Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 145 (2022) 105464

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Generalizability assessment of COVID-19 3D CT data for deep learning-based disease detection

Maryam Fallahpoor [a,b], Subrata Chakraborty [a,c,*], Mohammad Tavakoli Heshejin [d], Hossein Chegeni [e], Michael James Horry [a], Biswajeet Pradhan [a,f,g]

[a] Center for Advanced Modelling and Geospatial Information Systems, Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW, 2007, Australia
[b] Department of Nuclear Medicine, Vali-Asr Hospital, Tehran University of Medical Sciences, Tehran, Iran
[c] School of Science and Technology, Faculty of Science, Agriculture, Business and Law, University of New England, Armidale, NSW, 2351, Australia
[d] Department of Information Technology, Telecommunication Company of Iran, Tehran, Iran
[e] Imaging Center, Iranmehr Hospital, Tehran, Iran
[f] Center of Excellence for Climate Change Research, King Abdulaziz University, P. O. Box 80234, Jeddah, 21589, Saudi Arabia
[g] Earth Observation Center, Institute of Climate Change, Universiti Kebangsaan Malaysia, 43600, UKM, Bangi, Selangor, Malaysia

ARTICLE INFO

Keywords:
3D convolutional neural network
3D CT scan
Deep learning
COVID-19
Generalizability
Lung involvement detection

ABSTRACT

Background: Artificial intelligence technologies in classification/detection of COVID-19 positive cases suffer from generalizability. Moreover, accessing and preparing another large dataset is not always feasible and time-consuming. Several studies have combined smaller COVID-19 CT datasets into "supersets" to maximize the number of training samples. This study aims to assess generalizability by splitting datasets into different portions based on 3D CT images using deep learning.

Method: Two large datasets, including 1110 3D CT images, were split into five segments of 20% each. Each dataset's first 20% segment was separated as a holdout test set. 3D-CNN training was performed with the remaining 80% from each dataset. Two small external datasets were also used to independently evaluate the trained models.

Results: The total combination of 80% of each dataset has an accuracy of 91% on Iranmehr and 83% on Moscow holdout test datasets. Results indicated that 80% of the primary datasets are adequate for fully training a model. The additional fine-tuning using 40% of a secondary dataset helps the model generalize to a third, unseen dataset. The highest accuracy achieved through transfer learning was 85% on LDCT dataset and 83% on Iranmehr holdout test sets when retrained on 80% of Iranmehr dataset.

Conclusion: While the total combination of both datasets produced the best results, different combinations and transfer learning still produced generalizable results. Adopting the proposed methodology may help to obtain satisfactory results in the case of limited external datasets.

1. Introduction

The ongoing global COVID-19 pandemic presents governments and healthcare clinics with immense financial and human resources challenges [1] due to an increased demand for medical professionals, many of whom have succumbed to the disease. Clinical resources are tailored for non-pandemic operations, and long-term maintenance of extra staff in service for pandemic situations is not economically feasible. Consequently, global medical systems have been overwhelmed since the onset of the current COVID-19 pandemic [2]. This lack of resources has led to repeated failings to meet the diagnostic and therapeutic needs of the public. The effectiveness of global vaccination efforts [3] is reduced by new variants of COVID-19 [4,5]. Ultimately, controlling the spread of the disease is a long-term goal, and clinics will remain under pressure for the foreseeable future.

Clinical studies show that approximately 2–8% of patients infected with COVID-19 will develop severe pneumonia [6,7], being the primary cause of COVID-19 related death [8]. Although reverse transcription-polymerase chain reaction (RT-PCR) is considered as the gold standard test for COVID-19 detection [9], challenges include high

* Corresponding author. School of Science and Technology, Faculty of Science, Agriculture, Business and Law, University of New England, Armidale, NSW, 2351, Australia.
E-mail address: subrata.chakraborty@une.edu.au (S. Chakraborty).

https://doi.org/10.1016/j.compbiomed.2022.105464
Received 4 October 2021; Received in revised form 25 March 2022; Accepted 25 March 2022
Available online 1 April 2022

false-negative rates, time-consuming processing, and low sensitivity (60–71%) hinder the use of this technique [10]. X-ray imaging has also been used as a common, fast, and cost-effective imaging tool [11] for COVID-19 detection. However, X-ray images are imprecise, and radiologist interpretations of X-ray images are error-prone [12,13]. In contrast to standard chest X-ray images, the sensitivity of the chest Computed Tomography (CT) image for COVID-19 pulmonary infection is above 94% [10,14–16], with radiological features showing as small patches and ground-glass opacities associated with COVID-19 [17]. Therefore, it has been proposed that chest CT could potentially be employed as the primary diagnostic modality for examining patients with COVID-19 [18].

AI-based computer vision techniques may be used to automate the screening and classification of large volumes of COVID-19 patient images. For instance, such systems could be used to measure a patient's COVID-19 pulmonary involvement score as an objective measure of infection severity or to detect other clinically important COVID-19 associated lung diseases [19]. If the accuracy/sensitivity of the AI is sufficient, then the number of human radiologist-based CT interpretations will be reduced, resulting in less pressure on those radiologists [20]. Of the numerous computer vision systems to have been developed, the deep learning (DL) approach is promising since it does not require handcrafted feature extraction [21–23]. Manual feature extraction is a process that hinders the performance of deep learning techniques in new situations and complicated data sets. Several studies have previously shown the potential for DL classification in screening chest CT scans to diagnose patients with pulmonary COVID-19 involvement [16,24–26].

Three-dimensional imaging data improve the radiological assessment of lung diseases since the spatial relationship between CT slices provides valuable information relating to the extent of lung involvement and, therefore, more accurate disease diagnosis. Particularly in the COVID-19 diagnosis, 3D CT image superiority over chest X-ray, mainly due to the higher sensitivity [27] and resolution [28]. Despite the advantages of 3D CT images, there are relatively few studies on the application of 3D convolutional neural network (CNN) for COVID-19 classification [29–32] compared to a huge number of 2D CT studies [33–36]. This is mainly due to more process-intensive computing requirements [37] and the time-consuming process of 3D model training, which is in the order of days or weeks.

A key challenge in relation to deep learning algorithms for COVID-19 classification is the developed model's generalization capability with respect to external datasets [38]. Although AI model generalization is critical to clinical adoption [39,40], there are few published studies that perform validation of COVID-19 classification metrics against external datasets. One reason for this is the relative scarcity of large, high-quality labeled datasets needed for training deep learning models.

Several studies have combined smaller COVID-19 CT datasets into "supersets" to maximize the number of training samples for deep learning models. Our main contribution in this work is to thoroughly assess generalizability using varied combinations of two large COVID-19 CT image datasets using state-of-the-art 3D convolutional neural networks (CNN), showing that a combination of datasets can assist generalization. We further determine the optimal " combination" characteristics of these datasets.

2. Related work

Upon the global outbreak of the recent COVID-19 pandemic, the need for computer-aided diagnosis methods has significantly increased [19,20,41,42]. Most studies conducted on automated COVID-19 diagnosis from CT images using a single, internal dataset for training, validation, and testing deep learning models, resulting in high classification metrics [29,43]. It is not possible to assess whether these results are driven by classifier sensitivity to disease pathology or bias introduced by class imbalance, patient selection, or confounding bias. This is particularly a concern where disease-positive and negative disease patients have been sourced independently, potentially introducing systematic differences in image classes related to CT acquisition apparatus, operational parameters, and regional patient morphological differences. Such biases have been found to result in considerably lower classification metrics when these models are tested against external datasets [44,45].

A small number of studies focused on investigating the generalization of AI-based COVID-19 diagnosis [24,41,46,47]. Harmon et al. [24] combined four datasets into combinations of training, validation, and testing image corpora by excluding one dataset consisting of 147 patients as a holdout test set. They used DenseNet-121 as 3D CNN and implemented both lung segmented and full 3D image classification, considering one complete volume at a fixed size. They achieved 90.8% accuracy, 84% sensitivity, and 93% specificity. In this study, a total combination of datasets was performed, and the results were tested on a comparably small dataset. The authors considered one fixed dataset as a test dataset and there is a lack of external validation on each four datasets to demonstrate their network capability to achieve similar results. In the current study, we seek the results of one trained dataset when tested on the other datasets and the effect of data augmentation on generalization.

In separate work, Nguyen et al. [46] used four different datasets, including one internal dataset at UT Southwestern (UTSW) (337 patients) and three external datasets: 1) China Consortium of Chest CT Image Investigation (CC–CCII) [30], 2) COVID-CT set [48] and 3) MosMedData [49]. They implemented nine combinations of these datasets for two classes of COVID-19 positive and COVID-19 negative cases. They both trained the different combinations and tested on an external test dataset and trained the different combinations and tested on a holdout test set from one of the datasets used for training. They used different models for training on 3D CT images from which the best results were for the models trained on multiple datasets and evaluated on a test set from one of the datasets used for training (accuracy of 86–97%). Despite these promising internal classification results, classification metrics for these models were reduced to pure chance when evaluated against an external dataset, with an AUC of 0.5 calculated for all models. Nguyen et al. [46] adjusted the disease positive probability threshold to maximize accuracy in their simulations, thereby tightly binding model performance to the test dataset. This study did not segment the lung field from the CT images to reduce signal noise from features including ribs/bone and surrounding areas. In the present study, we have used 0.5 as the disease positive probability threshold for all models to decouple results from datasets, and lung segmentation was performed in the preprocessing part of the current study.

More recently, two comprehensive studies addressed generalization aspect of COVID-19 classification task. Li et al. [50], proposed the contrastive multi-task convolutional neural network (CMT-CNN) as a multi-task framework to increase generalizability. The authors stated that there is no need for further annotation to improve generalization using CMT-CNN. They used 3D volumes of CT images from two datasets: one from CC-CCII2 [30] with 4356 CT images and one from their hospital consisting of 402 CT images from 108 COVID-19 diagnosed patients confirmed by RT-PCR test. For X-ray, they used three datasets, including two public datasets from Cohen et al. [51] and Kaggle [52] and one from their hospital-based dataset with 231 COVID-19 cases in total. They used Mendeley Data website [53], containing 4007 pneumonia and 1583 normal cases as their normal control instances. Certain augmentation methods, including distortion, painting, and perspective transformations, improved representational learning capability. The results of their study indicate 5.49–6.45% generalization accuracy improvement for CT and 0.96–2.42% for X-ray images.

In another study, Aversano et al. [54] combined three pre-trained deep neural networks, including VGG-19 [55], Xception [56], and ResNet-50 [57], evolved with a direct coding scheme based on genetic programming to develop an ensemble classifier for each lung lobe

(superior, middle, and inferior). The main parts of their proposed ensemble architecture are multiple deep neural networks based on pre-trained models and a voting strategy. For the training phase, they used two volumetric CT datasets, Extensive COVID-19 X-Ray and CT Chest Images Dataset [58] and Coronavirus (COVID-19) CC-19 dataset [59], then clustered them into three sub-datasets comprising images of each lung lobe. To evaluate the results on external data, they used SARS-COV-2 Ct-Scan Dataset [60]. The pre-trained transfer learning CNN models combined with VGG-19, ResNet-50, and Xception were re-trained for the binary classification of CT images of COVID-19 versus normal cases. The genetic algorithm in this study executes an evolutionary process to identify the best architecture adaptation of the pre-trained models. The evaluation results on the external test dataset showed F1 score of 0.903 while it was 0.94–0.95 for their integrated dataset. In Refs. [50,54] studies, whole datasets were considered training and test datasets to assess generalizability. There is a lack of true external validation for each dataset (i.e., considering each dataset as an external test dataset in different simulations), and the applicability of trained models to real-life clinical situations is unknown.

A few previous studies have assessed the generalizability of CNN models trained on 2D CT slices and X-ray images. In a study by Silva et al. [40], that was performed on 2D CT slices, EfficientCovidNet, was proposed along with a voting-based approach and a cross-dataset analysis for COVID-19 detection. They evaluated EfficientCovidNet on three setups and with the two largest public CT datasets, including a cross-dataset analysis. The results of this study indicated the accuracy drops from 87.68 to 56.16% for the external COVID-19 test set.

Ahmed et al. [39] demonstrated a significant gap between the model tested on before-seen data (same source) and the model tested on external data in COVID-19 detection from X-ray images. Their developed model reached the AUC of 1.00 when tested on seen data while it was only 0.38 on external data. Hence, they recommended further investigations into finding/focusing on features that can be generalized across datasets.

Bassi et al. [44] tested the effect of segmentation on X-ray image classification of COVID-19, normal, and pneumonia cases. It was shown that segmenting lung has a positive effect on the model generalization capability, increasing the mean accuracy score on the external test dataset by 4.7% and the Bayesian estimation means by 4.4%. The results when tested on the external dataset, showed 85% sensitivity for COVID-19 detection in the case of the segmented lung being used while it was 81% for non-segmented lung. They stated that the improvement in accuracy might be due to the attention of DNN to the lung region. Lung segmentation can also reduce dataset bias and improve generalization.

The key focus of the current study is to assess the generalization of computer vision models trained on 3D CT images for automated COVID-19 diagnosis. According to the literature above, it can be seen that most available studies have selected one fixed dataset as their external test dataset. Therefore, there is a need for the study to test external validation on each dataset involved in the study since the results may vary significantly on different external test sets. Hence, we dedicated a part of this study to investigating this issue.

Another research gap identified in the above studies is that all available datasets were combined together for training and testing. Although a large number of data leads to more accurate results, the results from combinations of different data portions have not been investigated yet. We addressed the analysis of the results acquired from different combinations of dataset portions in both fully trained and transfer learning approaches. In the case of satisfactory results, the need for large dataset combinations is alleviated.

Finally, we have found that lung field segmentation plays a pivotal role in promoting model generalizations and recommend that this procedure be a standard part of the 3D CT image preprocessing pipeline for CNN-based COVID-19 diagnosis from medical images.

3. Materials and methods

The flowchart in Fig. 1 summarizes the procedures implemented in this study, and each step is described below.

3.1. Patients and dataset

Our study employs four independently sourced datasets. The first dataset was collected from Iranmehr hospital, located in Tehran, Iran, and we name this dataset as “Iranmehr”. Digital Imaging and Communications in Medicine (DICOM) data of chest CT images of 1110 patients were collected from Iranmehr hospital picture archiving and communication system (PACS). This dataset was collected from February 2020 to March 2020, when COVID-19 was at its peak. Imaging was done on GE Medical Brightspeed 16 detector multislice CT scan machine; low dose spiral high-resolution CT imaging technique was employed. CT images were collected as a screening protocol before hospitalization of the patients for COVID-19 infection detection. Pulmonary COVID-19 involvement score was based on the interpretation of two expert independent radiologists who had access to clinical data of the patients. Radiology specialists validated the gathered data, so only normal and COVID-19 patients were included. Iranmehr Hospital specialists supervised the collection of all patient data. Data was collected under the policies of Iranmehr hospital, which allow anonymized data to be used for research purposes. The data collection, subsequent anonymization (done onsite under strict supervision), and usage for this study were undertaken with proper authorization and following international data privacy standards. The second dataset was sourced from hospitals located in Moscow and made available by Morozov et al. [49]. This dataset has been assessed and labeled by expert radiologists according to COVID-19 lung involvement and grouped into four classes at 25% intervals. The first class, named CT-0 contains 254 images with no lung involvement representing a normal CT image. Classes CT-1 to CT-4 represent 25%–100% lung involvement and contain 854 images. This dataset is referred to as “Moscow” in this paper. We used two additional external test datasets to assess the validity of our results. First, the low-dose and ultra-low-dose (LDCT) [61] containing CT images of 104 COVID-19 positive cases, and 56 normal cases, were collected in Babak Imaging Center, Tehran, Iran. The second dataset is the 3DLSC-COVID dataset [62] which is publicly available and contains 100 COVID-19 positive cases and 96 normals. The LDCT image format is DICOM and 3DLSC is NIFTI.

3.2. Preprocessing

The matrix size of all CT images was 512 × 512 pixels, but they had different slice numbers. So, after loading DICOM CT images, they were initially resampled and interpolated to have the same slice number. We prepared two forms of datasets, including cropped and non-cropped images, to assess the effect of cropping in the training phase. For cropping sets, all images were cropped to remove surrounding areas that are not significant and then resampled to have the same size as 128 × 128 × 60. All CT images were resampled to a resolution of 1 mm × 1 mm × 1 mm and intensity clipped to (−1000, 400) Hounsfield Unit (HU) range which is considered as HU window for lung. We are not interested in HU values above 400, which are bony structures. The values below −1000

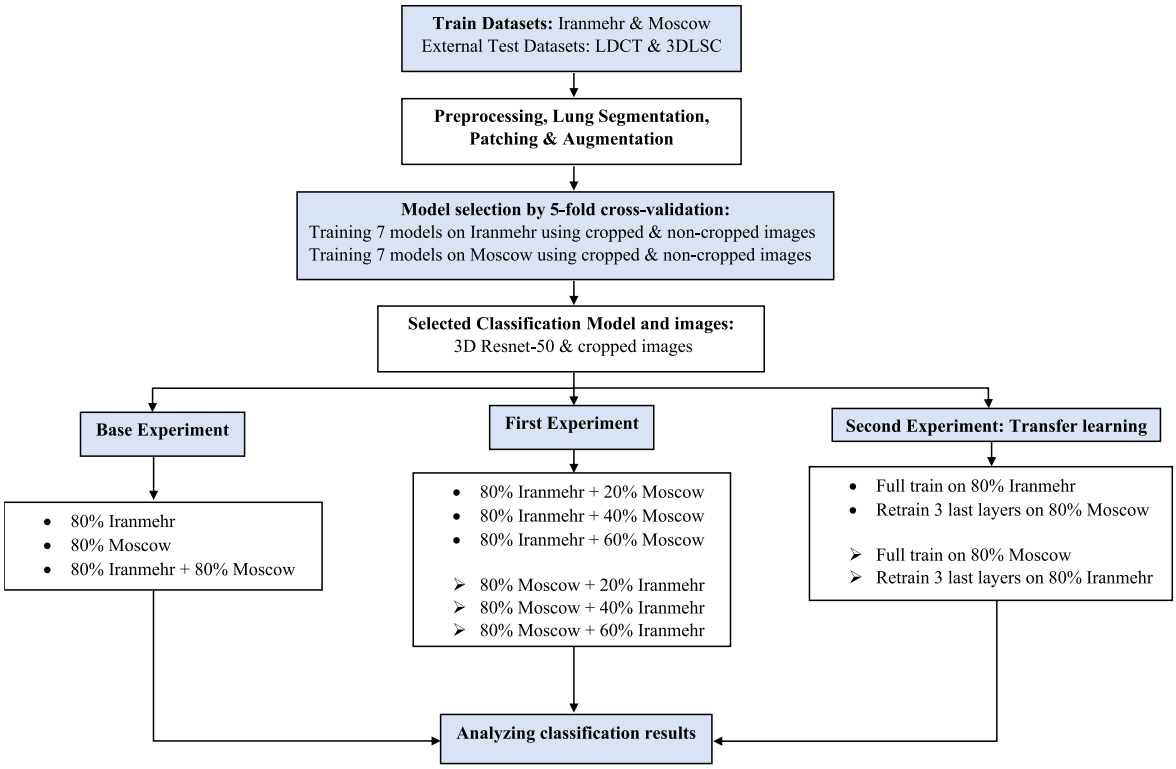


Fig. 1. The experimental workflow implemented in this study.

are also out of the range of the lung's HU. For non-cropping sets, we applied the same approach for resampling and kept the whole field of view (FOV) of the image. We assigned 1 for positive pulmonary COVID-19 involvement and 0 for normal images. We saved the preprocessed images as NumPy arrays to be fed as a network's input. The input of the 3D networks must have the same slice number. So, for 3D CNNs, resampling is of great importance. Additionally, cropping and intensity clipping remove the less useful parts of the image, resulting in more efficient training.

3.2.1. Segmentation

The segmentation results showed that the lung field from the CT image improved the results of classification and generalizability. Our simulations demonstrated that models trained with segmented lung CT images had results approximately 5% better than with non-segmented lung CT images. Fig. 2(a) and (b) illustrate the feature map output for segmented and non-segmented lungs, respectively. It can be seen from the figures that useless areas such as ribs, spine, and surrounding tissues exist in non-segmented lungs and affect the classification results. We tested three different algorithms on our four datasets to assess whether we could have one single lung segmentation approach. The segmentation methods include DSB Lung Segmentation Algorithm from Kaggle [63], an algorithm developed by Zuidhof [64], and a U-net based lung segmentation developed by Hofmanninger [65]. Nevertheless, as can be seen in Fig. 3, for each data format, one type of segmentation method performs better. This is probably due to the Neuroimaging Informatics Technology Initiative (NIFTI) format of images compared to the DICOM format. The reason for this might be the loss of some information during the conversion of original DICOM images to NIFTI format. The DSB algorithm failed to segment peripheral parts of the lung which have COVID-19 involvement. Therefore, we applied a U-net based lung segmentation module on Iranmehr and LDCT datasets to have 3D segmented lung area. For segmentation of Moscow and 3DLSC datasets, we used the Zuidhof method. Fig. 4 shows the result of the segmented lung used in the present study. On the other hand, the Zuidhof is not accurate as the Hofmanninger approach, and there were 17 out of 1110 images from the Moscow dataset that were not segmented properly. We used original CT images for these 17 non-segmented cases. Next, normalization, zero-centering, and shuffling were done in a pre-processing part of the task.

3.2.2. Patching and augmentation

Our trial-and-error experimentations found that simulations with data augmentation outperformed simulations without data by 2–5%. Random data augmentation prevents early overfitting and improves model performance. Furthermore, data augmentation produces different shapes and orientations of the images while still being recognizable, allowing the model to learn more features. Data augmentation steps were employed using random noise (mean: 0, standard deviation: 0.08), translation (shift with the size of random integer number between $(-0.1, 0.1) \times$ patch size in the x-direction), random rotation (random rotation between 0° and 360°), distort elastic (alpha: 100, sigma: 10), flip (in the direction of x and z-axis), 90-degree rotation (which provides random rotation of 90°, 180°, 270°), and scaling (zoom with the random size between 0.6 and 1.2). All the augmentations were applied in "on the fly" mode in the generator to prevent the network from overfitting.

We applied patching to train input images of a size that covers most of the lung and is a reasonable size for patches. Logically, since the test is performed on the full image of the patient, not just a patch, the patch size should be large enough to cover most of the lungs. Testing on the full image provides the most accurate results. When a patch is normal, it means the patient's decision was normal, however, there may be one

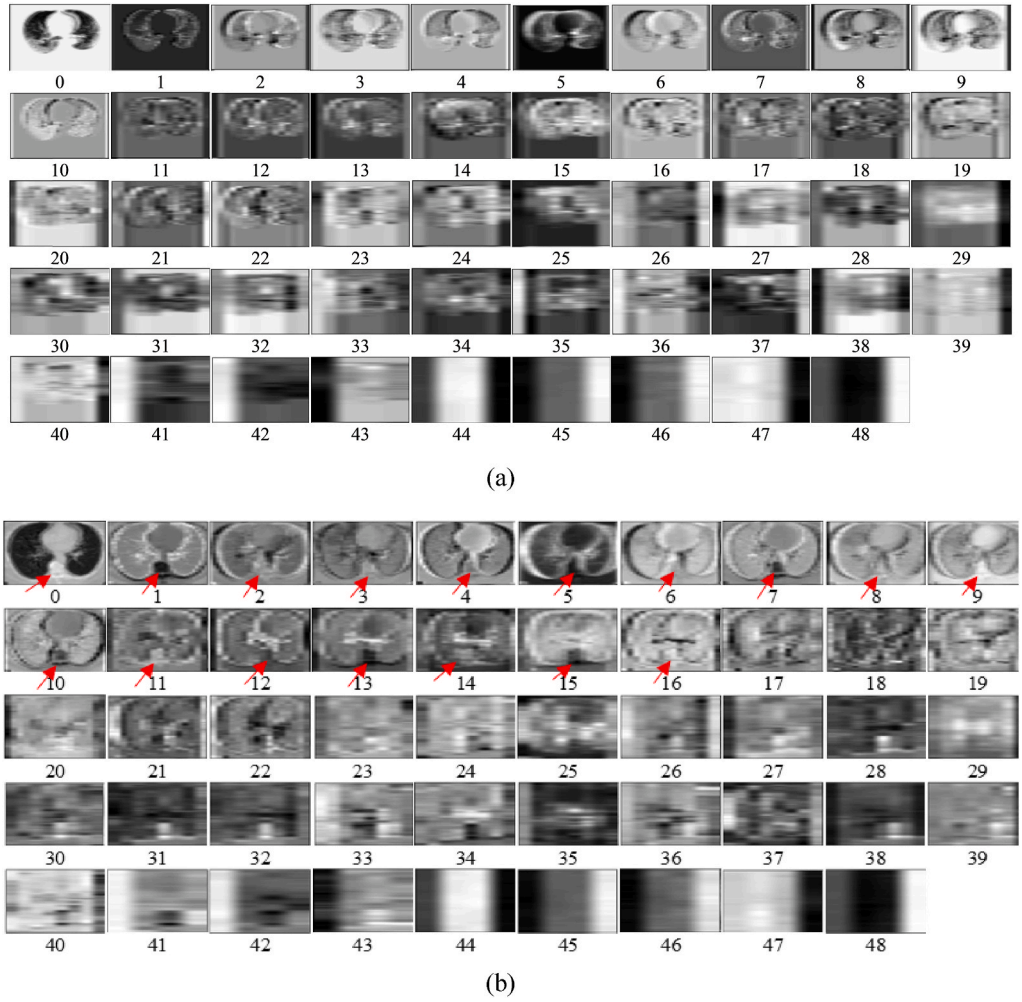


Fig. 2. (a) A typical segmented CT slice of feature map output from each convolutional layer of 3D ResNet-50 for a COVID-19 positive case; (b) A typical non-segmented CT slice of feature map output from each convolutional layer of 3D ResNet-50 for a COVID-19 positive case. Red arrows show the spine as an example surrounding area. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

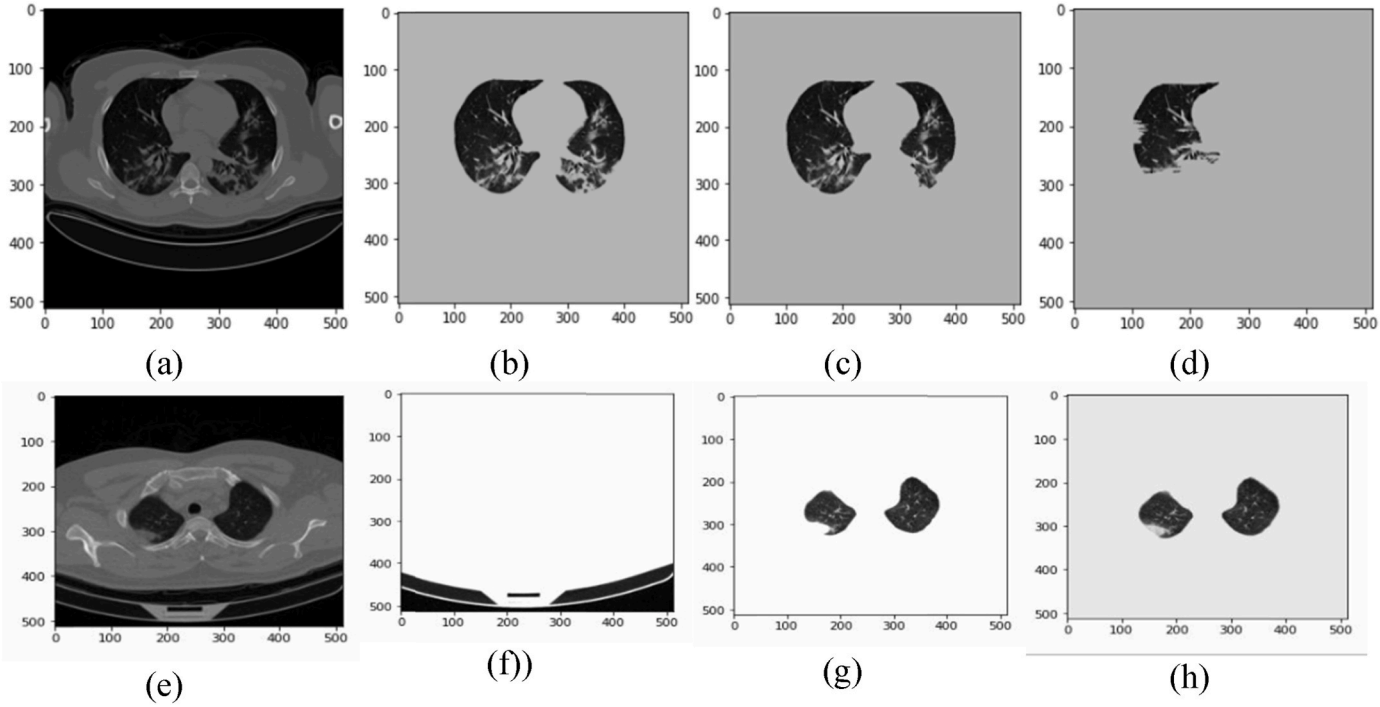


Fig. 3. Results of different segmentation methods on DICOM and NIFTI image format. (a) original NIFTI image segmented by (b) Zuidhof method, (c) DSB algorithm, and (d) Hofmanninger method. (e) Original DICOM image segmented by (f) Zuidhof method, (g) DSB algorithm, and (h) Hofmanninger method.

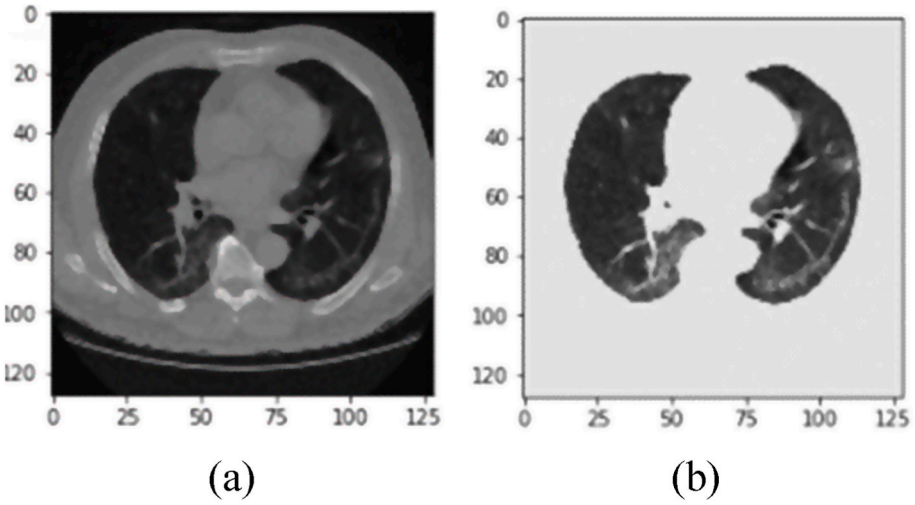


Fig. 4. A typical slice of a) chest CT image and b) segmented lung. This slice is for a COVID-19 positive case.

patch that is detected as COVID-19, so the full image is COVID-19. From different trials for hyper-parameters' testing, the patch-size of 115 × 115 × 55 was applied for images since it had up to 2% better performance than other patch sizes. For each dimension of x, y, and z, a random number was selected from the difference between the original size of the image and the patch size. The patch produced was from that random number to the patch size, plus that random number in each dimension. For big patch sizes, the overlap would be very high. However, in the generator part of our network, we first patched the data and then implemented augmentation on each patch to have augmented data as much as possible. Our trials found that this method improved the model performance compared to when patching was performed after data augmentation. The reason is that each patch undergoes a set of random augmentation and would be unique. The adopted approach for patching is shown in pseudocode below:

```
FUNCTION generate_train_data(x_inp, y_inp, patchsize, is_Moscow=False):
    WHILE True:
        selected_pat_x, selected_pat_y <- select_image_randomly()
        patch_number = 6 IF is_Moscow AND selected_pat_y ==1 ELSE 16 ENDIF
    For i = 1 to patch_number:
        row <- np.random.choice(range( IMAGE_X_SIZE-patchsize[0]))
        column <- np.random.choice(range( IMAGE_Y_SIZE-patchsize[1]))
        slice <- np.random.choice(range( IMAGE_Z_SIZE-patchsize[2]))
        xr <- selected_pat_x[row:row+patchsize[0],column:column+patchsize[1],
slice:slice+patchsize[2]]
            FOR each augmentation on the augmentation methods
                xr <- augment_patch_randomly(xr, augmentation)
        ENDFOR
            yield (xr,selected_pat_y)
        ENDFOR
    ENDWHILE
ENDFUNCTION
```

3.3. Classification

We used Tensorflow [66] library as the platform. Different l1 and l2 regularization at variable values were tested, and the l2 regularization at 0.001 was optimal and used. Furthermore, dropout at 0.2 was employed for further regularization and "Adam" [67] as model optimizer. To overcome the imbalanced number of COVID-19 and normal cases, we set a number of 6 patches per image for the majority class and 16 patches per image for the minority class. Also, we applied this approach to overcome the overfitting of the training. We saved the best weights of the trained model, so that overfitting didn't affect the simulations, even if we had overfitting in certain scenarios.

The network was trained on advanced GPUs provided by the UTS Interactive High-Performance Computing (iHPC).

3.4. Model selection

For all the experiments in the present work, we employed k-fold running. Using the k-fold technique, the dataset is randomly partitioned into k groups or folds of roughly equal size. In order to test the model performance, the first fold is kept, and the model is trained using k-folds. Validation is repeated k times, and each time a different fold or a different set of data points is used. Seven common models, including

Table 1
Accuracy (mean ± std) for 5-fold cross-validation on cropped and non-cropped images of Iranmehr and Moscow datasets.

	Iranmehr cropped	Iranmehr non-cropped	Moscow cropped	Moscow non-cropped
Densenet-169	**0.942 ± 0.012**	0.938 ± 0.004	0.857 ± 0.027	0.843 ± 0.019
ResNet-50	0.939 ± 0.014	0.938 ± 0.006	**0.864 ± 0.028**	**0.855 ± 0.012**
Resnext-50	0.940 ± 0.011	0.932 ± 0.006	0.856 ± 0.029	0.837 ± 0.007
Densenet-201	0.937 ± 0.012	**0.942 ± 0.006**	0.862 ± 0.032	0.834 ± 0.015
Resnet-152	0.935 ± 0.012	0.937 ± 0.006	0.859 ± 0.035	0.846 ± 0.013
Seresnet-152	0.936 ± 0.013	0.925 ± 0.009	0.857 ± 0.024	0.838 ± 0.015
Seresnext-50	0.934 ± 0.013	0.931 ± 0.006	0.856 ± 0.027	0.845 ± 0.013

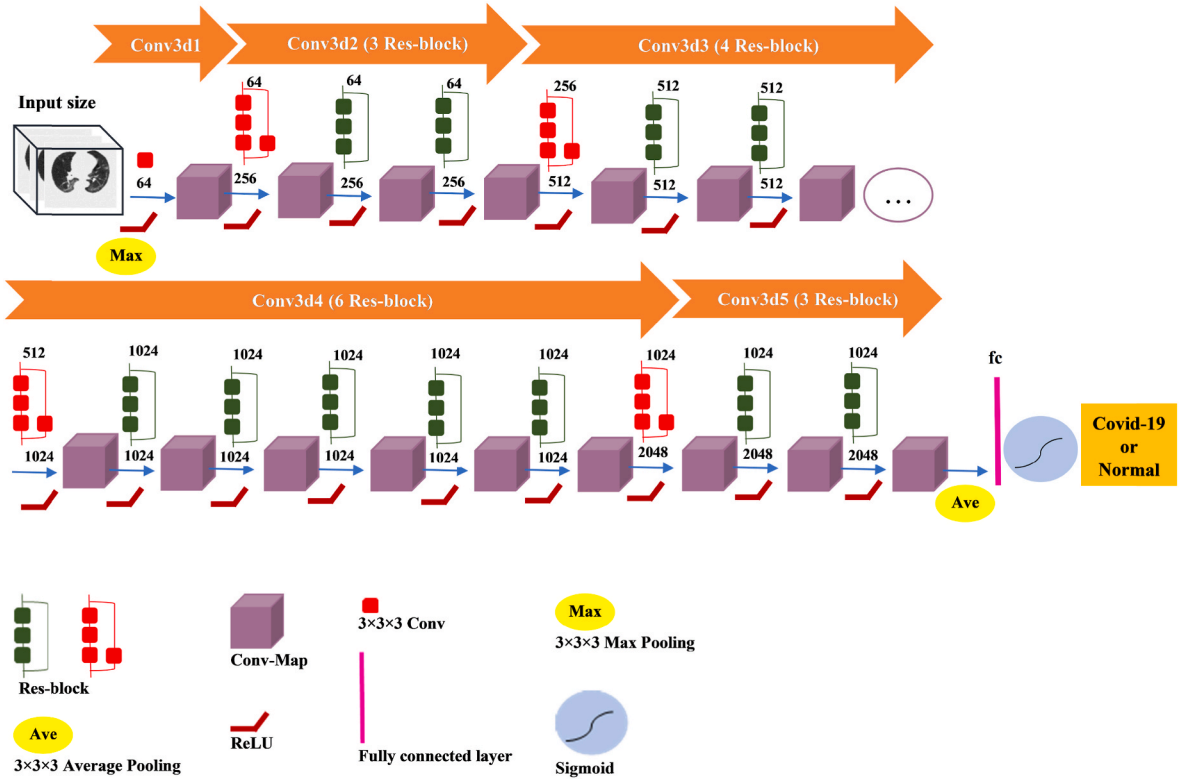


Fig. 5. Architecture of 3D ResNet-50. The segmented lung images are fed to the model, and the model output would be the predicted probability of COVID-19 positive or normal.

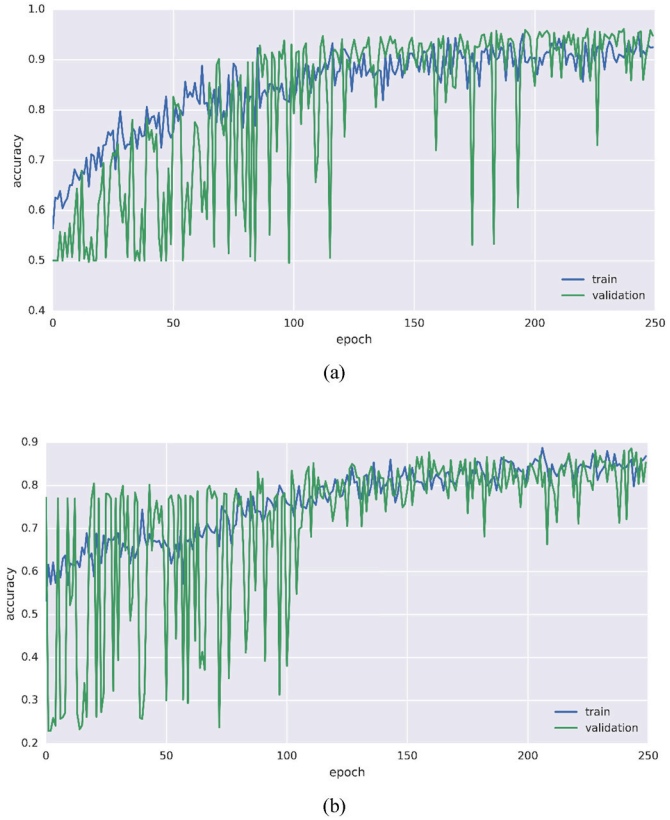


Fig. 6. Learning curve for training on (a) Iranmehr data using ResNet-50, and (b) Moscow data using ResNet-50.

ResNet-50, ResNet-152, DenseNet-169, DenseNet-201, Resnext-50, Seresnext-50, and Seresnet-152 were assessed for generalizability. In the present study, binary classification was performed using common models that have been used in previous studies [68–70].

Furthermore, for our 3-dimensional data, more complex models, such as ResNets and DenseNets, provided better results than simpler models, such as VGG. We used 1110 3D CT images of Iranmehr and Moscow dataset in this stage and fed 100% of the data for training. Training was carried out on one dataset and was tested against another whole dataset, i.e., training on 1110 Iranmehr dataset and testing against 1110 Moscow dataset, and training on 1110 Moscow dataset and testing on 1110 Iranmehr dataset. Several hyper-parameters were tested, including learning rate, different patch sizes, and the number of training iterations. After parameter tuning, training was performed using seven mentioned models for an initial learning rate of 10^{-4} and 250 epochs for 5-fold as the best-selected hyper-parameters. We selected ResNet-50 as the 3D model for the classification due to the better results (Table 1) and lower runtime, consistent with previous studies into COVID-19 classification from 3D CT images [71]. An overview of Table 1 shows that ResNet-50 outperforms other models. In particular, the results of the ResNet-50 are better than the DenseNet-169 on the Moscow dataset in terms of accuracy and standard deviation. Regarding the comparison of cropped and non-cropped images, all models had better results on cropped images, except for ResNet-152 and DenseNet-201.

According to our inspections, compared to DenseNet-169 (and likewise other models) ResNet-50 focuses on the inner structure of the lung rather than borders. It might be a reason that ResNet-50 outperforms other models on our datasets.

From Table 1, we can see that ResNet-50 performs the best on the Moscow dataset, specifically on cropped images. For the Iranmehr dataset, ResNet-50 performed considerably well compared to the best-performed model, DenseNet-169. We want an to highlight that the Moscow dataset is in NIFTI image format. On the other hand, the

Table 2
Training and Testing data percentage for the base experiment.

Test No.	Training data	Testing data
1	80% Iranmehr	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
2	80% Moscow	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
3	80% Iranmehr +80% Moscow	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC

Table 3
Training and Testing data percentage for the first experiment.

Test No.	Training data	Testing data
1	Iranmehr 80% + Moscow 20%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
2	Iranmehr 80% + Moscow 40%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
3	Iranmehr 80% + Moscow 60%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
4	Moscow 80% + Iranmehr 20%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
5	Moscow 80% + Iranmehr 40%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
6	Moscow 80% + Iranmehr 60%	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC

Iranmehr dataset image type is DICOM which retains image details better. Hence, it is of better quality. Our key reason for selecting ResNet-50 is that it performs best with the lower quality NIFTI image type from the Moscow dataset while performing really well with the better quality DICOM image type from the Iranmehr dataset. Furthermore, all deep learning models performed better on the DICOM image type, Iranmehr dataset, than the NFTI image type, which is the Moscow dataset.

In the generalizability assessment approach, all models were evaluated by the mean of k-fold running parameters on external tests. Since our study is a binary simulation, we used the sigmoid activation function [72]. Fig. 5 illustrates 3D ResNet-50 structure [73], and Fig. 6 shows the resultant learning curves. The "jitter" that is noticeable on the earlier training epochs is caused by fluctuations in training loss which is the consequence of training a very large network (ResNet 50 with approximately 50 million trainable parameters) using datasets of 1110 3D CT scans. Also, COVID-19 lung involvement is not apparent, or it may be very subtle in some patient CT slices. As a result, when there are several such slices in a training batch, the validation loss for that batch will be very small because gradient descent is minimal. As training proceeds, the neural network becomes more tolerant of these adversarial images resulting in a smoother training curve in the later training epochs.

3.5. Generalization

Based on 1110 3D CT images of the Moscow dataset, we randomly used the same number from the Iranmehr dataset. The allocation to training, validation, and test groups and splitting was done randomly. We split each dataset into five segments of 20% each, with equal distribution of normal and COVID-19 in each segment. We separated one 20% segment from each dataset as holdout test set and kept it the same for all experiments, and all training parts were performed with the remaining 80% from each dataset. The generalization evaluation was carried out in three experiments, as follows.

Table 4
Training and Testing data for the second experiment (transfer learning).

Test No.	Training Data	Testing Data
1	80% Iranmehr	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC
2	80% Moscow	20% Iranmehr holdout set 20% Moscow holdout set LDCT 3DLSC

3.5.1. Base experiment

Models were fully trained in 3 experiments using 80% Iranmehr, 80% Moscow, and 80% Iranmehr +80% Moscow data, respectively. These models were tested with 20% holdout sets separately from Iranmehr and Moscow sets. The details shown in Table 2 serve as the base experiment results for generalization tests.

3.5.2. First experiment

Models trained with 80% of Iranmehr dataset with the addition of an increasing portion of Moscow dataset (20%, 40%, 60%). Similarly, Models were trained with 80% Moscow dataset and an increasing portion of the Iranmehr dataset (20%, 40%, 60%). All models were then tested on one 20% holdout set from each dataset. Additionally, they were tested on two external datasets to evaluate the effect of adding different combinations to the dataset. Details of these six tests are given in Table 3.

Table 5
Different model results when trained with Iranmehr dataset and tested against Moscow dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	**0.814**	0.777	0.888	0.858	87 ± 0.01
DenseNet-201	0.802	0.762	0.876	0.846	86 ± 0.01
ResNet-152	0.796	0.775	**0.896**	0.858	88 ± 0.01
ResNet-50	0.800	0.799	0.818	0.861	**89 ± 0.01**
ResnNext-50	0.788	**0.813**	0.765	**0.862**	89 ± 0.02
Seresnet-152	0.800	0.763	0.889	0.849	88 ± 0.00
Seresnext50	0.792	0.758	0.9	0.847	88 ± 0.01

Table 6
Different model results when trained with Iranmehr dataset and tested against LDCT dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	**0.935**	0.901	0.982	**0.943**	0.96 ± 0.01
DenseNet-201	0.917	0.875	**0.978**	0.927	95 ± 0.02
ResNet-152	0.916	0.892	0.960	0.932	96 ± 0.01
ResNet-50	0.920	0.892	0.964	0.933	96 ± 0.01
ResnNext-50	0.910	**0.907**	0.971	**0.943**	**96 ± 0.00**
Seresnet-152	0.897	0.878	0.957	0.924	95 ± 0.01
Seresnext50	0.912	0.886	0.964	0.930	94 ± 0.00

Table 7
Different model results when trained with Iranmehr dataset and tested against 3DLSC dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	**0.889**	0.794	**0.970**	0.870	93 ± 0.02
DenseNet-201	0.869	0.78	0.910	0.832	93 ± 0.03
ResNet-152	0.871	0.808	0.958	0.873	**94 ± 0.01**
ResNet-50	0.857	0.798	0.922	0.850	94 ± 0.03
ResnNext-50	0.848	0.842	0.887	0.861	94 ± 0.03
Seresnet-152	0.866	**0.844**	0.922	**0.879**	94 ± 0.02
Seresnext50	0.863	0.786	0.943	0.853	94 ± 0.02

3.5.3. Second experiment

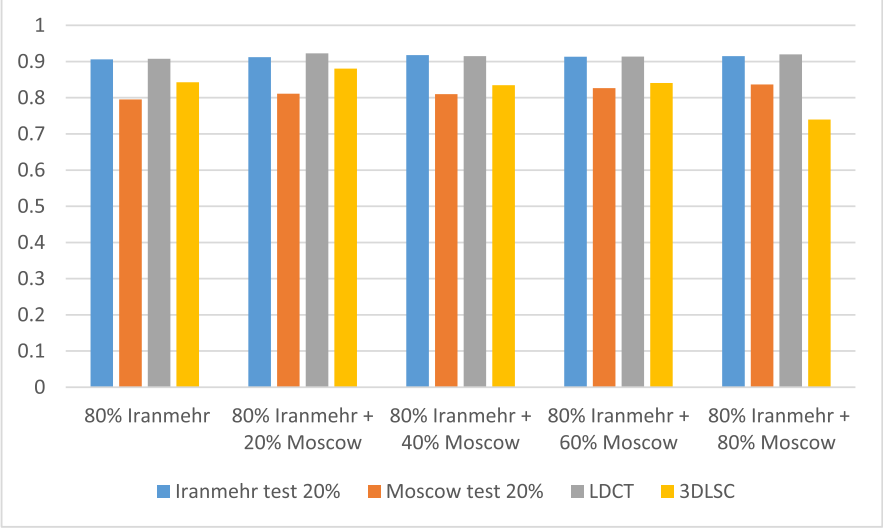
Transfer learning was employed to check how it helps with generalization. This task included three steps for each dataset of Moscow and Iranmehr datasets.

1. Using weights of the full trained model with 80% of one dataset.
2. Retraining three last layers with 80% of the other dataset.
3. Testing on Iranmehr and Moscow holdout test data, and two external datasets: LDCT 3DLSC (See Table 4).

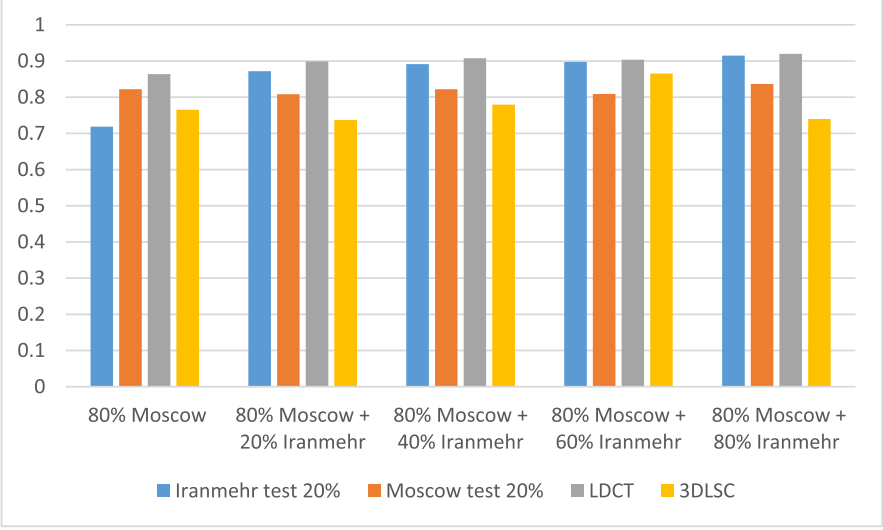
This experiment used the last Conv3D layer with 1048576 trainable parameters, the last batch normalization with 8192 trainable parameters, and the fully connected layer with 2049 trainable parameters. For instance, when we trained 80% of the Moscow dataset using transfer learning, we loaded the weight obtained from 80% full training of the Iranmehr dataset. A similar approach for transfer learning uses 80% of the Iranmehr dataset. Based on trial and error, we selected three last layers of the model to have transfer learning with best possible accuracy.

3.6. Evaluation metrics

Classification performance for all trained models was evaluated by several statistical measures: accuracy (the percentage of correctly classified test cases, Eq. (1)), sensitivity (the percentage of correct COVID-19 detected cases, Eq. (2)), specificity (the percentage of correct normal classified cases, Eq. (3)), F1-score [74], Eq. (4), and area under the curve (AUC) of receiver operating characteristic (ROC) which is the true positive rate (TPR) against false positive rate (FPR) [75]. The mathematical formulation of the statistical measures is given below.

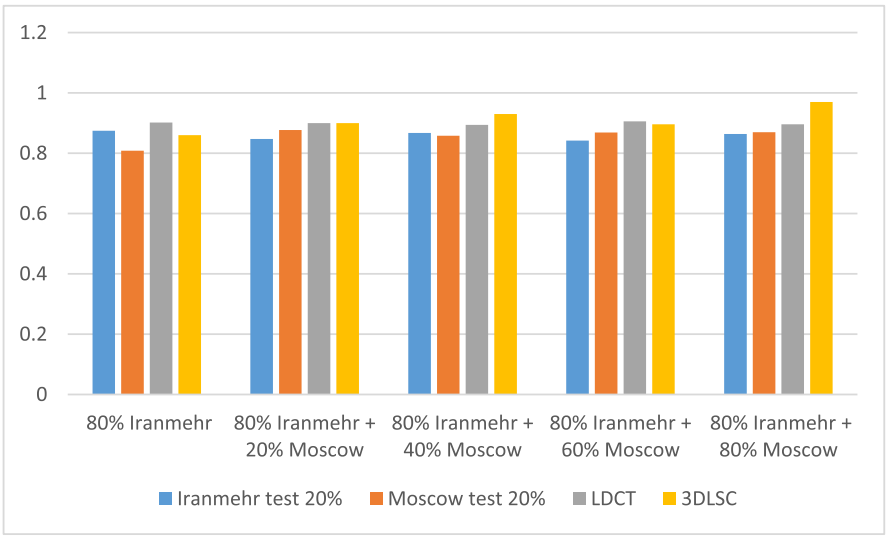


(a)

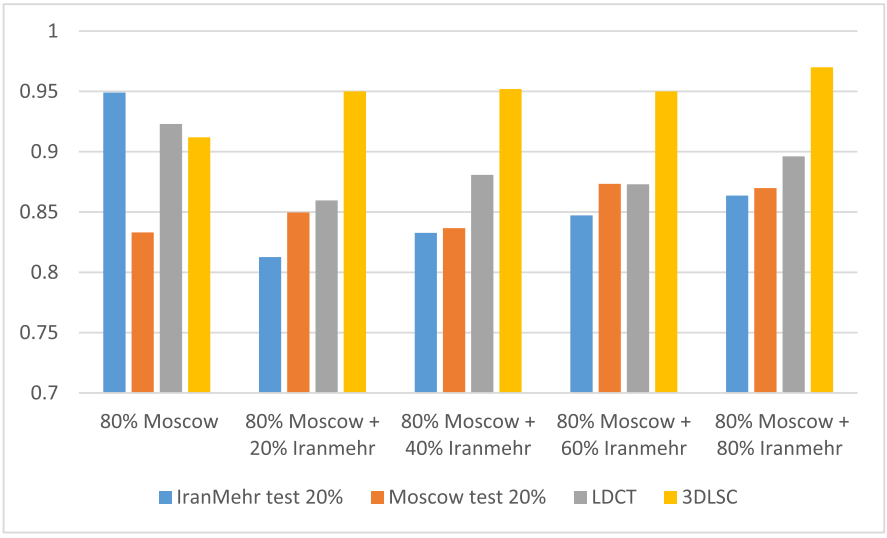


(b)

Fig. 7. Accuracy results for a different portion of datasets when (a) 80% of the dataset is Iranmehr and splits of Moscow are added, and (b) 80% of the dataset is Moscow and splits of Iranmehr are added.



(a)



(b)

Fig. 8. Sensitivity results for a different portion of datasets when (a) 80% of the dataset is Iranmehr and splits of Moscow are added, and (b) 80% of the dataset is Moscow and splits of Iranmehr are added.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4}$$

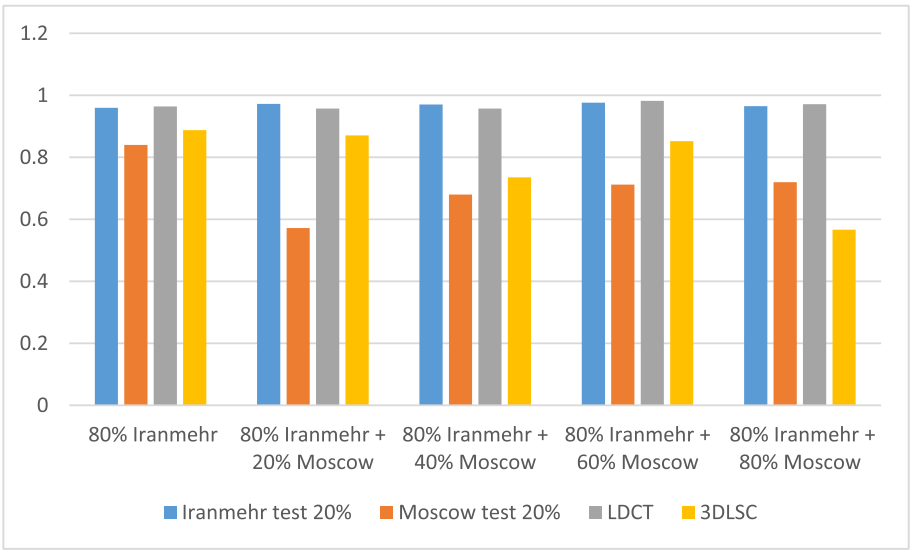
where, TP is the number of true positives, TN is true negative, FP stands for the number of false positives, and FN indicates the number of false negatives.

4. Results

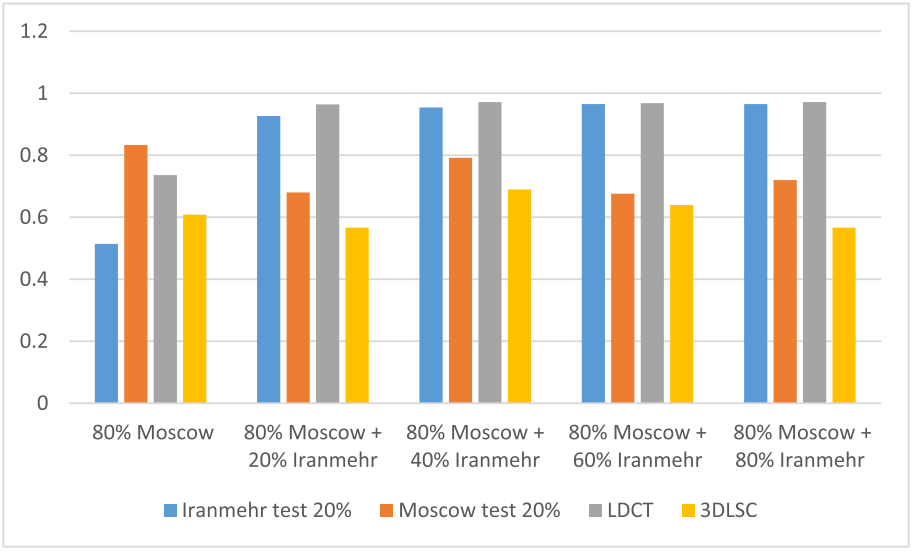
The statistic and AUC results for model selection (external-validation evaluation) for training on Iranmehr dataset and tested on Moscow dataset, LDCT, and 3DLSC are presented in Tables 5–7, respectively. The reverse process, i.e., training on the Moscow dataset and testing on the Iranmehr dataset, LDCT, and 3DLSC are presented in 8–10, respectively. According to the tables, the results of training on Iranmehr and testing on LDCT are higher than other test datasets. However, it should be noted that compared to LDCT and 3DLSC, Moscow and Iranmehr datasets contain 1110 images, which increases the testing validity.

The accuracy of a different combination of datasets of experiment phases is given in Table 11. According to Table 11, the combination of 80% of one dataset with the addition of different of the other has close accuracy to the accuracy of the total combination. Table 12 presents the AUC results of different combinations of datasets of experiment phases. According to Table 12, all AUC results are near to the AUC of the total combination. In Table 13, the statistical transfer learning results are presented. By general overview of Table 13, it can be found that for all metrics except for specificity, the results of retraining on 80% Moscow dataset using the weights of a full run of Iranmehr dataset are higher compared to the retraining on 80% Iranmehr dataset using the weights of Moscow dataset full run.

The comparison diagrams of accuracy, sensitivity, specificity, and F1 score are presented in Figs. 7–10. Additionally, Fig. 11 presents the



(a)



(b)

Fig. 9. Specificity results for a different portion of datasets when (a) 80% of the dataset is Iranmehr and splits of Moscow are added, and (b) 80% of the dataset is Moscow and splits of Iranmehr are added.

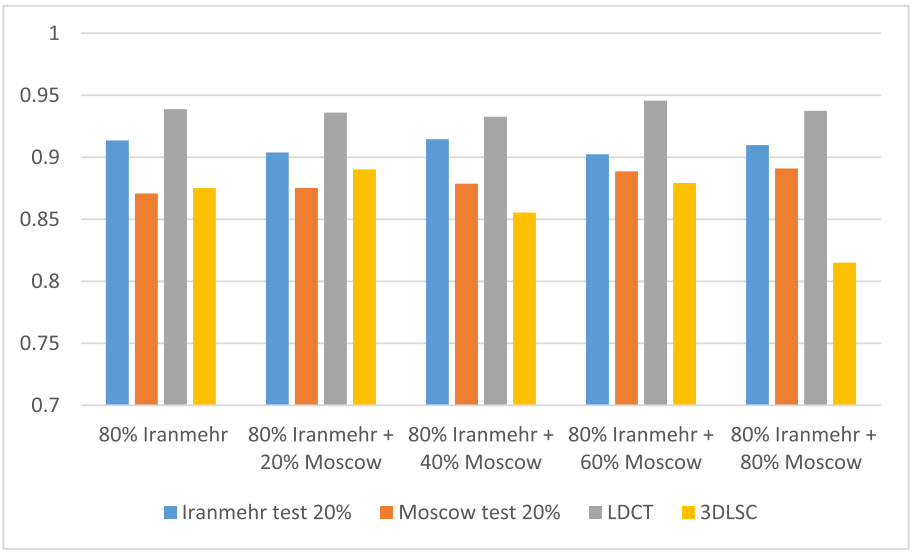
confusion matrix results for all three experiments. According to Fig. 7, the accuracy of different combinations of datasets almost smoothly grows. We can see that the combination of 80% from one dataset with the addition of a different portion of the other datasets performs similarly on test sets. Fig. 8 illustrates the sensitivity of different combinations. Based on the results shown in Fig. 8, for all the combinations, the results of testing on holdout test sets are considerably different from each other. Regarding to the specificity, it can be seen from Fig. 9 that the combination above 80% from one dataset and 40% of the other have similar results to the total combination. For F1 score, as shown in Fig. 10, the combination of 80% of one dataset added to 40% of the other reaches the results near to the total combination.

In Fig. 11(a–c), the confusion matrix results are presented when different combinations and transfer learning results are tested against the unseen holdout dataset. The highest number of TPs belongs to the total combination, and other combinations have close results when tested on holdout test set of Moscow and Iranmehr. However, when testing on the external datasets, 3DLSC, we can see that the numbers of FPs are high in combination of 80% Moscow and 20% Iranmehr. Following the total combination, 80% Moscow and 20% Iranmehr and 80% Moscow added to 20% Iranmehr have the lowest number of FNs when tested on Iranmehr holdout test set.

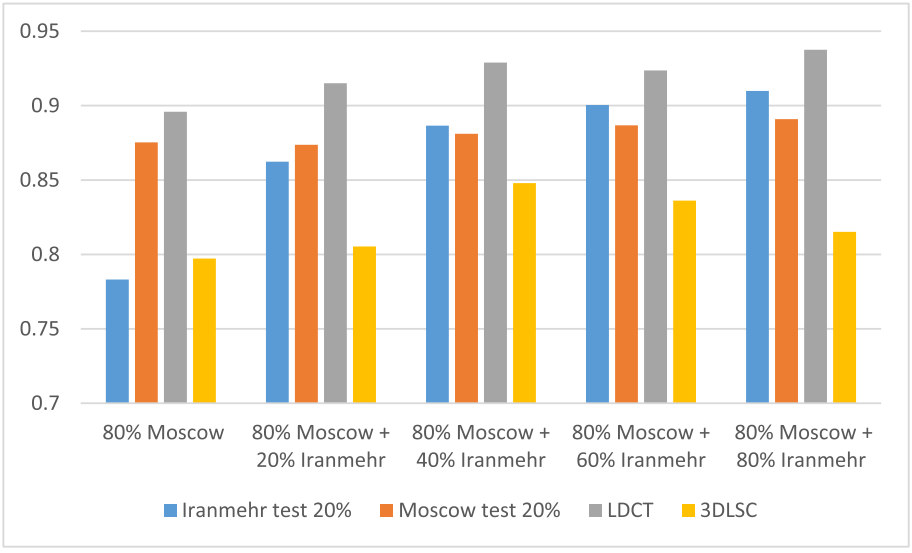
5. Discussion

The results of AI-based models seem to be more reliable when they use 3D CT images and are tested for generalizability. The reason is that more features can be extracted in whole 3D slices compared to 2D implementations [50,54]. As many COVID-19 CT images show, not all slices of a patient's image contain involvement. Therefore, considering the slice-based classification of COVID-19 and normal cases may not be as realistic as whole CT slices for each patient. This is especially true when the involvement is very small, and its detection is possible only when the slice is compared with neighbor slices.

According to previous studies [24,44], and in our many experimental trials, lung segmentation improves the results of COVID-19 and normal cases' classification and should be considered in preprocessing. This is probably due to the fact that it prevents the model from focusing on



(a)



(b)

Fig. 10. F1-score results for a different portion of datasets when (a) 80% of the dataset is Iranmehr and splits of Moscow are added, and (b) 80% of the dataset is Moscow and splits of Iranmehr are added.

unwanted targets like bone and soft tissue. Segmentation results are also affected by image type. We used two different segmentation approaches to segment lung from NIFTI and DICOM CT images since no single segmentation method works for all image formats. Also, a patch-based approach both for the compensation of imbalanced classes and to overcome overfitting showed the capacity to be taken into account for 3D medical datasets which may suffer from a low number of images and imbalanced classes. Through trial and error in the current study, it has been shown that patch methods improved results by up to 2% over non-patches methods.

Based on our model selection simulations, which can be considered external-validation evaluation, the generalizability of the 3D ResNet-50 and procedures undertaken in this study indicate that the accuracy when trained with Iranmehr and tested on external datasets is above 78% with the AUC of around 0.90. According to the statistics presented in Tables 5–10, the Iranmehr dataset produces a generalizable model. This may be due to the precise data categorization in Iranmehr dataset as COVID-19 and normal patients for the training phase. Moreover, a general overview of Tables 5–10 reveals that Iranmehr and LDCT datasets have better results compared to Moscow and 3DLSC. This also may be related to the NIFTI format of these datasets, which seems to affect classification results compared to the DICOM format. In the NIFTI format, the file no longer contains the granular and detailed information to benefit from DICOM's broad and complex header structures. NIFTI images are more pixelated, and the conversion also affects the image. These pixelized images of NIFTI format make the inner structure of feature maps blurrier, making it more difficult to extract features from them. It is one of our findings that when combining different datasets, we need to be aware of image types and analyze how they impact the performance of deep learning models. Since two out of four external test datasets have a large number of images (1110 images), and experiments were carried out in a 5-fold approach, the test results are reproducible.

The main purpose of this study is to evaluate the effect of different portion combinations of datasets on generalizability. This study confirms that, although the total combination produced the best results with less overfitting (as shown in Fig. 12), different combinations of datasets provide close results. Moreover, in many studies, especially for the tasks related to medical images, accessibility, preparation, preprocessing may

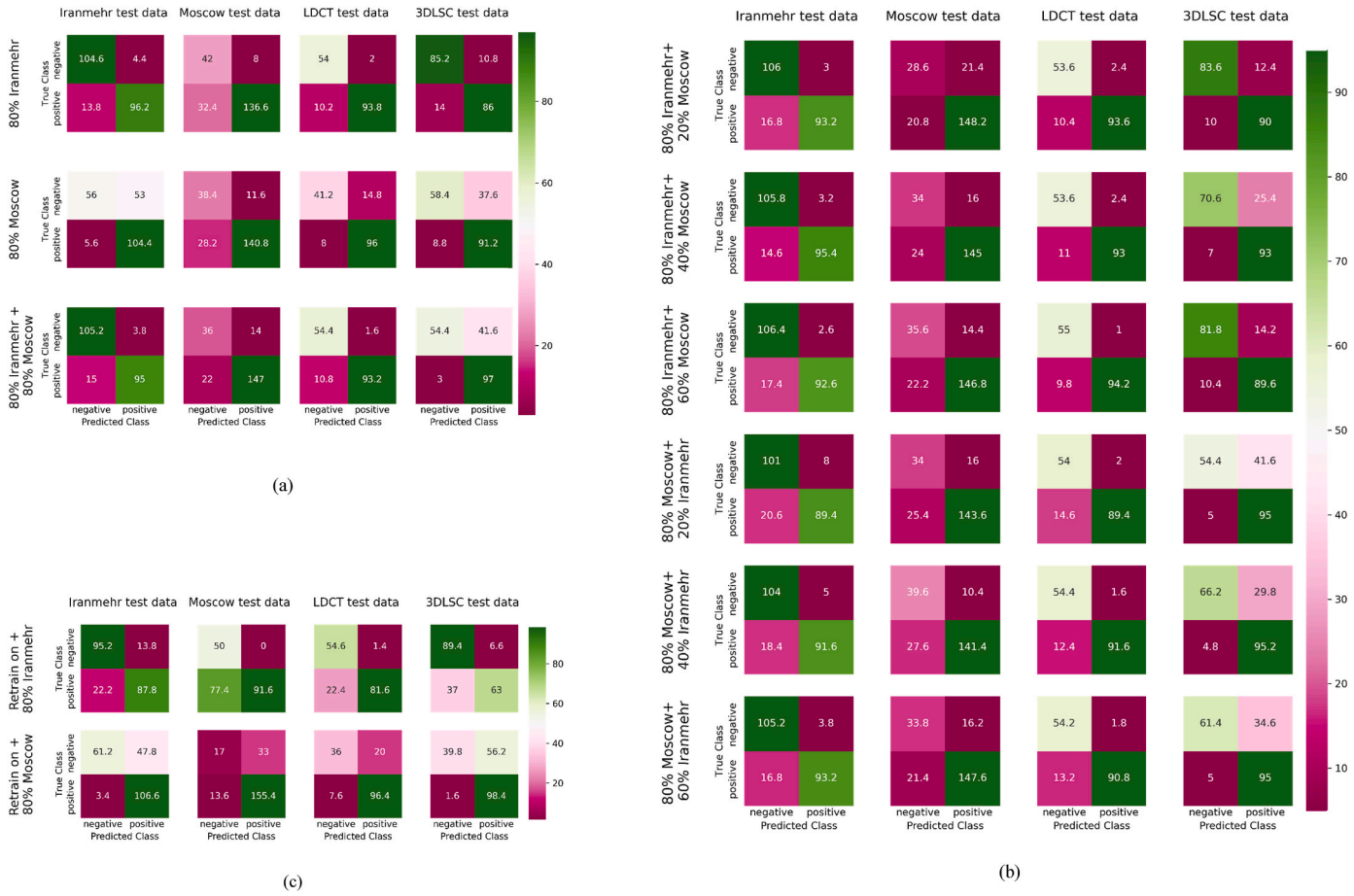


Fig. 11. The results of confusion matrix for (a) base experiment; (b) first experiment; and (c) second experiment (transfer learning).

Table 8
Different model results when trained with Moscow dataset and tested against Iranmehr dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	**0.765**	0.943	0.507	0.765	92 ± 0.01
DenseNet-201	0.716	0.921	0.636	0.809	92 ± 0.01
ResNet-152	0.677	0.952	0.552	0.795	92 ± 0.02
ResNet-50	0.741	0.946	0.595	0.805	92 ± 0.02
ResnNext-50	0.699	**0.96**	0.492	0.781	92 ± 0.02
Seresnet-152	0.716	0.953	0.548	0.797	92 ± 0.01
Seresnext50	0.745	0.918	**0.681**	**0.822**	**93 ± 0.01**

Table 9
Different model results when trained with Moscow dataset and tested against LDCT dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	0.880	0.915	0.764	0.898	94 ± 0.01
DenseNet-201	0.827	0.905	0.835	0.909	94 ± 0.02
ResNet-152	0.847	**0.930**	0.782	0.910	**95 ± 0.01**
ResNet-50	**0.889**	0.913	**0.857**	**0.918**	**95 ± 0.01**
ResnNext-50	0.850	0.923	0.757	0.900	94 ± 0.01
Seresnet-152	0.877	0.909	0.775	0.897	94 ± 0.01
Seresnext50	0.846	0.894	0.814	0.898	94 ± 0.01

impose difficulties and sometimes be computationally expensive, especially in training on 3D images [76]. With this aim, we divided two available 3D datasets, i.e., Iranmehr and Moscow, into the five 20% portions, and we evaluated the different combination results.

Table 10
Different model results when trained with Moscow dataset and tested against 3DLSC dataset.

3D Models	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DenseNet-169	0.728	0.886	0.672	**0.805**	90 ± 0.03
DenseNet-201	0.715	0.882	**0.693**	0.811	90 ± 0.02
ResNet-152	0.710	**0.954**	0.472	0.777	91 ± 0.02
ResNet-50	0.771	0.91	0.662	0.817	**92 ± 0.02**
ResnNext-50	0.737	0.9	0.591	0.786	89 ± 0.03
Seresnet-152	**0.765**	0.888	0.662	**0.805**	90 ± 0.03
Seresnext50	0.727	0.928	0.587	0.800	**92 ± 0.02**

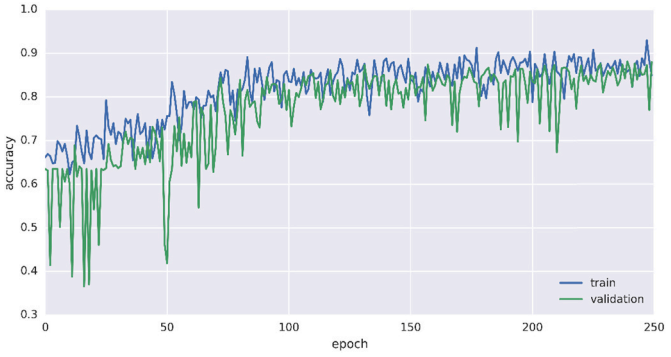


Fig. 12. Learning curve for training on 80% Iranmehr 80% Moscow data using ResNet-50.

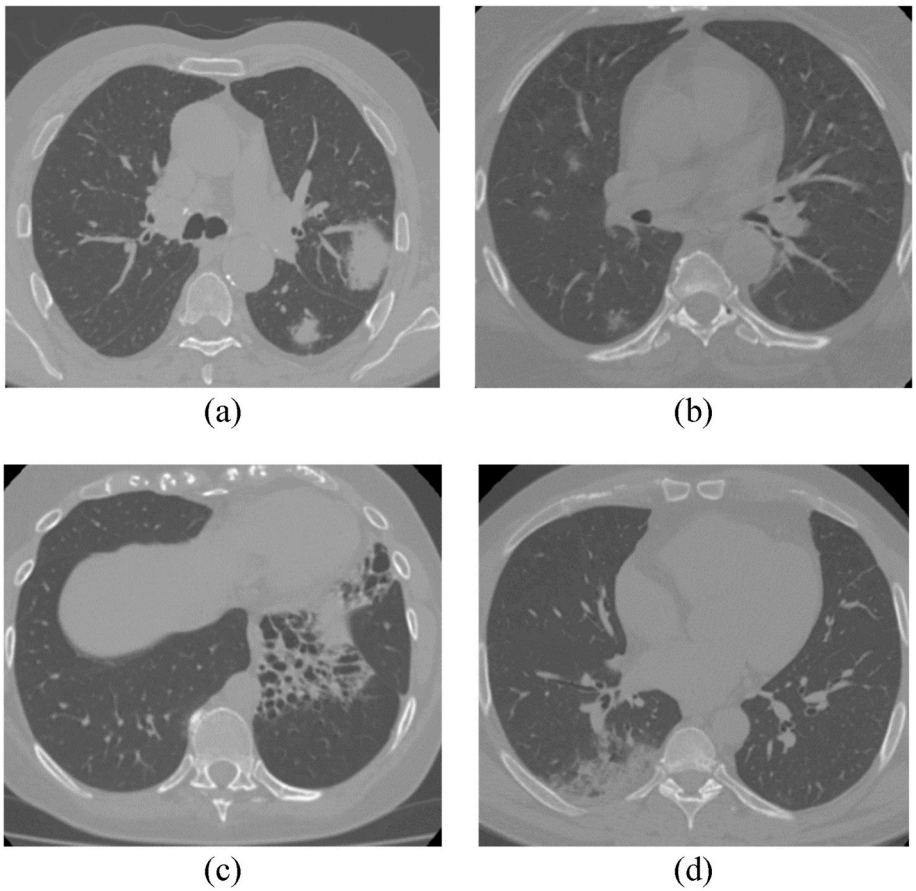


Fig. 13. Some suspected cases involved in Moscow normal dataset folder. (a) case 7 diagnosed with patchy consolidation, which can be pneumonia including COVID-19 or tumoral lesions, (b) case 34 is more probably a COVID-19 case with small involvement (c) case 68 diagnosed with honeycombing fibrosis at left lung lower lobe, and (d) case 77, diagnosed with wedge consolidation at base of the right lung which may arise due to the pulmonary thromboemboli or segmental pneumonia.

Table 11
Accuracy (%) of trained ResNet-50 for base and first experiments.

Training setting	Iranmehr test data (20%)	Moscow test data (20%)	LDCT	3DLSC
80% Iranmehr	90.5	79.5	90.7	84.28
80% Moscow	71.8	82.1	86.3	76.5
80% Iranmehr + 20% Moscow	91.2	81.0	92.2	88.0
80% Iranmehr + 40% Moscow	91.7	81.0	91.4	83.4
80% Iranmehr + 60% Moscow	91.3	82.6	91.3	84.0
80% Moscow + 20% Iranmehr	87.2	80.8	89.8	73.6
80% Moscow + 40% Iranmehr	89.1	82.1	90.7	77.9
80% Moscow + 60% Iranmehr	89.7	80.9	90.3	86.5
80% Iranmehr + 80% Moscow	91.5	83.6	91.9	73.9

Table 12
AUC ± std of trained ResNet-50 for base and first experiments.

Training setting	Iranmehr test data (20%)	Moscow test data (20%)	LDCT	3DLSC
80% Iranmehr	0.096 ± 0.00	0.87 ± 0.03	0.96 ± 0.01	0.95 ± 0.02
80% Moscow	0.92 ± 0.01	0.91 ± 0.01	0.94 ± 0.01	0.90 ± 0.02
80% Iranmehr + 20% Moscow	0.95 ± 0.01	0.86 ± 0.02	0.95 ± 0.01	0.96 ± 0.01
80% Iranmehr + 40% Moscow	0.95 ± 0.01	0.87 ± 0.03	0.95 ± 0.01	0.96 ± 0.01
80% Iranmehr + 60% Moscow	0.94 ± 0.01	0.89 ± 0.01	0.96 ± 0.01	0.95 ± 0.02
80% Moscow + 20% Iranmehr	0.92 ± 0.01	0.88 ± 0.01	0.94 ± 0.00	0.94 ± 0.01
80% Moscow + 40% Iranmehr	0.94 ± 0.00	0.89 ± 0.01	0.95 ± 0.01	0.95 ± 0.02
80% Moscow + 60% Iranmehr	0.94 ± 0.00	0.88 ± 0.02	0.95 ± 0.00	0.95 ± 0.03
80% Iranmehr + 80% Moscow	0.95 ± 0.01	0.90 ± 0.02	0.95 ± 0.01	0.95 ± 0.01

Regarding the accuracy, according to the results presented in Fig. 7, the combination of 80% of one dataset and 40% or 60% of the other reaches to acceptable results close to the results obtained from the total. It seems that when only 20% of each dataset is added to the other, the model encounters new features trying to learn them. However, since the number of images in the 20% portion is much lower than that of the 80% portion, the bias occurs, and the results on holdout test sets have a higher difference compared to other combinations. According to Fig. 8 (a), the sensitivity of different combinations succeeded in learning most of the features, and the results are near the total combination when the training set is Iranmehr. However, when the training set is Moscow, while the behavior of different combinations is similar, the results above 80% and 40% resemble those of the total combination (see Fig. 8(b)). In terms of specificity, Fig. 9 (a and b) demonstrate that for different combinations, the test on the Iranmehr holdout test set and LDCT is more successful than that on the Moscow holdout test set and 3DLSC.

It was observed that specificity is dramatically low. According to our

Table 13
Statistics of trained ResNet-50 for transfer learning.

		Training setting	
	Test sets	Retrain on 80% Iranmehr	Retrain on 80% Moscow
Accuracy (%)	Iranmehr test data (20%)	83.2	76.8
	Moscow test data (20%)	64.2	79.1
	LDCT	**85**	**82.3**
	3DLSC	77.6	71.2
Sensitivity (%)	Iranmehr test data (20%)	**79.8**	96.9
	Moscow test data (20%)	54.2	91.9
	LDCT	78.4	92.6
	3DLSC	63	**98.4**
Specificity (%)	Iranmehr test data (20%)	87.3	56.1
	Moscow test data (20%)	**100**	34
	LDCT	97.5	**64.2**
	3DLSC	93.1	41.4
F1-Score	Iranmehr test data (20%)	82.9	80.8
	Moscow test data (20%)	70.2	**86.9**
	LDCT	**87.2**	87.5
	3DLSC	74.2	77.4
AUC	Iranmehr test data (20%)	89 ± 0.00	**0.95 ± 0.00**
	Moscow test data (20%)	0.90 ± 0.00	0.86 ± 0.01
	LDCT	**0.93 ± 0.00**	0.94 ± 0.01
	3DLSC	86 ± 0.00	0.94 ± 0.01

radiologists, and as shown in Fig. 13, there are some cases in CT0 (normal dataset) of Moscow dataset that is not normal lungs, so they can affect the classification results. We didn't remove any case from the Moscow dataset to avoid data manipulation. Also, we see that for almost all metrics, the Moscow dataset has an adverse effect. This indicates that public datasets still cannot be treated as ideal as real clinical data. Specifically, in specificity results illustrated in Fig. 9, the results of specificity are much lower when tested on holdout 20% of Moscow test set and 3DLSC, even for total combination.

The high accuracy results in each combination or high F1-score (Fig. 10 a and b) show the network's capability either for COVID-19 detection or screening of similar disease type and the capability for screening. Another reason to demonstrate this capability is the much higher TP for each combination (Fig. 11(a–c)).

Several studies have combined smaller COVID-19 CT datasets into "supersets" to maximize the number of training samples for deep learning models. Previous studies have not investigated the effect of combining CT corpora in this manner. In this study, we proved that combining datasets is an effective approach to training deep learning models for COVID-19 detection for the CT imaging mode. We found a "saddle point" at the 80:40% mix of datasets for the datasets investigated. According to our interpretation, 80% of a primary dataset is adequate for fully training a model, and the additional fine-tuning using 40% of a secondary dataset helps the model generalize to a third, unseen dataset.

Our second experiment used transfer learning as an alternative deep learning approach for training models [77]. It is clear that when we use the results of the full run for a similar image type, the results are better. It means that pretraining on medical images are more suitable for retraining medical images than using pretrained JPEG images such as ImageNet weights of generic images [78]. Therefore, we used the weights from the full run of 80% of each dataset in retraining the last three layers using the other 80% dataset. According to Table 13, the results show that when the weights come from the more accurate full run (here, full run using 80% Iranmehr), the result of retraining is better. However, the result of transfer learning is still lower than that of total combination full run, i.e., the full run of 80% of one dataset with the addition of 80% of the other dataset. Nevertheless, given the results presented in Table 13, the transfer learning technique allows for fewer data and faster training while providing close AUC and accuracy to those of total combination full run.

6. Limitations

The available large 3D datasets in DICOM or NIFTI format were really rare. The other 3D image format, like JPEG, if they exist, suffers from low resolution compared to DICOM or NIFTI, and also, they are not considered as clinical image formats to be a real reflection of what is used in reality. Consequently, this study was performed only on two large datasets available to us for training and two small datasets for testing in which their image format was DICOM/NIFTI. In further studies, more 3D CT images are needed to be done before this could be part of a clinical workflow. Besides, there is a need for one accurate approach for lung segmentation that can give accurately segmented lungs from different image formats. Therefore, we used two segmentation approaches for two image types, which made the preprocessing part different for each image format.

7. Conclusion

This study thoroughly assessed the impact of image preprocessing, different 3D CNN models, and different combinations of datasets on generalizability. The results indicated that the different combinations of 3D CT images, lung segmentation to improve the signal-to-noise ratio, and patching in the training process to avoid overfitting and dataset imbalance improve generalizability. Also, in the absence of a large dataset, we showed that combining 80% of one dataset with a 40% or greater contribution from another dataset produces results comparable to the total combination. This is potentially very helpful in clinical applications which are often constrained by the lack of a sizable external dataset and/or the difficulty of time-consuming image preparation and labeling processes. Although the total combination obtained the best results, we found diminishing returns after the combination of 80% and 40%, respectively. About dataset types, studies on real clinical data are more reliable with regard to biases that may exist in public datasets since most of the public datasets are stored in formats other than DICOM, which is used in real clinical modalities, and there are also doubts about their correct diagnosis. We saw that public Moscow dataset training results were considerably lower than those for Iranmehr data. The same was true for LDCT and 3DLSC. This was because of some mistakenly grouped cases in the Moscow dataset, adversely affecting transfer learning results. The results of the transfer learning approach in this study highly suggest that weights of training on a precisely sorted dataset should be used to produce more accurate results in future studies. The simulations were run on our own medical data without using the pre-trained weights of models. Given the limited access to medical data, we attempted to achieve good results on our new datasets by cropping, segmentation, patching, etc., to improve training performance. This study would be useful in practice where datasets are limited and may be collected in different locations and settings.

Funding

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

Declaration of competing interest

The authors declare no conflict of interest.

Ethical standard

This study was carried out under the ethics approval from the University of Technology Sydney, Australia, under "UTS HREC REF NO. ETH21-6536".

Summary

For detection of COVID-19 positive cases, artificial intelligence (AI)-based technologies can be of immense assistance. However, the results obtained from one specific dataset may not be consistent with the external dataset. Furthermore, a second large dataset is often not accessible, and it is very time-consuming to prepare it. We have studied how COVID-19 CT image datasets can be combined to assess generalizability with unique combinations of two large data sets using state-of-the-art 3D convolutional neural networks (CNN), which shows that a combination of datasets can assist generalization and determining the optimal "combined" characteristics.

In this study, 1110 3D CT images were used from two large datasets; one from Iranmehr hospital, Tehran, Iran, and one from publicly available datasets from hospitals located in Moscow, Russia. We split each dataset into five segments of 20% each, with equal distribution of normal and COVID-19 in each segment. We separated one 20% segment from each dataset as holdout test set and kept it the same for all experiments, and all training parts were performed with the remaining 80% from each dataset. Two small external test datasets including LDCT and 3DLSC were used to evaluate the validity of the present study. We carried out three experiments as follows: (a) Base experiment: Models were fully trained in 3 experiments using 80% Iranmehr, 80% Moscow, and 80% Iranmehr + 80% Moscow data, respectively. These models were tested with 20% holdout sets separately from Iranmehr and Moscow sets. (b) First experiment: Models trained with 80% of Iranmehr dataset + increasing portion of Moscow dataset (20%, 40%, 60%). Similarly, Models were trained with 80% Moscow dataset + an increasing portion of Iranmehr dataset (20%, 40%, 60%). All models then tested with one 20% holdout set from each dataset and two small external test sets(c) Second experiment (transfer learning): weights of the full trained model with 80% of one dataset were used. Three last layers (including a fully connected layer, batch normalization, and a 3D convolution layer) were retrained with 80% of the other dataset. All simulations were carried out in a 5-fold manner to increase the reliability of the results. The results were tested on Iranmehr and Moscow holdout test data, LDCT, and 3DLSC datasets.

The total combination of 80% of each dataset has an accuracy of 91% on Iranmehr and 83% on Moscow holdout test datasets. The results show that all other combinations provided near to that of the total combination of datasets. The results of the transfer learning study led to the accuracy of 83% on Iranmehr holdout test set and 85% on LDCT external test set when retrained on 80% of Iranmehr dataset.

The results of splitting dataset demonstrated that although the total combination of 80% of both datasets has the best results; the different combinations of datasets still lead to acceptable results in terms of generalizability. Transfer learning performed in this study showed lower statistical results compared to the total combination, but still gave satisfactory results. Adopting the proposed methodology described in this work may help to obtain satisfactory results for normal/abnormal lung screening and in the case of limited, sparse external datasets.

Acknowledgments

We would like to thank personnel in Iranmehr Hospital for their help during data collection and Dr. Abbasi for his generous help and support. Thanks to the University of Technology Sydney (UTS) for the International Research Scholarship and the UTS President's Scholarship for the PhD study.

References

[1] M.L. Ranney, V. Griffeth, A.K. Jha, Critical supply shortages—the need for ventilators and personal protective equipment during the Covid-19 pandemic, N. Engl. J. Med. 382 (18) (2020) e41, https://doi.org/10.1056/NEJMp2006141.
[2] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study, Lancet 395 (10223) (2020) 507–513, https://doi.org/10.1016/S0140-6736(20)30211-7.
[3] U. COVID, About Variants of the Virus that Causes COVID-19 [Online]. Available: https://stacks.cdc.gov/view/cdc/104698, 2021.
[4] J. Lopez Bernal, N. Andrews, C. Gower, E. Gallagher, R. Simmons, S. Thelwall, J. Stowe, E. Tessier, N. Groves, G. Dabrera, Effectiveness of covid-19 vaccines against the B. 1.617. 2 (Delta) variant, N. Engl. J. Med. (2021), https://doi.org/10.1056/NEJMoa2108891.
[5] M. Kandeel, M.E.M. Mohamed, H.M. Abd El-Lateef, K.N. Venugopala, H.S. El-Beltagi, Omicron variant genome evolution and phylogenetics, J. Med. Virol. (2021), https://doi.org/10.1002/jmv.27515.
[6] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, R. Agha, World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19), Int. J. Surg. 76 (2020) 71–76, https://doi.org/10.1016/j.ijsu.2020.02.034.
[7] C.-C. Lai, Y.H. Liu, C.-Y. Wang, Y.-H. Wang, S.-C. Hsueh, M.-Y. Yen, W.-C. Ko, P.-R. Hsueh, Asymptomatic carrier state, acute respiratory disease, and pneumonia due to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): facts and myths, J. Microbiol. Immunol. Infect. 53 (3) (2020) 404–412, https://doi.org/10.1016/j.jmii.2020.02.012.
[8] Z. Xu, L. Shi, Y. Wang, J. Zhang, L. Huang, C. Zhang, S. Liu, P. Zhao, H. Liu, L. Zhu, Pathological findings of COVID-19 associated with acute respiratory distress syndrome, Lancet Respir. Med. 8 (4) (2020) 420–422, https://doi.org/10.1016/S2213-2600(20)30076.
[9] M. Shen, Y. Zhou, J. Ye, A.A.A. Al-Maskri, Y. Kang, S. Zeng, S. Cai, Recent advances and perspectives of nucleic acid detection for coronavirus, J. Pharm. Anal. 10 (2) (2020) 97–101, https://doi.org/10.1016/j.jpha.2020.02.010.
[10] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest CT for COVID-19: comparison to RT-PCR, Radiology 296 (2) (2020) E115–E117, https://doi.org/10.1148/radiol.2020200432.
[11] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, Viral pneumonia screening on chest X-ray images using confidence-aware anomaly detection, arXiv preprint arXiv:2003.12338 (2020).
[12] W.C. Serena Low, J.H. Chuah, C.A.T. Tee, S. Anis, M.A. Shoaib, A. Faisal, A. Khalil, K.W. Lai, An overview of deep learning techniques on chest X-ray and CT scan identification of COVID-19, Comput. Math. Methods Med. 2021 (2021), https://doi.org/10.1155/2021/5528144.
[13] A. Ulhaq, J. Born, A. Khan, D.P.S. Gomes, S. Chakraborty, M. Paul, Covid-19 control by computer vision approaches: a survey, IEEE Access 8 (2020) 179437–179456, https://doi.org/10.1109/ACCESS.2020.3027685.
[14] H. Asefi, A. Safaie, The role of chest CT scan in diagnosis of COVID-19, Front. Emerg. Med. 4 (2s) (2020) e64, https://doi.org/10.22114/ajem.v4i2s.451, e64.
[15] S. Ahuja, B.K. Panigrahi, N. Dey, V. Rajinikanth, T.K. Gandhi, Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices, Appl. Intell. 51 (1) (2021) 571–585, https://doi.org/10.1007/s10489-020-01826-w.
[16] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E.F. Fang, W. Menpes-Smith, J. Xia, Weakly supervised deep learning for covid-19 infection detection and classification from ct images, IEEE Access 8 (2020) 118869–118883, https://doi.org/10.1109/ACCESS.2020.3005510.
[17] Y. Huang, W. Cheng, N. Zhao, H. Qu, J. Tian, CT screening for early diagnosis of SARS-CoV-2 infection, Lancet Infect. Dis. 20 (9) (2020) 1010–1011, https://doi.org/10.1016/S1473-3099(20)30241-3.
[18] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, J. Liu, Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: relationship to negative RT-PCR testing, Radiology 296 (2) (2020) E41–E45, https://doi.org/10.1148/radiol.2020200343.
[19] A. Shoeibi, M. Khodatars, R. Alizadehsani, N. Ghassemi, M. Jafari, P. Moridian, A. Khadem, D. Sadeghi, S. Hussain, A. Zare, Automated Detection and Forecasting of Covid-19 Using Deep Learning Techniques: A Review, 2020 *arXiv preprint arXiv: 2007.10785*.
[20] D.L. Rubin, Artificial intelligence in imaging: the radiologist's role, J. Am. Coll. Radiol. 16 (9) (2019) 1309–1317, https://doi.org/10.1016/j.jacr.2019.05.036.
[21] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron. Mark. (2021) 1–11, https://doi.org/10.1007/s12525-021-00475-2.
[22] N. Bhatt, N. Bhatt, P. Prajapati, Deep learning: a new perspective, Int. J. Eng. Technol. Manag. Appl. Sci. (IJLTEMAS) 6 (6) (2017) 136–140.
[23] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (1) (2021) 1–74, https://doi.org/10.1186/s40537-021-00444-8.
[24] S.A. Harmon, T.H. Sanford, S. Xu, E.B. Turkbey, H. Roth, Z. Xu, D. Yang, A. Myronenko, V. Anderson, A. Amalou, Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets, Nat. Commun. 11 (1) (2020) 1–7, https://doi.org/10.1038/s41467-020-17971-2.
[25] A. Amyar, R. Modzelewski, H. Li, S. Ruan, Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: classification and segmentation, Comput. Biol. Med. 126 (2020) 104037, https://doi.org/10.1016/j.compbiomed.2020.104037.

[26] A.A. Ardakani, A.R. Kanafi, U.R. Acharya, N. Khadem, A. Mohammadi, Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks, Comput. Biol. Med. 121 (2020) 103795, https://doi.org/10.1016/j.compbiomed.2020.103795.
[27] A. Borakati, A. Perera, J. Johnson, T. Sood, Diagnostic accuracy of X-ray versus CT in COVID-19: a propensity-matched database study, BMJ Open 10 (11) (2020) e042946, https://doi.org/10.1136/bmjopen-2020-042946.
[28] M. Z. Che Azemin, R. Hassan, M. I. Mohd Tamrin, and M. A. Md Ali, "COVID-19 deep learning prediction model using publicly available radiologist-adjudicated chest X-ray images as training data: preliminary findings," Int. J. Biomed. Imag., vol. 2020, 2020, doi: 10.1155/2020/8828855.
[29] X. He, S. Wang, X. Chu, S. Shi, J. Tang, X. Liu, C. Yan, J. Zhang, G. Ding, Automated model Design and benchmarking of 3D deep learning models for COVID-19 detection with chest CT scans, arXiv preprint arXiv:2101.05442 (2021).
[30] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography, Cell 181 (6) (2020) 1423–1433, https://doi.org/10.1016/j.cell.2020.04.045, e11.
[31] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19, IEEE Rev. Biomed. Eng. 14 (2020) 4–15, https://doi.org/10.1109/RBME.2020.2987975.
[32] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT, Radiology (2020), https://doi.org/10.1148/radiol.2020200905.
[33] T. Javaheri, M. Homayounfar, Z. Amoozgar, R. Reiazi, F. Homayounieh, E. Abbas, A. Laali, A.R. Radmard, M.H. Gharib, S.A.J. Mousavi, CovidCTNet: an open-source deep learning approach to diagnose covid-19 using small cohort of CT images, NPJ Digit. Med. 4 (1) (2021) 1–10, https://doi.org/10.1038/s41746-021-00399-3.
[34] H. Ko, H. Chung, W.S. Kang, K.W. Kim, Y. Shin, S.J. Kang, J.H. Lee, Y.J. Kim, N. Y. Kim, H. Jung, COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation, J. Med. Internet Res. 22 (6) (2020) e19569, https://doi.org/10.2196/19569.
[35] V. Shah, R. Keniya, A. Shridharani, M. Punjabi, J. Shah, N. Mehendale, Diagnosis of COVID-19 using CT scan images and deep learning techniques, Emerg. Radiol. 28 (3) (2021) 497–505, https://doi.org/10.1007/s10140-020-01886-y.
[36] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, M. Kaur, Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning, J. Biomol. Struct. Dyn. (2020) 1–8, https://doi.org/10.1080/07391102.2020.1788642.
[37] Z. Han, B. Wei, Y. Hong, T. Li, J. Cong, X. Zhu, H. Wei, W. Zhang, Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning, IEEE Trans. Med. Imag. 39 (8) (2020) 2584–2594, https://doi.org/10.1109/TMI.2020.2996256.
[38] E.J. Topol, Is my cough COVID-19? Lancet 396 (10266) (2020) 1874, https://doi.org/10.1016/S0140-6736(20)32589-7.
[39] K.B. Ahmed, G.M. Goldgof, R. Paul, D.B. Goldgof, L.O. Hall, Discovery of a generalization gap of convolutional neural networks on COVID-19 X-rays classification, IEEE Access 9 (2021) 72970–72979, https://doi.org/10.1109/ACCESS.2021.3079716.
[40] P. Silva, E. Luz, G. Silva, G. Moreira, R. Silva, D. Lucio, D. Menotti, COVID-19 detection in CT images with deep learning: a voting-based scheme and cross-datasets analysis, Inform. Med. Unlocked 20 (2020) 100427, https://doi.org/10.1016/j.imu.2020.100427.
[41] A.A. Ardakani, R.M. Kwee, M. Mirza-Aghazadeh-Attari, H.M. Castro, T.Y. Kuzan, K. M. Altintoprak, G. Besutti, F. Monelli, F. Faeghi, U.R. Acharya, A practical artificial intelligence system to diagnose COVID-19 using computed tomography: a multinational external validation study, Pattern Recogn. Lett. 152 (2021) 42–49, https://doi.org/10.1016/j.patrec.2021.09.012.
[42] A. Gudigar, U. Raghavendra, S. Nayak, C.P. Ooi, W.Y. Chan, M.R. Gangavarapu, C. Dharmik, J. Samanth, N.A. Kadri, K. Hasikin, Role of artificial intelligence in COVID-19 detection, Sensors 21 (23) (2021) 8045, https://doi.org/10.3390/s21238045.
[43] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, C. Zheng, A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT, IEEE Trans. Med. Imag. 39 (8) (2020) 2615–2625, https://doi.org/10.1109/TMI.2020.2995965.
[44] P.R. Bassi, R. Attux, COVID-19 detection using chest X-rays: is lung segmentation important for generalization? arXiv preprint arXiv:2104.06176 (2021).
[45] G. Maguolo, L. Nanni, A critic evaluation of methods for covid-19 automatic detection from x-ray images, Inf. Fusion 76 (2021) 1–7, https://doi.org/10.1016/j.inffus.2021.04.008.
[46] D. Nguyen, F. Kay, J. Tan, Y. Yan, Y.S. Ng, P. Iyengar, R. Peshock, S. Jiang, Deep learning-based COVID-19 pneumonia classification using chest CT images: model generalizability, arXiv preprint arXiv:2102.09616 (2021).
[47] M.J. Horry, S. Chakraborty, B. Pradhan, M. Fallahpoor, H. Chegeni, M. Paul, Factors determining generalization in deep learning models for scoring COVID-CT images, Math. Biosci. Eng. 18 (6) (2021) 9264–9293, https://doi.org/10.3934/mbe.2021456.
[48] M. Rahimzadeh, A. Attar, S.M. Sakhaei, A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset, Biomed. Signal Process Control 68 (2021) 102588, https://doi.org/10.1016/j.bspc.2021.102588.
[49] S. Morozov, A. Andreichenko, I. Blokhin, P. Gelezhe, A. Gonchar, A. Nikolaev, N. Pavlov, V. Chernina, V. Gombolevsky, Mosmeddata: data set of 1110 chest ct scans performed during the covid-19 epidemic, Digital Diagnostics 1 (1) (2020) 49–59, https://doi.org/10.17816/DD46826.
[50] J. Li, G. Zhao, Y. Tao, P. Zhai, H. Chen, H. He, T. Cai, Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19, Pattern Recogn. 114 (2021) 107848, https://doi.org/10.1016/j.patcog.2021.107848.
[51] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: prospective predictions are the future," arXiv preprint arXiv: 2006.11988, 2020.
[52] COVID-19 X rays. https://www.kaggle.com/andrewmvd/convid19-x-rays, 2020.
[53] Mendeley data. https://data.mendeley.com. (Accessed March 2020).
[54] L. Aversano, M.L. Bernardi, M. Cimitile, R. Pecori, Deep neural networks ensemble to detect COVID-19 from CT scans, Pattern Recogn. 120 (2021) 108135, https://doi.org/10.1016/j.patcog.2021.108135.
[55] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[56] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1251–1258.
[57] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[58] W. El-Shafai, F. Abd El-Samie, Extensive COVID-19 X-Ray and CT Chest Images Dataset, vol. 3, Mendeley Data, 2020.
[59] R. Kumar, A.A. Khan, J. Kumar, A. Zakria, N.A. Golilarz, S. Zhang, Y. Ting, C. Zheng, W. Wang, Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging, IEEE Sensor. J. (2021), https://doi.org/10.1109/JSEN.2021.3076767.
[60] P. Angelov, E. Almeida Soares, SARS-CoV-2 CT-scan dataset: a large dataset of real patients CT scans for SARS-CoV-2 identification, medRxiv (2020), https://doi.org/10.1101/2020.04.24.20078584.
[61] H. Shahin. COVID-19 low-dose and ultra-low-dose CT SCANS, doi: 10.21227/sed8-6r15.
[62] W. Xiaofei. 3DLSC-COVID, doi: https://dx.doi.org/10.21227/mxb3-7j48.
[63] K.S. Mader, DSB lung segmentation algorithm. https://www.kaggle.com/kmader/dsb-lung-segmentation-algorithm, 2017.
[64] G. Zuidhof, Full preprocessing tutorial, in: https://www.kaggle.com/gzuidhof/full-preprocessing-tutorial, 2017.
[65] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, G. Langs, Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem, Eur. Radiol. Exp. 4 (1) (2020) 1–13, https://doi.org/10.1186/s41747-020-00173-2.
[66] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Tensorflow: a system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.
[67] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[68] Y. Seol, Y. Kim, Y. Kim, Y. Cheon, K. Kim, A study on 3D deep learning-based automatic diagnosis of nasal fractures, Sensors 22 (2) (2022) 506, https://doi.org/10.3390/s22020506.
[69] S. Chaudhary, S. Sadbhawna, V. Jakhetiya, B.N. Subudhi, U. Baid, S.C. Guntuku, Detecting covid-19 and community acquired pneumonia using chest ct scan images with deep learning," ICASSP IEEE Int. Conf. Acoust, Speech and Signal Processing (ICASSP) (2021) 8583–8587, https://doi.org/10.1109/ICASSP39728.2021.9414007. IEEE.
[70] S.-Y. Lee, H. Kang, J.-H. Jeong, D.-y. Kang, Performance evaluation in [18F] Florbetaben brain PET images classification using 3D Convolutional Neural Network, PLoS One 16 (10) (2021) e0258214, https://doi.org/10.1371/journal.pone.0258214.
[71] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy, Radiology 296 (2) (2020) E65–E71, https://doi.org/10.1148/radiol.2020200905.
[72] S. Sharma, S. Sharma, Activation functions in neural networks, Data Sci. 6 (12) (2017) 310–316.
[73] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6546–6555.
[74] A. Bhandari, Everything you should know about confusion matrix for machine learning. https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/. (Accessed April 2020).
[75] J. Fan, S. Upadhye, A. Worster, Understanding receiver operating characteristic (ROC) curves, Can. J. Emerg. Med. 8 (1) (2006) 19–20, https://doi.org/10.1017/S1481803500013336.
[76] M.J. Willemink, W.A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R.M. Summers, D.L. Rubin, M.P. Lungren, Preparing medical imaging data for machine learning, Radiology 295 (1) (2020) 4–15, https://doi.org/10.1148/radiol.2020192224.
[77] M.J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, N. Shukla, COVID-19 detection through transfer learning using multimodal imaging data, IEEE Access 8 (2020) 149808–149824, https://doi.org/10.1109/ACCESS.2020.3016780.
[78] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conf. Comput. Vis. Pattern Recognit, Ieee, 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.