value < 0.05) and $|\beta_{11}| > |\beta_{31}|$, we say that x is associated with y, mediated through m, which provides a potential mechanism explanation of how x has influence on y through m. The indirect effect of x on y through m is defined as $\beta_{21} \times \beta_{32}$.

For multiple mediation analysis, we first perform a pair-wise correlation analysis of the significant mediators from the univariate mediation analysis using the R package, corrplot [42], to control the potential confounding influence on the multiple mediation analysis. The mediator pairs that have absolute correlation coefficient greater than 0.8 are first identified. Then, one phenotype within each of these correlated pairs is removed. The filtering criteria include both less indirect effect or less commonly used in medical research. The remaining mediators with the two independent variables (age and the underlying diseases) are input into a multiple mediation model for further identifying the indirect effect when controlling for each other using R package lavaan [43. Lavaan is a tool for structure equation modeling (SEM) which is a very general and powerful multivariate technique. SEM uses conceptual model, path diagram and linked regression-style equations to model complex relationships among a network of variables. Thus, it allows multiple independent variables and mediators, even multiple dependent variables in the model [43–45]. We build our equation system as below for our special case (two independent variables, one dependent variable, and several mediators linked to different independent variables.):

$$\begin{aligned} y &= \theta_{00} + \theta_{01}x_1 + \theta_{02}x_2 + \varepsilon_0 \\ M_c &= \theta_{c0} + \theta_{c1}x_1 + \theta_{c2}x_2 + \varepsilon_c \\ M_a &= \theta_{a0} + \theta_{a1}x_1 + \varepsilon_a \\ M_u &= \theta_{u0} + \theta_{u2}x_2 + \varepsilon_u \\ y &= \theta_{(n+1)0} + \theta_{(n+1)1}x_1 + \theta_{(n+1)2}x_2 + \theta_{(n+1)3} \\ M_1 &+ \ldots + \theta_{(n+1)(n+2)}M_n + \varepsilon_{n+1} \end{aligned} \tag{15}$$

where x_1 is the age, x_2 is the underlying disease. M_c is the significant mediators for both age and underlying disease. M_a represents the significant mediators for age, while M_u refers to the significant mediators for underlying diseases. The θs are the coefficients which are estimated and tested when the model is fit. ε are the residuals.

Sensitivity analyses

A series of sensitivity analyses are performed to further support our conclusions. These analyses include: a three-fold cross validations performed using both single SSInfNet and multi SSInfNet to ensure that the performance is consistent, a comparison with transfer learning- based FCN8 segmentation network [46], further experiments on other independent datasets [47] to show the generalization ability of our models, ablation studies to explore which techniques (generative adversarial image inpainting, focal loss, and lookahead optimizer) we use in the multi SSInfNet model contribute to the improved performance, and a computation cost analysis to show the difference between the different models' computation efficiency. The details of these analyses could be found in Additional file 1: Sensitivity Analysis.

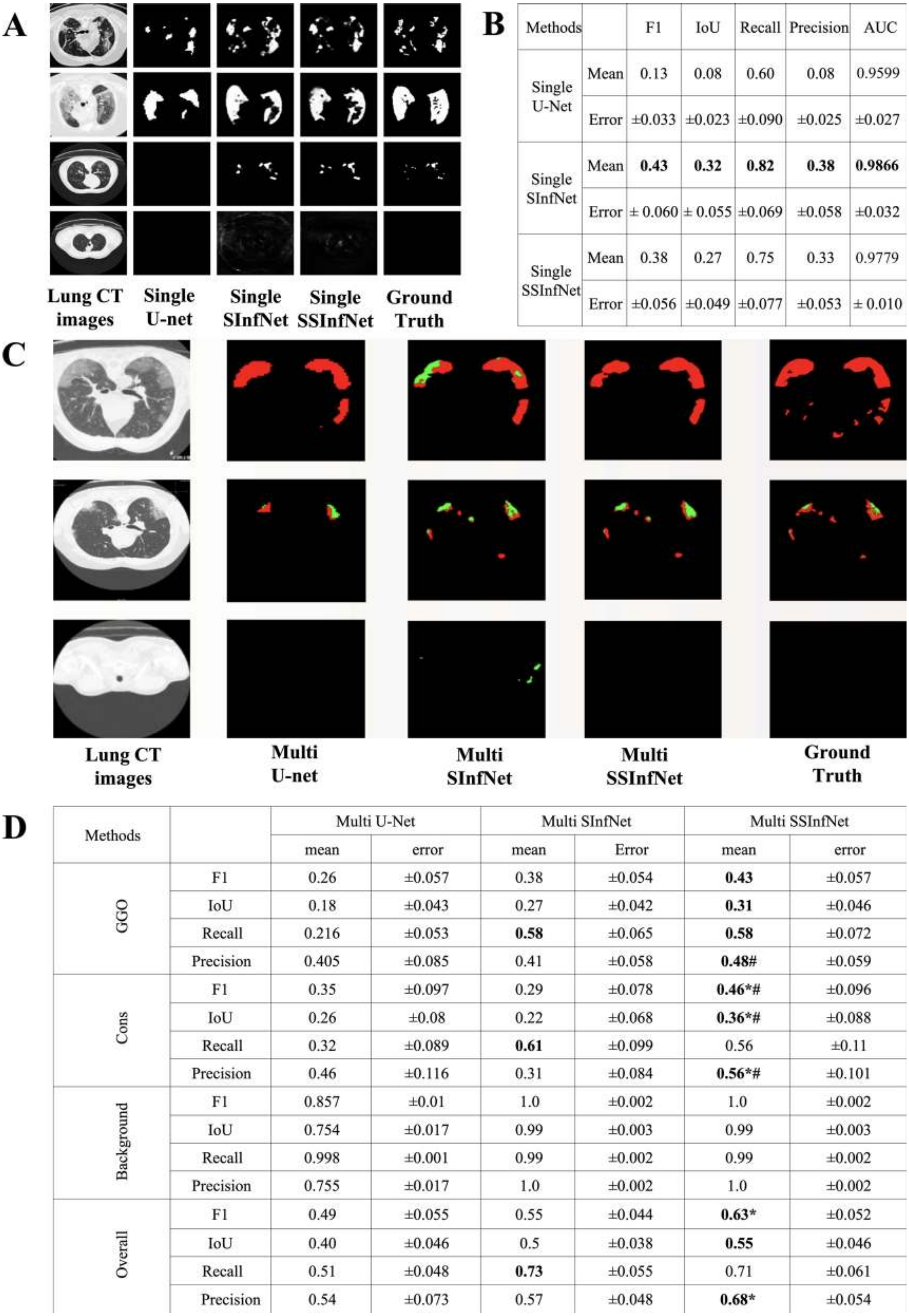
Results

Single SSInfNet

The segmentation performance of the proposed single SSInfNet and the two baseline models (single U-net and single SInfNet) can be found in Fig. 3A and B. In this stage, the models do not segment either GGO or consolidation. They segment and represent the entire infected region as one overall lesion. U-Net [48] and supervised

(See figure on next page.)

Fig. 3 Visual comparison and quantitative comparison of segmentation results among different networks. **A** Four examples of the original lung CT images, their overall segments predicted by three different networks and the ground truth overall lesion annotation. The two baseline models are the single U-net and the single SInfNet (supervised COVID-19 lung infection segmentation) model. The proposed model is the single SSInfNet (self-supervised COVID-19 lung infection segmentation) model. **B** The mean and error of five quantitative model performance metrices calculated from the 35 test samples. **C** Three examples of the original lung CT images, their GGO and consolidation segments predicted by three different networks and the ground truth lesion annotations. The two baseline models are the multi U-net and the multi SInfNet models. The proposed model is the multi SSInfNet model. **D** The mean and error of five model performance metrics calculated from the 35 test samples. The Overall showed the averaged performance for GGO, consolidation, and background



Methods		F1	IoU	Recall	Precision	AUC
Single U-Net	Mean	0.13	0.08	0.60	0.08	0.9599
	Error	±0.033	±0.023	±0.090	±0.025	±0.027
Single SInfNet	Mean	**0.43**	**0.32**	**0.82**	**0.38**	**0.9866**
	Error	± 0.060	± 0.055	±0.069	±0.058	±0.032
Single SSInfNet	Mean	0.38	0.27	0.75	0.33	0.9779
	Error	±0.056	±0.049	±0.077	±0.053	± 0.010

Methods		Multi U-Net		Multi SInfNet		Multi SSInfNet	
		mean	error	mean	Error	mean	error
GGO	F1	0.26	±0.057	0.38	±0.054	**0.43**	±0.057
	IoU	0.18	±0.043	0.27	±0.042	**0.31**	±0.046
	Recall	0.216	±0.053	**0.58**	±0.065	**0.58**	±0.072
	Precision	0.405	±0.085	0.41	±0.058	**0.48#**	±0.059
Cons	F1	0.35	±0.097	0.29	±0.078	**0.46*#**	±0.096
	IoU	0.26	±0.08	0.22	±0.068	**0.36*#**	±0.088
	Recall	0.32	±0.089	**0.61**	±0.099	0.56	±0.11
	Precision	0.46	±0.116	0.31	±0.084	**0.56*#**	±0.101
Background	F1	0.857	±0.01	1.0	±0.002	1.0	±0.002
	IoU	0.754	±0.017	0.99	±0.003	0.99	±0.003
	Recall	0.998	±0.001	0.99	±0.002	0.99	±0.002
	Precision	0.755	±0.017	1.0	±0.002	1.0	±0.002
Overall	F1	0.49	±0.055	0.55	±0.044	**0.63***	±0.052
	IoU	0.40	±0.046	0.5	±0.038	**0.55**	±0.046
	Recall	0.51	±0.048	**0.73**	±0.055	0.71	±0.061
	Precision	0.54	±0.073	0.57	±0.048	**0.68***	±0.054

Fig. 3 (See legend on previous page.)

InfNet [22 (SInfNet) were selected as baseline models for comparing performance with our proposed SSInfNet. U-Net is a classical CNN and is often used as baseline or backbone of segmentation networks [49–53], while the SInfNet is our backbone model and was developed to solve the same COVID-19 segmentation problem. Five classical metrices (F1, IoU, Recall, Precision and AUC of the receiver operating characteristic) were used to quantitatively measure the networks' performances. As the prediction is at the pixel level, we calculated the performance metrices at the sample level instead of the entire test set. Therefore, the mean and error for each of these performance metrices in the entire test set were shown in Fig. 3B. Observed from Fig. 3B**,** the proposed single SSInfNet and the baseline single SInfNet achieved comparable performances. The overall AUC and error based on the single SSInfNet is comparable to that of the single SInfNet (Additional file 1: Figures S3B, S4; Table S2), and both models outperform the baseline single U-net (Additional file 1: Figure S4) in terms of the overall AUC.

Even though the baseline single SInfNet has better mean values for F1, IoU, and Recall, the self-supervised approach helps create robustness and consistency in the model to better handle outliers. This can be observed by the fact that in Fig. 3A, the baseline single SInfNet overestimated the overall infected region of an outlier image (the last row) while the single SSInfNet did a better job at predicting outliers, where its prediction is more closely related to the ground truth than that of the baseline single SInfNet.

Even though the baseline single SInfNet has better mean values for F1, IoU, and Recall, the self-supervised InfNet approach helps handle some outliers in a better way. This can be observed by the fact that in Fig. 3A, the baseline single SInfNet overestimated the overall infected region of an outlier image (the last row) while the single SSInfNet did a better job at predicting outliers, where its prediction is more closely related to the ground truth than that of the baseline single SInfNet.

Multi SSInfNet

Figure 3C and D show the multi SSInfNet performance. This is the second stage of solving the proposed multi segmentation problem. In this stage, the network breaks down the previous overall segments predicted by the single SSInfNet into two parts, the GGO and the consolidation. The overall segments from the single SSInfNet act as a prior and is fed, along with the CT lung images, into the multi SSInfNet. The output is a pixel level 3-channel matrix with each cell in one channel annotating the probability of being GGO, another channel annotating the probability of being consolidation, and the last channel annotating probability of being background. Again, the multi U-net and multi SInfNet were used as baseline models for comparison. The proposed multi SSInfNet was able to achieve better performance than the baseline multi U-net and multi SInfNet. As visualized in Fig. 3C, the multi SSInfNet achieved better performance in evaluating the GGO and the consolidation areas of the CT lung images than the multi SInfNet and the multi U-net, in terms of predicting the visible most similar segments to the ground truth. However, as Fig. 3D shows, the baseline multi SInfNet achieved a better recall than the rest of the networks. But, as we can see in Fig. 3C, multi SInfNet predicted more consolidation segments, even in areas that were not infected. On the third row of Fig. 3C, the multi SInfNet overestimated the consolidation region in a CT lung image from a healthy individual. Since recall is the true positives over the total actual consolidation area, the multi SInfNet seems to overestimate the consolidation area which results in a higher recall than the other networks. This explains well why the precision for the SInfNet is lower, as most of its prediction of the consolidation area is not accurate. Hence, this decreases the performance of the multi SInfNet while the proposed multi SSInfNet handled this problem very well.

Mediation analysis

A total of 204 averaged image phenotypes for 1338 patients were created from the output of the proposed multi SSInfNet using Python's radiomic package PyRadiomics [41]. The PyRadiomic algorithm needs a mask and the original image as input, and it outputs a list of continuous measures as image phenotypes. According to different measure approach, these image phenotypes can be categorized into first order features, Gray Level Co-occurrence Matrix (GLCM) features, Gray Level Dependence Matrix (GLDM) features, Neighboring Gray Tone Difference Matrix (NGTDM) features and so on. Three kinds of image segments (GGO, consolidation, and overall lesion) were separately input into the PyRadiomic algorithm as the mask of the original image. Each mask contributed 68 image phenotypes in four of the above defined phenotype categories, as listed in Additional file 1: Table S1.

R's package, lavaan, was used to conduct the univariate mediation analysis for each combination of a given independent variable, dependent variable, and image phenotype. The independent variables include age (continuous), gender (binary), and the number of underlying diseases (continuous). The dependent variable is the binarized COVID-19 severity. All consolidation-based image phenotypes were not significant (adjusted p-value > 0.05) in the univariate mediation analysis. Therefore, the results of the consolidation- and overall lesion (consolidation + GGO)- based image phenotype analyses are not reported here. Among all 68 GGO- based image phenotypes, 37 of them have significant (adjusted p-value < 0.05) mediation effects on COVID-19 severity. Thirty-two of these 37 mediators were mediating the indirect effect of underlying disease on COVID-19 severity, while 27 of these 37 mediators were mediating the indirect effect of age on COVID-19 severity. The results, including the categories and names of the image phenotypes, the adjusted p-value of the indirect effects and the estimated coefficients for these image phenotypes, can be found in Fig. 4A (for age variable) and Fig. 4B (for underlying disease variable), respectively.

The correlations among these mediators were calculated based on their patient-level averaged feature values, and the results are demonstrated in Fig. 5. Mediators with high absolute correlation coefficients are more linearly dependent and hence have similar effect on COVID-19 severity. So, the correlation between each pair of the image phenotypes was compared and one of the two phenotypes was removed if their absolute correlation coefficient was greater than 0.8. It was suggested by the PyRadiomics that some features are the confounding of the segment area [41. Thus, only the area variable was kept among those phenotypes. Since the entropy measures the uniformity and is more widely incorporated in medical image phenotype related researches [54], the entropy phenotype was kept while the uniformity phenotype was removed. The MCC (Maximal Correlation Coefficient) and IMC1 (Informational Measure of Correlation) measures the complexity of the texture. We decided to keep IMC1 because it has been widely used and reported in lung cancer studies [55, 56]. After the correlation filtering, 3 mediators (Entropy, Kurtosis, and Skewness) were left for the underlying disease variable and 5 mediators (Mean, Area, Entropy, Kurtosis, and IMC1) were left for the age variable. The remaining 3 and 5 mediators were input into the mediation analysis models with multiple mediators. The results can be found in the path plot (Fig. 6).

Sensitivity analyses

From the results of the three-fold cross-validation (Additional file 1: Table S2), we can see that the baseline single SInfNet performs the best for most of the performance metrics. Self-supervised SSInfNet does not show an improvement in the single segmentation for the CT lung images. However, as shown in Additional file 1: Table S3, the multi SSInfNet shows a better performance than both the multi U-Net and the multi SInfNet in the three-fold cross-validation, suggesting that the multi SSInfNet can generalise well in segmenting GGO and consolidation regions of the CT lung images. These three-fold cross-validation based results are consistent with those shown in Fig. 3B (single InfNet) and 3D (multi InfNet) based on the training, test, and validation strategy used to develop the models.

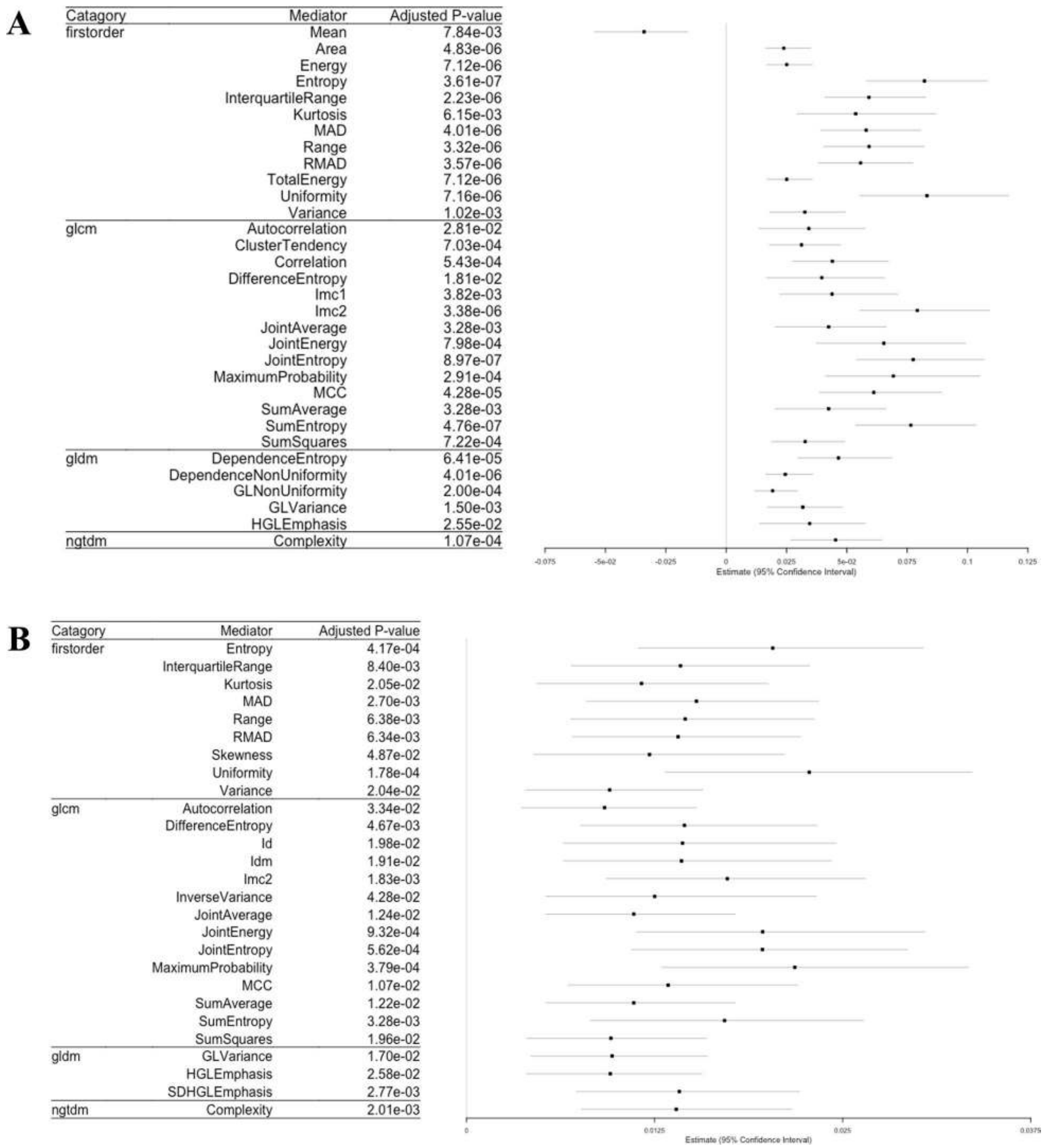
As shown in the Additional file 1: Table S4, the proposed new method, multi SSInfNet, achieves the best performance among the multi FCN8 (with pre-trained weights), mullti U-Net, and the baseline multi SInfNet models. Further experiments on other independent datasets were applied to evaluate the generalization ability of our models (Additional file 1: Tables S5, S6). For the independent data set 2 (see Additional file 1: Additional data sets), due to the nature of the very small dataset on which we tested the methods, the dataset did not replicate a good generalisation behaviour from the methods that were trained on (Additional file 1: Table S5A (Single InfNet) and S5B (Multi InfNet)). However, for a relatively larger independent data set 3 (see Additional file 1: Additional data sets), we can see that its results (Additional file 1: Table S6B) are consistent with our current dataset where our multi SSInfNet shows a better performance in segmenting the CT lung images (Fig. 3D) than the baseline multi SInfNet (Fig. 3B). The results that we obtained from the baseline SInfNet and the SSInfNet can be found in Additional file 1: Table S6A.

The results of the ablation studies can be found in Additional file 1: Table S7. We can see that all the additional techniques added on the baseline SInfNet have improved performance on the segmentation of the CT lung images. They compensate with each other and then achieve a higher performance.

The computational cost analysis of different models is shown in the Additional file 1: Table S8. Overall, the time taken to process 1 image for baseline multi SInfNet is 1.06 times longer than the time taken to process 1 image for multi SSInfNet. The multi FCN8 has the best computation efficiency with 0.74 times shorter time than the time taken by the baseline multi SInfNet.

Discussion

Lung CT imaging was proposed as a backup diagnosis and monitoring tool for COVID-19 in emergency breakouts when PCR kits are not available [57, 58]. Others also suggested that PCR testing and lung CT imaging should be used together for routine COVID-19 diagnosis and prognosis to enhance the clinical protocol of COVID-19 [58, 59]. Some radiologists suggested to avoid using confirmative statement about COVID-19 identification from lung CT imaging because CT may not be able to distinguish among different viruses [60], but a recent study showed that 7 radiologists successfully identified COVID-19 distinguished from other viral infections with 93–100% specificity based on lung CT imaging features [61]. In many studies, lung CT imaging showed



A

Catagory	Mediator	Adjusted P-value
firstorder	Mean	7.84e-03
	Area	4.83e-06
	Energy	7.12e-06
	Entropy	3.61e-07
	InterquartileRange	2.23e-06
	Kurtosis	6.15e-03
	MAD	4.01e-06
	Range	3.32e-06
	RMAD	3.57e-06
	TotalEnergy	7.12e-06
	Uniformity	7.16e-06
	Variance	1.02e-03
glcm	Autocorrelation	2.81e-02
	ClusterTendency	7.03e-04
	Correlation	5.43e-04
	DifferenceEntropy	1.81e-02
	Imc1	3.82e-03
	Imc2	3.38e-06
	JointAverage	3.28e-03
	JointEnergy	7.98e-04
	JointEntropy	8.97e-07
	MaximumProbability	2.91e-04
	MCC	4.28e-05
	SumAverage	3.28e-03
	SumEntropy	4.76e-07
	SumSquares	7.22e-04
gldm	DependenceEntropy	6.41e-05
	DependenceNonUniformity	4.01e-06
	GLNonUniformity	2.00e-04
	GLVariance	1.50e-03
	HGLEmphasis	2.55e-02
ngtdm	Complexity	1.07e-04

B

Catagory	Mediator	Adjusted P-value
firstorder	Entropy	4.17e-04
	InterquartileRange	8.40e-03
	Kurtosis	2.05e-02
	MAD	2.70e-03
	Range	6.38e-03
	RMAD	6.34e-03
	Skewness	4.87e-02
	Uniformity	1.78e-04
	Variance	2.04e-02
glcm	Autocorrelation	3.34e-02
	DifferenceEntropy	4.67e-03
	Id	1.98e-02
	Idm	1.91e-02
	Imc2	1.83e-03
	InverseVariance	4.28e-02
	JointAverage	1.24e-02
	JointEnergy	9.32e-04
	JointEntropy	5.62e-04
	MaximumProbability	3.79e-04
	MCC	1.07e-02
	SumAverage	1.22e-02
	SumEntropy	3.28e-03
	SumSquares	1.96e-02
gldm	GLVariance	1.70e-02
	HGLEmphasis	2.58e-02
	SDHGLEmphasis	2.77e-03
ngtdm	Complexity	2.01e-03

Fig. 4 Significant image phenotypes in the univariate mediation analyses. **A** Forest plot showing the 32 mediators of age's indirect effect on COVID-19 risk. **B** Forest plot showing the 27 mediators of underlying disease's indirect effect on COVID-19 risk

comparable sensitivity with PCR in diagnosis of COVID-19. The sensitivity of PCR ranges from 42 to 71% [57, 60, 62], while the reported sensitivity of lung CT imaging-based diagnosis for COVID-19 varies from 60 to 90% [57, 60, 62–64]. There are also some concerns about the cost of lung CT imaging as the PCR test is much cheaper than lung CT scan [65]. Although the cost of healthcare services is complex, the major cost of a medical imaging-based test is spent on using radiologists for image reading [66]. Hence, it is a promising area to introduce AI into

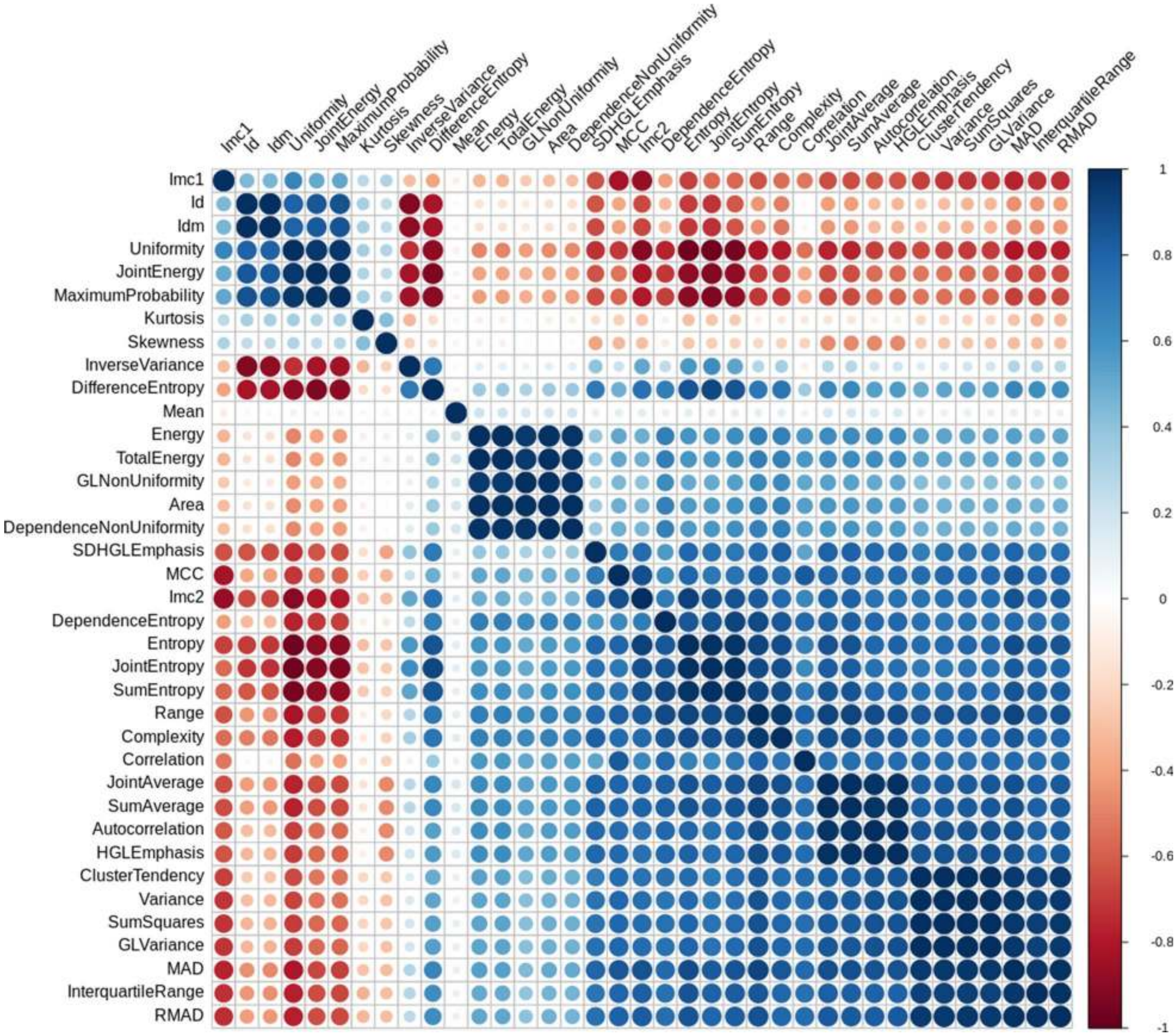


Fig. 5 Hierarchical clustering of the ordered correlation matrix of the 37 image phenotypic mediators from the univariate mediation analysis. The color represents the correlation coefficient

lung CT imaging-based diagnosis to assist radiologists. Furthermore, it is possible to increase the sensitivity and specificity of PCR alone or human involved lung CT imaging COVID-19 diagnosis.

When introducing AI into healthcare, the accuracy of the model is not the whole picture. The interpretability and transparency of the model should always be kept in mind. Lung CT image segmentation, in its very core spirit, is to split human understandable regions from the less informative background regions to assist people's decision-making. Therefore, lung CT image segmentation model makes more clinical sense than a binary classifier which takes the raw lung CT image as input and gives a simple yes or no answer especially. The computationally segmented lung regions could be further analyzed for biological explanations and diagnostic values.

InfNet is able to achieve a competent performance on segmentation of infected region for CT lung images. The authors of the InfNet further extended the network by introducing semi-supervised learning to InfNet. They generated pseudo labels and utilized the pseudo labels to train their model using a two-step strategy. However, their method of semi-supervised by generating pseudo-labels when used in the dataset that we used takes a few months to finish generating the pseudo-labels. This is not feasible in real world applications. Hence, here we

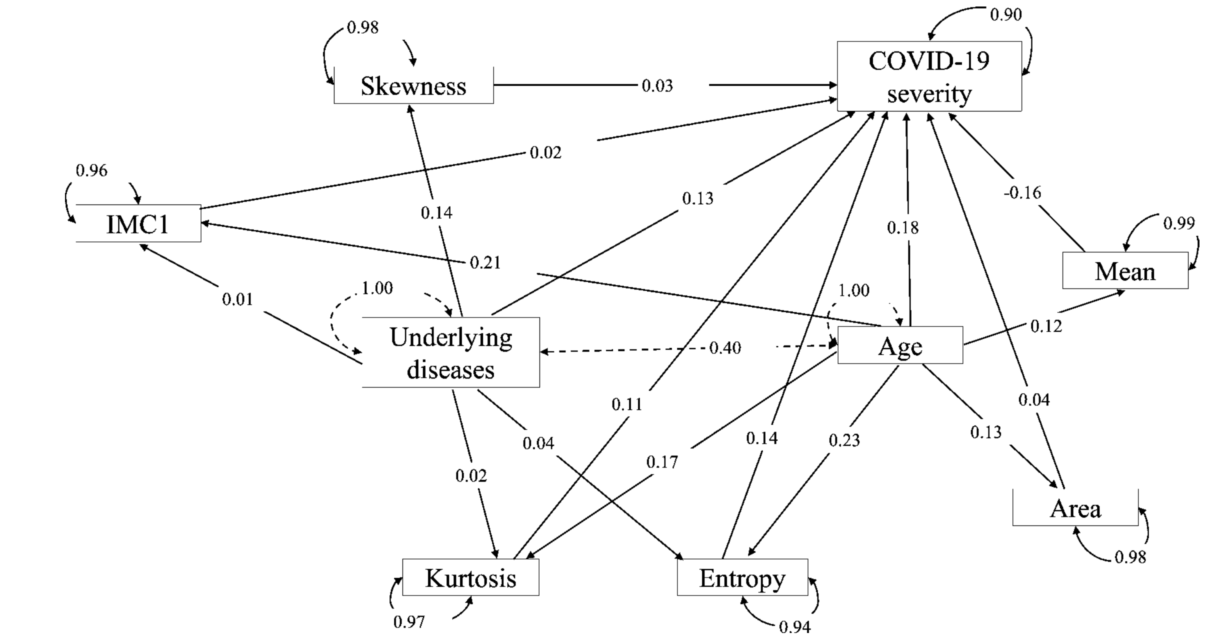


Fig. 6 Path plot of the mediation analysis model with multiple mediators. The standardized effect estimates of each variable are shown on the edges of the paths. The mediators are Entropy, Kurtosis, Skewness, Mean, Area, and IMC1. Dependent variable is the COVID-19 severity, and independent variables are the Underlying disease and Age. Curves with arrowheads on both sides is the standardized residual variance. Solid curve is for dependent variable and mediators. Dash line curve is for independent variables. Straight dash line represents the standardized covariance of two independent variables. Straight solid line with arrowhead on one side is the standardized effect estimate

propose to use a self-supervised learning strategy. The self-supervised method creates a huge speed up improvement when compared to their method of semi-supervised learning.

To increase the model performance, self-supervised image inpainting was used in this study for model pretraining, focal loss was used to replace the traditional cross entropy loss, Lookahead optimizer was used along with SGD to manage the training iteration. The integration of these advanced techniques achieved the best model performance, as compared to other baseline models. The proposed Multi SSInfNet model is better at dealing with outliers and makes fewer false positive in predicting the minority class, which, for COVID-19, is the consolidation lesion.

To enhance the interpretation of the proposed Multi SSInfNet model, we further extracted the lung imaging phenotypes from the output of the model and applied statistical mediation analysis to explore the potential causal association of the patients' age, gender, and underlying diseases with COVID-19 severity through the identified lung CT imaging mediators. We showed that 8 image phenotypes from the predicted GGO segments were significantly correlated with COVID-19 severity and the age or underlying disease(s) of a patient with COVID-19. The entropy and kurtosis of the computational GGO segments have a positive mediation effect on both underlying diseases caused COVID-19 severity and age caused COVID-19 severity. Entropy represents the uncertainty of the pixel values within the GGO segment. A higher entropy indicates a more chaotic GGO segment [41]. Kurtosis measures the peakedness of the GGO pixel value distribution. A higher kurtosis implies that the GGO pixel values are concentrated towards the tails rather than towards the mean [41]. This means that elders or people with sever underlying diseases will probably have more chaotic and peakier distributed GGO segments, and thus will suffer from severer COVID-19. In addition to these two lung image phenotypes, the area and IMC1 of the computational GGO segments have a positive mediation effect on age caused COVID-19 severity, which suggests that elders often have bigger GGO lesions and more correlate distributed probability of the pixel values [41] within the GGO regions. Skewness of computational GGO segments also has a positive mediation effect on underlying diseases caused COVID-19 severity, which suggests that underlying diseases could cause asymmetrically distributed pixel values within GGO regions, thus leading to severer COVID-19. Interestingly, these lung image phenotypes have also been reported as potential image biomarkers for lung adenocarcinoma [67, non-small cell lung cancer [68], pulmonary interstitial pneumonia [69], and so on [70, 71].

One limitation of this study is that we do not have the resource to recruit radiologists into our experiment. We also do not have failure and success information of the PCR test and lung CT image test for the same patients. These could be future directions for institutes that are able to get these resources.

Conclusion

In conclusion, our work carefully considers several aspects of AI-based COVID-19 imaging diagnosis and prognosis, in terms of the model performance, model interpretability, and biological mechanism of the computational segmental image phenotypes associated with COVID-19. A series of sensitivity analyses have shown the robustness and generalizability of our proposed method. The clinical explanation of the computational GGO segment is also well addressed. Eight GGO segment-based image features have been identified as potential image biomarkers for COVID-19 severity. Comparing with previous works, our model shows better performance and is well interpreted both clinically and statistically.

Abbreviations

AI: Artificial intelligence; AUC: Area under the curve; CNN: Convolutional neural network; CT: Computerized tomography; GAN: Generative adversarial network; GGO: Ground-glass opacity; GLCM: Gray level co-occurrence matrix; GLDM: Gray level dependence matrix; ICTCF: Integrative resource of lung CT images and clinical features; InfNet: COVID-19 lung infection segmentation network; IoU: Intersection over union; Med-Seg: Medical segmentation; NGTDM: Neighboring gray tone difference matrix; PCR: Polymerase chain reaction; ROC: Receiver operating characteristic; SE: Standard error; SEM: Structure equation modeling; SGD: Stochastic gradient descent; SInfNet: Supervised COVID-19 lung infection segmentation network; SSInfNet: Self-supervised COVID-19 lung infection segmentation network.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-021-02992-2.

Additional file 1: **Figure S1**. Architecture of the supervised InfNet. **Figure S2**. A is the original architecture of the SInfNet. B is the architecture of our self-supervised InfNet model. Highlighted purple block is the difference between the original single SInfNet and the single SSInfNet. **Figure S3**. A is the architecture of the original multi SInfNet model. B is the architecture of our self-supervised multi InfNet model. Highlighted green block is the difference between the original multi SInfNet and our self-supervised multi SSInfNet. **Figure S4**. ROC for single InfNet. **Algorithm S1**. SSInfNet. **Table S1**. Image phenotypes. **Table S2**. The three-fold cross-validation performance of single networks. It should be noted that the data were obtained by combining the training, testing, and validation set from the Med-Seg (medical segmentation) COVID-19 dataset, and then splitting the combined data into 3 folds. **Table S3**. The three-fold cross validation performance of multi networks. **Table S4**. Comparison with transfer learning based FCN8 network. Quantitative result of Ground-glass Opacities & Consolidation on the test data set of the Med-Seg (medical segmentation) COVID-19 dataset. Prior was obtained from the single segmentation InfNet. **Table S5**. Model performance on independent COVID-19 CT Dataset set 2. **Table S6**. Model performance on the independent COVID-19 CT Data set 3. **Table S7**. Results of ablation studies. The performance of the ablation of our proposed multi-SSInfNet. Multi-SSInfNet refers to the self-supervised SInfNet with Focal Loss and Lookahead optimizer. We tried a variety of the model with a subtraction of the different technologies to carry out the ablation. **Table S8**. Computational costs of processing one image

Acknowledgements

We thank for the organizers who made the Meg-Seg data at http://medicalsegmentation.com/covid19/ and ICTCF data at http://ictcf.biocuckoo.cn/ online

Authors' contributions

PH, CL, DF, and QL designed the project. DF, QL and JZ developed the model and analyzed the data. PH and CL supervised the study. DF, QL were major contributors in writing the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported in part by Natural Sciences and Engineering Research Council of Canada (NSERC) and University of Manitoba. PH is the holder of Manitoba Medical Services Foundation (MMSF) Allen Rouse Basic Science Career Development Research Award.

Availability of data and materials

Code for SSInfNet is available at https://github.com/darylfung96/self-supervised-CT-segmentation.

Declarations

Ethics approval and consent to participate

We analyzed the publicly available data sets. The ethics approval and the consent to participate were done in original papers.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada. [2] Department of Biochemistry and Medical Genetics, University of Manitoba, 745 Bannatyne Avenue, Winnipeg, MB R3E 0J9, Canada. [3] CancerCare Manitoba Research Institute, CancerCare Manitoba, Winnipeg, MB R3E 0W3, Canada.

Received: 5 March 2021 Accepted: 17 July 2021
Published online: 26 July 2021

References

1. Disease outbreak news. WHO|Novel coronavirus—China. WHO. 2020 [cited 2020 Sep 22]. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/
2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20:533–4.
3. Aleta A, Martín-Corral D, y Piontti AP, Ajelli M, Litvinova M, Chinazzi M, et al. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. Nat Hum Behav. 2020;4:964–71. https://doi.org/10.1038/s41562-020-0931-9.
4. Harmon SA, Sanford TH, Xu S, Turkbey EB, Roth H, Xu Z, et al. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat Commun. 2020;11:1–7. https://doi.org/10.1038/s41467-020-17971-2.

5. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, et al. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. 2020. http://arxiv.org/abs/2003.05037
6. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ. 2020;369:18.
7. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell. 2020;2:283–8. https://doi.org/10.1038/s42256-020-0180-7.
8. Nikpouraghdam M, Jalali Farahani A, Alishiri GH, Heydari S, Ebrahimnia M, Samadinia H, et al. Epidemiological characteristics of coronavirus disease 2019 (COVID-19) patients in IRAN: a single center study. J Clin Virol. 2020;127:104378.
9. Banerjee A, Pasea L, Harris S, Gonzalez-Izquierdo A, Torralbo A, Shallcross L, et al. Estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study. Lancet. 2020;395:1715–25.
10. Remy-Jardin M, Tillie-Leblond I, Szapiro D, Ghaye B, Cotte L, Mastora I, et al. CT angiography of pulmonary embolism in patients with underlying respiratory disease: impact of multislice CT on image quality and negative predictive value. Eur Radiol. 2002;12:1971–8. https://doi.org/10.1007/s00330-002-1485-0.
11. Harris A, Kamishima T, Hao HY, Kato F, Omatsu T, Onodera Y, et al. Splenic volume measurements on computed tomography utilizing automatically contouring software and its relationship with age, gender, and anthropometric parameters. Eur J Radiol Elsevier. 2010;75:e97-101.
12. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. Annu Rev Psychol. 2007;58:593–614.
13. Chen, X., Yao, L., Zhou, T., Dong, J. & Zhang, Y. Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images. Pattern Recogn 2021;113:107826. https://doi.org/10.1016/j.patcog.2021.107826.
14. Li Y, et al. Efficient and Effective Training of COVID-19 Classification Networks With Self-Supervised Dual-Track Learning to Rank. IEEE J Biomed Health Info 2020; 24(10):2787-2797. https://doi.org/10.1109/JBHI.2020.3018181
15. Shahinfar S, Meek P, Falzon G. "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. Ecol Inform. 2020;57:101085.
16. Voulodimos A, Protopapadakis E, Katsamenis I, Doulamis A, Doulamis N. A few-shot U-net deep learning model for COVID-19 infected area segmentation in CT images. Sensors. 2021;21(6):2215.
17. Voulodimos A, Protopapadakis E, Katsamenis I, Doulamis A, Doulamis N. Deep learning models for COVID-19 infected area segmentation in CT images. medRxiv. 2020. https://doi.org/10.1101/2020.05.08.20094664.
18. 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. 2021;48:1197–210.
19. Lizancos Vidal P, de Moura J, Novo J, Ortega M. Multi-stage transfer learning for lung segmentation using portable X-ray devices for patients with COVID-19. Expert Syst Appl. 2021;173:114677.
20. Aslan MF, Unlersen MF, Sabanci K, Durdu A. CNN-based transfer learning–BiLSTM network: a novel approach for COVID-19 infection detection. Appl Soft Comput. 2021;98:106912.
21. Katsamenis I, Protopapadakis E, Voulodimos A, Doulamis A, Doulamis N. Transfer learning for COVID-19 pneumonia detection and classification in chest X-ray images. In: Proceedings of the PCI. 2020. p. 170–174. https://doi.org/10.1145/3437120.3437300.
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 2020;39:2626–37. https://doi.org/10.1109/TMI.2020.2996645.
23. Yao Q, Xiao L, Liu P, Kevin Zhou S. Label-free segmentation of COVID-19 lesions in lung CT. IEEE Trans Med Imaging 2021. https://doi.org/10.1109/TMI.2021.3066161.
24. Wang Y, Zhang J, Kan M, Shan S, Chen X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2020. pp. 12272–12281.
25. Bertalmio M, Sapiro G, Caselles V, Ballester C. Image inpainting. In: Proceedings of the SIGGRAPH. 2000. pp. 417–424. https://doi.org/10.1145/344779.344972.
26. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. Context encoders: feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 2536–2544.
27. Liu G, Reda FA, Shih KJ, Wang T-C, Tao A, Catanzaro B. Image inpainting for irregular holes using partial convolutions. Proc. ECCV 2018;89–105. https://doi.org/10.1007/978-3-030-01252-6_6.
28. Xie J, Xu L, Chen E. Image denoising and inpainting with deep neural networks. In: NIPS. 2021. pp. 350–358. https://proceedings.neurips.cc/paper/2012/file/6cdd60ea0045eb7a6ec44c54d29ed402-Paper.pdf.
29. Hassan ET, Abbas HM, Mohamed HK. Image inpainting based on image segmentation and segment classification. In: Proceedings of ICCSCE 2013. 2013. p. 28–33. https://doi.org/10.1109/ICCSCE.2013.6719927
30. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS. Generative image inpainting with contextual attention. openaccess.thecvf.com. 2010;4:34–40. https://github.com/
31. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. 2014;1–9. http://arxiv.org/abs/1406.2661
32. Lavanya M, Kannan PM. Lung lesion detection in CT scan images using the Fuzzy Local Information Cluster Means (FLICM) automatic segmentation algorithm and back propagation network classification. Asian Pacific J Cancer Prev. 2017;18:3395–9.
33. Collins J, Stern EJ. Chest radiology: the essentials. Lippincott Williams & Wilkins; 2008.
34. Dahnert WF. Radiology review manual, 8e. Lippincott Williams & Wilkins; 2017. ISBN 9781496360694.
35. Robin Smithuis O van D and CS-P. The radiology assistant : basic interpretation. https://radiologyassistant.nl/chest/hrct/basic-interpretation
36. COVID-19—Medical segmentation. http://medicalsegmentation.com/covid19/
37. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell 2017;42:318–27. https://doi.org/10.1109/TPAMI.2018.2858826
38. Bottou L, Curtis FE, Nocedal J. Optimization methods for large-scale machine learning. SIAM Rev. 2018;60(2):223–311.
39. Zhang MR, Lucas J, Hinton G, Ba J. Lookahead optimizer: k Steps forward, 1 step back. Adv Neural Inf Process Syst. 2019;32:1–19.
40. Ning W, Lei S, Yang J, Cao Y et al.. Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. Nat Biomed Eng 2020;4:1197–1207. https://doi.org/10.1038/s41551-020-00633-5
41. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77:e104–7.
42. Wei T, Simko V. R package "corrplot": visualization of a correlation matrix. 2017.
43. Rosseel Y. Lavaan: An R package for structural equation modeling. J Stat Softw 2012;48:1–36. https://doi.org/10.18637/jss.v048.i02.
44. Gunzler D, Chen T, Wu P, Zhang H. Introduction to mediation analysis with structural equation modeling. Shanghai Arch Psychiatry. 2013;25:390–4.
45. Muthen B. Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus. Reliab Risk Saf Back to Futur. 2010;106–13.
46. Hooda R, Mittal A, Sofat S. Lung segmentation in chest radiographs using fully convolutional networks. Turkish J Electr Eng Comput Sci 2019;27:710–22. https://doi.org/10.3906/elk-1710-157.
47. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell Cell Press. 2020;181:1423-1433.e11.
48. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI 2015. 2015. pp. 1–8. https://doi.org/10.1007/978-3-319-24574-4_28.
49. Zeng Z, Xie W, Zhang Y, Access YL-I, 2019 U. RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images. IEEE Access 2019;7:21420–8. https://doi.org/10.1109/ACCESS.2019.2896920.
50. Wang C, Macgillivray T, Macnaught G, Yang G, Newby D. A two-stage 3D Unet framework for multi-class segmentation on full resolution image.

In: Proceedings of STACOM 2018. 2018. pp. 191–199. https://doi.org/10.1007/978-3-030-12029-0_21.
51. Weng Y, Zhou T, Li Y, Access XQ-I, 2019 U. Nas-unet: Neural architecture search for medical image segmentation. IEEE Access 2019;7:44247–57. https://doi.org/10.1109/ACCESS.2019.2908991.
52. Li X, Chen H, Qi X, Dou Q, Fu C-W, Heng P-A. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans Med Imaging 2018;37:2663–74. https://doi.org/10.1109/TMI.2018.2845918.
53. Chen W, Liu B, Peng S, Sun J, Qiao X. S3D-UNET: Separable 3D U-Net for brain tumor segmentation. In: Proceedings of BrainLes 2018. 2018. pp. 358–368. https://doi.org/10.1007/978-3-030-11726-9_32.
54. Dudewicz EJ, van der Meulen EC. Entropy-based tests of uniformity. J Am Stat Assoc. 1981;76:967.
55. van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Lambin P. Feature selection methodology for longitudinal cone-beam CT radiomics. Acta Oncol (Madr). 2017;56:1537–43. https://doi.org/10.1080/0284186X.2017.1350285.
56. Lee SH, Cho HH, Lee HY, Park H. Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer. Cancer Imaging. 2019;19:54. https://doi.org/10.1186/s40644-019-0239-z.
57. 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. 2020;296:E32-40. https://doi.org/10.1148/radiol.2020200642.
58. Dai WC, Zhang HW, Yu J, Xu HJ, Chen H, Luo SP, et al. CT Imaging and differential diagnosis of COVID-19. Can Assoc Radiol J. 2020;71:195–200.
59. Shen C, Mark R, Kagetsu NJ, Becker AS, Bar-Yam Y. Combining PCR and CT testing for COVID. 2020. http://arxiv.org/abs/2006.02140
60. Simpson S, Kay FU, Abbara S, Bhalla S, Chung JH, Chung M, et al. Radiological society of north America expert consensus statement on reporting chest CT findings related to COVID-19. Endorsed by the society of thoracic Radiology, the American College of Radiology, and RSNA. Radiol Cardiothorac Imaging. 2020;2:e200152. https://doi.org/10.1148/ryct.2020200152.
61. Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. Radiology. 2020;296:E46-54.
62. Wen Z, Chi Y, Zhang L, Liu H, Du K, Li Z, et al. Coronavirus disease 2019: initial detection on chest CT in a retrospective multicenter study of 103 Chinese subjects. Radiol Cardiothorac Imaging. 2020;2:e200092. https://doi.org/10.1148/ryct.2020200092.
63. 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. 2020;296(2):E115–7. https://doi.org/10.1148/radiol.2020200432.
64. Inui S, Fujikawa A, Jitsu M, Kunishima N, Watanabe S, Suzuki Y, et al. Chest CT findings in cases from the cruise ship "Diamond Princess" with coronavirus disease 2019 (COVID-19). Radiol Cardiothorac Imaging. 2020;2:e200110. https://doi.org/10.1148/ryct.2020200110.
65. Singh N, Fratesi J. Chest CT imaging of an early Canadian case of COVID-19 in a 28-year-old man. CMAJ 2020;192:E455. https://doi.org/10.1503/cmaj.200431
66. Rubin GD. Costing in radiology and health care: rationale, relativity, rudiments, and realities. Radiology. 2017;282:333–47. https://doi.org/10.1148/radiol.2016160749.
67. Yuan M, Pu X-H, Xu X-Q, Zhang Y-D, Zhong Y, Li H, et al. Lung adenocarcinoma: assessment of epidermal growth factor receptor mutation status based on extended models of diffusion-weighted image. J Magn Reson Imaging. 2017;46:281–9. https://doi.org/10.1002/jmri.25572.
68. Weiss GJ, Ganeshan B, Miles KA, Campbell DH, Cheung PY, Frank S, et al. Noninvasive image texture analysis differentiates K-ras mutation from pan-wildtype NSCLC and is prognostic. PLoS One 2014;9:e100244. https://doi.org/10.1371/journal.pone.0100244.
69. Koyama H, Ohno Y, Yamazaki Y, Nogami M, Kusaka A, Murase K, et al. Quantitatively assessed CT imaging measures of pulmonary interstitial pneumonia: effects of reconstruction algorithms on histogram parameters. Eur J Radiol. 2010;74:142–6.
70. Schofield R, Ganeshan B, Fontana M, Nasis A, Castelletti S, Rosmini S, et al. Texture analysis of cardiovascular magnetic resonance cine images differentiates aetiologies of left ventricular hypertrophy. Clin Radiol. 2019;74:140–9.
71. Baek HJ, Kim HS, Kim N, Choi YJ, Kim YJ. Percent change of perfusion skewness and kurtosis: a potential imaging biomarker for early treatment response in patients with newly diagnosed glioblastomas. Radiology 2012;264:834–43. https://doi.org/10.1148/radiol.12112120.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.