Article

LGM: Novel large empirical study of deep transfer learning for COVID-19 classification based on CT and X-ray images

Mansour Almutaani[1], Turki Turki[1,*], and Y-h. Taguchi[2]

[1]Department of Computer Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Physics, Chuo University, Tokyo 112-8551, Japan
* Correspondence: tturki@kau.edu.sa (T.T.)

Abstract: Early and accurate prediction of COVID-19 based on medical images can speed up the diagnostic process and thereby mitigating the disease spread. Hence, developing AI-based models is an inevitable endeavor. The presented work, to our knowledge, is considered the largest empirical COVID-19 classification study using CT and X-ray in which we propose a novel computational framework constructing 10000 deep transfer learning (DTL) models as follows. First, we downloaded and processed 4481 CT and X-ray images pertaining to COVID-19 and non-COVID-19 patients, obtained from the Kaggle repository. Second, we provide processed images as input to four pre-trained deep learning models on more than million images from ImageNet database in which we froze the convolutional and pooling layers pertaining to the feature extraction part while unfroze and train the densely connected classifier with Adam optimizer. Third, we generate and take majority vote of 2, 3, and 4-combinations from the 4 DTL models, resulted in $\sum_{r=2}^{4}\binom{4}{r}$=11 DTL models. Then, we combine the 11 DTL models, followed by consecutively generating and taking the majority vote of $\sum_{r=2}^{11}\binom{11}{r}$= 2036 DTL models. Finally, we select 7953 DTL models from $\binom{2036}{2}$. Experimental results on the whole datasets using five-fold cross-validation demonstrate that best generated DTL model, named HC, achieving the best AUC of 0.909 when applied to the CT dataset while ConvNeXt yielded a higher marginal AUC of 0.933 compared to 0.93 for HX when considering the X-ray dataset. These promising results set the foundation for promoting large generation of models (LGM) in AI.

Keywords: COVID-19; medical imaging; classification; deep transfer learning; model generation; advanced AI concepts

1. Introduction and Related Work

Coronavirus disease 2019 (COVID-19) patients experience different symptoms such as loss of smell and/or taste, fever, headache, and shortness of breath, to name a few [1]. Human-to-human transmission can happen via (1) sneeze droplets carrying the SARCoV-2; (2) a direct contact with infected individual; and (3) touching a virus-containing surface followed by touching eyes or mouth [2]. To lower the risk of COVID-19 transmission, the center of control disease and prevention suggested prevention strategies including wearing masks, social distancing, and washing hands with soap for at least 20 seconds (see Figure 1) [3, 4]. Early detection of COVID-19 and thereby isolating the infected individuals is considered the most effective prevention strategy [5, 6]. Hence, researchers have proposed various computational methods based on deep learning to detect COVID-19 using various image sources. Ozturk et al. [7] proposed a deep learning approach (DarkCovidNet) composed of interleaved convolutional and maxpooling layers for feature extraction and densely connected classifier for the binary and multiclass COVID-19 classification tasks. DarkCovidNet and other nine deep learning (DL) models were applied to 635 (and 1125) X-ray test images from GitHub pertaining to binary (and multiclass)

COVID-19 classification. DarkCovidNet had the highest accuracy of 0.9808 (and 0.8702) for the binary (and multiclass) COVID-19 classification.

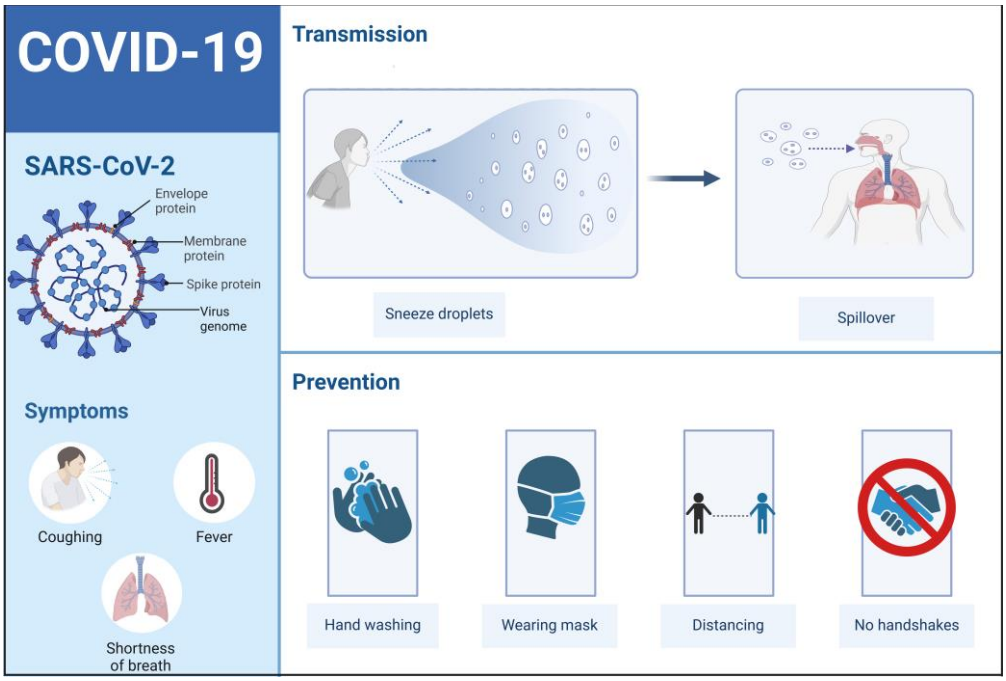

Figure 1. COVID-19 disease symptoms as well as transmission and prevention ways for the SARS-CoV-2 virus. Figure created with BioRender.com.

To address the COVID-19 classification task using X-ray images, Apostolopoulos et al. [8] employed five pre-trained DL models including MobileNetV2, VGG19, Inception, Xception, and InceptionResNetV2 applied to two testing data obtained from GitHub and Kaggle. In the first testing data, it had 1428 X-ray images for the binary class task pertaining to COVID-19 and non-COVID-19 images while the multiclass task was to discriminate between COVID-19, Pneumonia, and Normal images. VGG19 achieved the highest accuracy of accuracy of 0.9875 and 0.9348 for the binary and multiclass classification tasks, respectively. For the second testing data of 1442 test images, MobileNetv2 generated the highest accuracy of 0.9678 and 0.9472 for the binary and multiclass classification tasks, respectively. Ismael et al. [9] aimed to discriminate between Normal and COVID-19 X-ray images as follows. They employed five pre-trained DL models, namely VGG16, VGG19, ResNet18, ResNet50, and ResNet101 for feature extraction in which each was coupled with SVM to train and perform prediction. Data were downloaded COVID-19 X-ray images from GitHub, Kaggle, and Radiology Assistant. Results demonstrated that when ResNet50 applied 95 test images for feature extraction followed by SVM for classification, resulted in the highest accuracy of 0.947.

Ravi et al. [10] proposed a deep and machine learning approach to discriminate between COVID-19, non-COVID-19 X-ray and CT images. The proposed approach consisted of pre-trained EfficientNet for feature extraction followed by logistic regression as a machine learning (ML) model for classification. Compared to 26 DL models, EfficientNet+ML when applied to 2417 CT images, yielded the highest accuracy of 0.9946. Moreover, EfficientNet+ML generated the highest accuracy of 0.9948 when applied to 2864 COVID-19 X-ray images. Asif et al. [11] proposed a shallow CNN trained on COVID-19 X-ray images from Kaggle. The proposed shallow CNN was compared against 5 pre-trained DL models (i.e., NasNet, MobileNet, Xception, InceptionV3, and DenseNet201) in which all were applied to 636 COVID-19 X-ray test images. Results demonstrated that the proposed shallow CNN achieved the highest accuracy of 0.9968. Gupta et al. [12] proposed a deep learning model (DLM) for the task of COVID-19 classification based on

COVID-19 CT images from Kaggle. The proposed DLM is composed of a series of convolutional and maxpooling layers for feature extraction, followed by a densely connected classifier to discriminate between COVID-10 and non-COVID-19 CT images. Compared to two pre-trained models (DarkNet19 and MobileNetV2), results demonstrated that the proposed DLM when applied to 2481 yielded the highest accuracy of 0.9891.

Ayalew et al. [13] proposed an approach based on CNN for feature extraction coupled with SVM for COVID-19 classification. They downloaded COVID-19 X-ray images from GitHub. Then, applying CNN+SVM to 1200 X-ray images pertaining to COVID-19 and Normal cases. Results demonstrated that the proposed approach achieved the highest accuracy of 0.991. Constantinou et al. [14] employed 5 pre-trained DL models including InceptionV3, DenseNet169, ResNet101, DenseNet121 and ResNet50 for the multiclass COVID-19 classification task. They downloaded COVID-19 X-ray images from Kaggle. Results on 6788 test images demonstrated that ResNet101 achieved the highest accuracy of 0.96 when compared to the other DL models. Turki et al. [15] proposed a deep learning approach for COVID-19 classification based on transmission electron microscopy (TEM) images, obtained from the electron microscopy public image archive (EMPIAR). They downloaded and processed 286 images concerning healthy and SARS-CoV-2 infected alveolar type II cells. Results using five-fold cross-validation demonstrated the superiority of DenseNet121 under the transfer learning setting (named TFeDenseNet201), achieving the highest accuracy of 0.993. In Table 1, we provide a comparison of our proposed work against existing works for COVID-19 classification.

Although these existing studies aimed to detect COVID-19 based on medical images, the novelty in our study can be mainly ascribed to the following summarized contributions:

(1) Unlike existing studies for detecting COVID-19, to our knowledge, this is the first study to investigate large generation of AI models in which we propose a novel computational framework, resulted in constructing thousands of prediction models for the task of COVID-19 classification based on CT and X-ray images.

(2) For the COVID-19 classification task in this work, we employed four pre-trained DL models, including ConvNeXt, EfficientNetV2, DenseNet121, and ResNet34. Then, we downloaded 2481 and 2000 COVID-19 images from Kaggle pertaining to CT and X-ray scanning methods, respectively. For each pre-trained model, we froze the feature extraction part composed of convolutional and pooling layers while training from scratch the densely connected layer with Adam optimizer for the binary classification, resulting in 4 deep transfer learning (DTL) models. We take majority vote of 2, 3, and 4-combinations from the 4 DTL models, resulting in a total of $\sum_{r=2}^{4}\binom{4}{r}$=11 DTL models. Next, we take 2, 3, …, and 11-combinations from the 11 DTL models, resulting in $\sum_{r=2}^{11}\binom{11}{r}$= 2036 DTL models. Finally, we choose 7953 DTL models from $\binom{2036}{2}$.

(3) According to our large empirical study of 10004 models attributed to 11+2036+7953 generated DTL models plus 4 DTL models, the generated model named HC which corresponds to composition of DTL models obtained the highest AUC of 0.909 when COVID-19 CT images is used. For the COVID-19 X-ray images, ConvNeXt and the generated model HX obtained an AUC of 0.933 and 0.930, respectively. These results demonstrate that our computational framework (1) leads to promoting never-ending AI model generation; (2) expands the model space from 4 to 10000; and (3) aids to the search of high-performance models to tackle the studied classification task.

Table 1: Comparison of our proposed work against other studies. P stands for pneumonia. VP stands for viral pneumonia. BP stands for bacterial pneumonia. CAP stands for community acquired pneumonia. CT is computed tomography. TEM is transmission electron microscopy. ACC is accuracy. AUC is area under curve.

Year	Study	Data Source	Scanning	No. Models	Best Model	Class	Test Images	Results
2020	Ozturk et al. [7]	GitHub [16]	X-ray	10	DarkCovidNet	2-class (COVID-19, No-Findings)	635	ACC (0.9808)
						3-class (COVID-19,P; No-Findings)	1125	ACC (0.8702)
2020	Apostolopoulos et al. [8]	GitHub [16], Kaggle [17, 18]	X-ray	5	VGG19	2-class (COVID-19, non-COVID-19)	1428	ACC (0.9875)
						3-class (COVID-19, P, Normal)		ACC (0.9348)
					MobileNetv2	2-class (COVID-19, non-COVID-19)	1442	ACC (0.9678)
						3-class (COVID-19, P, Normal)		ACC (0.9472)
2020	Ouchicha et al. [19]	GitHub, Kaggle	X-ray	1	CVDNet	3-class (COVID-19, P, Normal)	2905	ACC (0.9669)
2021	Ismael et al. [9]	GitHub [16] Kaggle [17, 18], Radiology Assistant [20]	X-ray	5	ResNet50 Features + SVM	2-cass (COVID-19, Normal)	95	ACC (0.947)
2021	Ahamed et al. [21]	Kaggle [22-24]	CT	1	ResNet50V2	2-cass (COVID-19, Normal)	3000	ACC (0.99990)
						3-cass (COVID-19, Normal, CAP)		ACC (0.99012)
			Xray			2-cass (COVID-19, VP)	4593	ACC (0.98954)
						3-cass (COVID-19, BP, Normal)		ACC (0.97242)
						4-cass (COVID-19, BP, VP, Normal)		ACC (0.96452)
2021	Jain et al. [25]	Kaggle [26]	X-ray	3	Xception	3-class (COVID-19, P, Normal)	965	ACC (0.9797)
2022	Ravi et al. [10]	Mendeley [27]	CT	27	EfficientNet+ML	2-class (COVID-19, non-COVID-19)	2417	ACC: (0.9946)
			X-ray				2864	ACC (0.9948)
2022	Furtado et al. [28]	Multiple [29-35] + Local	CT	1	Cimatec-CovNet-19	2-class (COVID-19, non-COVID-19)	414	AUC (0.95)
2022	Asif et al. [11]	Kaggle [22, 23]	X-ray	6	Shallow CNN	2-claass (COVID-19, Normal)	636	ACC (0.9968)
2023	Gupta et al. [12]	Kaggle [36, 37]	CT	3	DLM	2-class (COVID-19, non-COVID-19)	2481	ACC (0.9891)
2023	Ayalew et al. [13]	GitHub [16, 38]	X-ray	1	CNN+SVM	2-class (COVID-19, Normal)	1200	ACC (0.991)
2023	Constantinou et al. [14]	Kaggle [39]	X-ray	5	ResNet101	3-class (COVID-19, non-COVID-19, Normal)	6788	ACC (0.96)
2024	Echtioui et al. [40]	Shared [41]	X-ray	4	DenseNet-169 + 2Dense	3-class (COVID-19, Normal, P)	231	ACC (0.9004)
					DenseNet-169 + 2Dense			
2024	Turki et al. [15]	EMPIAR [42, 43]	TEM	10	TFeDenseNet201	2-class (Infected, Control)	286	ACC (0.993)
2024	Proposed	Kaggle [36, 37]	CT	10004	HC	2-class (COVID-19, non-COVID-19)	2481	AUC (0.909)
		Kaggle [22]	X-ray	10004	ConvNeXt	2-class (COVID-19, non-COVID-19)	2000	AUC (0.933)

2. Materials and Methods

2.1. Data Preprocessing

The CT dataset (named SARS-COV-2 CT-Scan database) was downloaded from Kaggle https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset (accessed on 17 March 2024). We randomly selected and processed 2481 images in which the class label distribution is 1252 for COVID-19 and 1229 for non-COVID-19. The X-ray dataset (named COVID-19 Radiography Database) was downloaded from Kaggle at https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database (accessed on 17 March 2024). We randomly selected 2000 images in which 1000 images were categorized as COVID-19 and the other remaining 1000 images were categorized as non-COVID-19. All images were resized to 256 × 256 when addressing the COVID-19 classification task in this study.

2.2. Deep Transfer Learning

In Figure 2, we illustrate the deep transfer learning (DTL) part pertaining to our approach applied to CT images. First, we utilize four pre-trained deep learning (DL) models,

including ConvNeXt, EfficientNetV2, DenseNet121, and ResNet34. Each pre-trained DL model was induced after training on million images from ImageNet database to tackle the multiclass classification task of categorizing images into one of the 1000 predefined classes. In this study, we leave weights of the feature extraction part unchanged in which it is applied to COVID-19 images followed by flattening step and then performing training to update weights of subsequent layers in the classification part. The training phase incorporates Adam optimizer and binary cross entropy with logits loss function. Each DTL yields probabilities when performing predictions in which if the probability if greater than 0.5, then the image is classified as non-COVID-19. If the prediction value is less than or equal to 0.5, then the image is classified as COVID-19. Figure 3 illustrates the same DTL computations. The only difference is the dataset, which belongs to X-ray images instead of CT images.

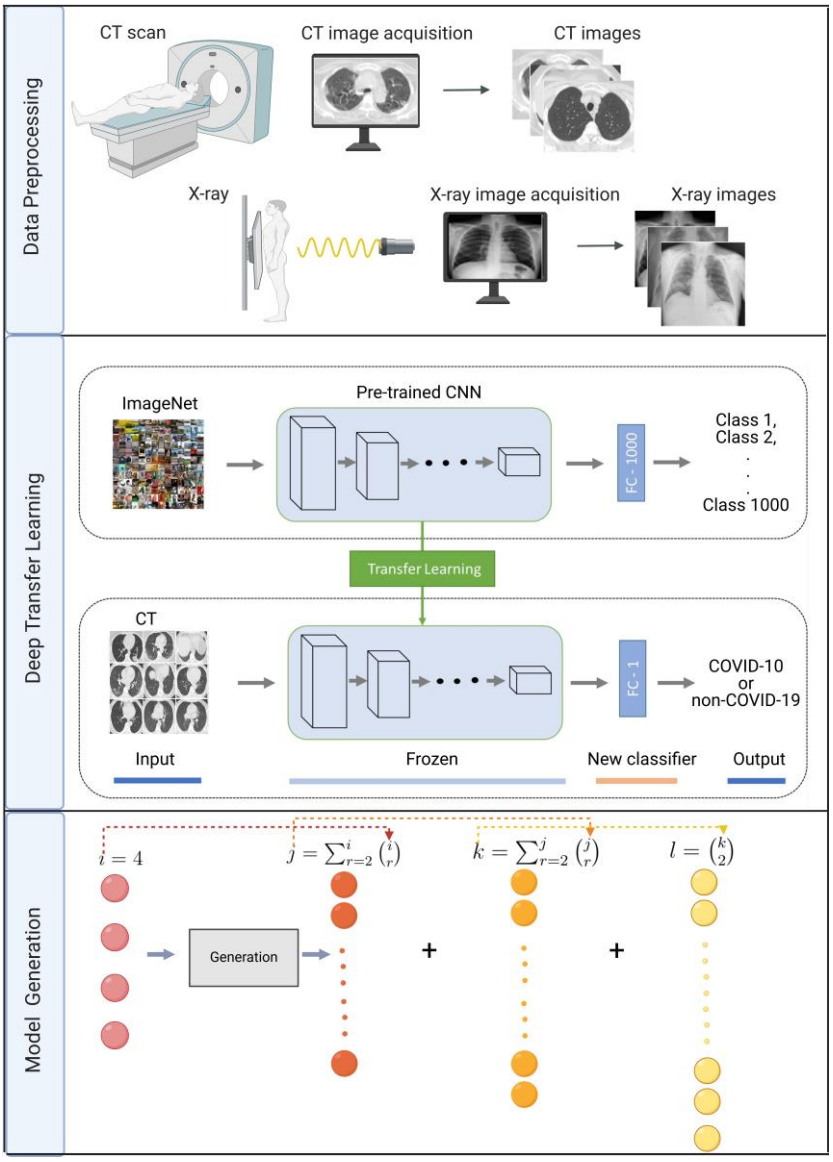

Figure 2. Flowchart showing our computational approach for COVID-19 classification. Data Preprocessing: Acquiring COVID-19 CT and X-ray images pertaining to COVID-19 and non-COVID-19 patients. Deep Transfer Learning: CT images are provided to a pre-trained CNN for feature extraction followed by training a new classifier to discriminate between COVID-19 and non-COVID-19 images. Model Generation: four models are provided as input to sequentially generate DTL models. Figure created with BioRender.com.

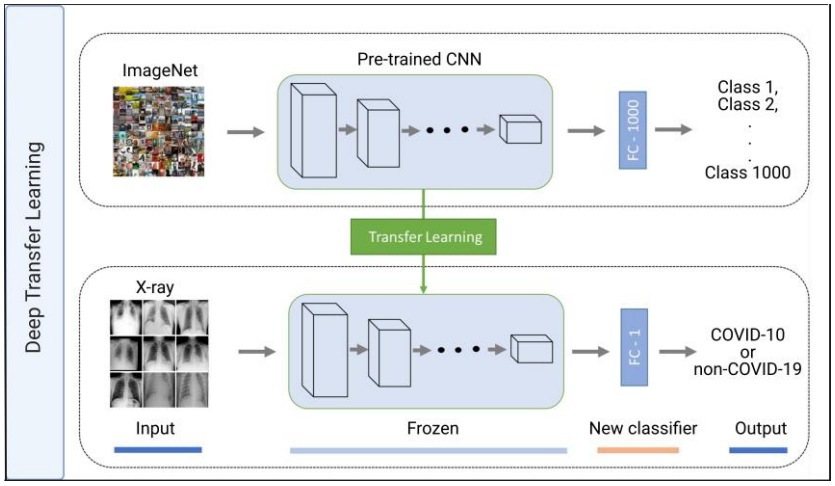

Figure 3. Deep transfer learning is performed in which a pre-trained CNN is applied for feature extraction followed by training a new classifier to discriminate between COVID-19 and non-COVID-19 images. Figure created with BioRender.com.

2.3. Model Generation

To generate a large number of DTL models performing prediction to a test example z, in our study, we start with 4 DTL models, namely $h_1(z)$, $h_2(z)$, $h_3(z)$, and $h_4(z)$ which correspond to models obtained from training ConvNext, EfficientNetV2, DenseNet121, and ResNet34, respectively. Let D1={$h_1(z)$, $h_2(z)$, $h_3(z)$, $h_4(z)$} where $hi(z): z \rightarrow$ {COVID-19, non-COVID-19} and we generate all 2, 3 and 4-combinaions from D1, resulting in S_i^1 for i=1…11 attributed to $\sum_{r=2}^{4}\binom{4}{r}$. It can be evident that the size of a given S_i^1 is upper bounded by 4 and lower bounded by 2 (i.e., $2 \leq |S_i^1| \leq 4$). Assuming that $h_i(z) \in S_i^1$ and we then perform majority prediction as shown in Equation 1 in which $I()$ is an indicator function returning 1 if true and 0 otherwise. Let D2={$h1_1(z)$, $h1_2(z)$,…, $h1_{11}(z)$}. Now, we generate all 2, 3, 4, 5, 6, 7, 8, 9, 10 , and 11-combinations from D2, resulting in S_i^2 for i = 1…2036 ascribed to $\sum_{r=2}^{11}\binom{11}{r}$. It can be seen that the size of a given S_i^2 is lower bounded by 2 and upper bounded by 11 (i.e., $2 \leq |S_i^2| \leq 11$). Let $h1_i(z) \in S_i^2$ and prediction is performed using Equation 2. We set D3 = {$h2_1(z)$, $h2_2(z)$,…, $h2_{2036}(z)$}. We only select 7953 from generated 2-combinations from D3. Therefore, we have S_i^3 for i = 1…7953 from $\binom{2036}{2}$ models. For $h2_i(z) \in S_i^3$, a prediction is performed using Equation 3. It can be shown that we had $|\{S_i^1\}_{i=1}^{11}| + |\{S_i^2\}_{i=1}^{2036}| + 7953 = 10000$ models.

$$h1_i(z) = \arg\max_y \sum_i I(h_i(z) = y) \tag{1}$$

$$h2_i(z) = \arg\max_y \sum_i I(h1_i(z) = y) \tag{2}$$

$$h3_i(z) = \arg\max_y \sum_i I(h2_i(z) = y) \tag{3}$$

3. Results

3.1. Classification Methodology

In our study, we adapted four pre-trained DL models including ConvNeXt, Efficient-NetV2, DenseNet121, and ResNet34. All these four pre-trained DL models were trained on more than a million images from the ImageNet database in which the number of

different class labels is equal to 1000. We froze all layers in feature extraction related to convolutional and pooling layers to extract features from the COVID-19 images generated via imaging techniques while updating weights in the subsequent layers pertaining to the classification part during the training phase. For all DTL models, we utilized the Adam optimizer with binary crossentropy with logits loss function. In terms of the optimization parameters when building models during the training phase, we used 32 for the batch size, 0.001 for the learning rate, and 10 for the number of epochs. When performing prediction to an unseen image, if the predicted value is greater than 0.5, then we designate the image as non-COVID-19. Otherwise, we designate the image as COVID-19. For all DTL including the 10000 constructed models in each imaging technique, we employed the following five performance measures to evaluate the studied models [44, 45]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Sensitivty = \frac{TP}{TP + FN} \tag{7}$$

$$Sepcificity = \frac{TN}{TN + FP} \tag{8}$$

where *TP* stands for true positive, referring to the number of non-COVID-19 images that were correctly classified as non-COVID-19. *FN* stands for false negative, referring to the number of non-COVID-19 images that were incorrectly classified as COVID-19. *TN* stands for true negative, referring to the number of COVID-19 images that were correctly classified as COVID-19. *FP* stands for false positive, referring to the number of COVID-19 images that were incorrectly classified as non-COVID-19.

To assess the performance on two studied datasets, we employed five-fold cross-validation as follows. For each dataset, we randomly assign images to five-folds. Then, we take examples in one fold for testing while images in the remaining four folds are used to train DTL models. Then, we perform prediction to images in the testing fold. We report five performance results pertaining to each DTL model. The same process in the first run is done for the four left runs in five-fold cross-validation in which we record testing results in those four runs. Finally, we take the average testing results of each performance metric and record results corresponding to five-fold cross-validation results.

3.2. Implementation Details

When running the experiments in this study using NVIDIA Tesla P100 GPU, the deep learning models are implemented in Python (version 3.10.13) [46] based on PyTorch library (version 2.1.2) [47, 48] and torchvision library (version 0.16.2) in which the installed CUDA version is 12.1 [49]. Data processing is achieved through various libraries including NumPy (version 1.26.4) and pandas (version 2.2.2) [50, 51]. The performance evaluation is performed through the sklearn library (version 1.2.2) [52].

3.3. Classification Results

3.3.1. Training Results

Figure 4a and b display the average training accuracy and loss over 10 epochs for the 4 DTL models during the running of five-fold cross-validation, obtained when applied to COVID-19 CT and X-ray datasets, respectively. The accuracy and loss moving in the

upward and downward directions, respectively, show the capability of models to learn and thereby will be able to classify unseen COVID-19 images during the testing phase. When considering the CT dataset (Figure 4a), ConvNeXt achieved the highest and lowest average accuracy and loss of 0.86 and 0.26, respectively, followed by DenseNet121, yielding an average accuracy (and loss) of 0.84 (and 0.30). EfficientNetV2 had the lowest and highest average accuracy and loss of 0.81 and 0.34, respectively. For the X-ray dataset as shown in Figure 4b, ConvNeXt achieved the highest average accuracy of 0.89 while attaining the lowest average loss of 0.20. DenseNet121 achieved an average accuracy and loss of 0.87 and 0.26, respectively. EfficientNetV2 yielded an average accuracy and loss of 0.84 and 0.31, respectively.

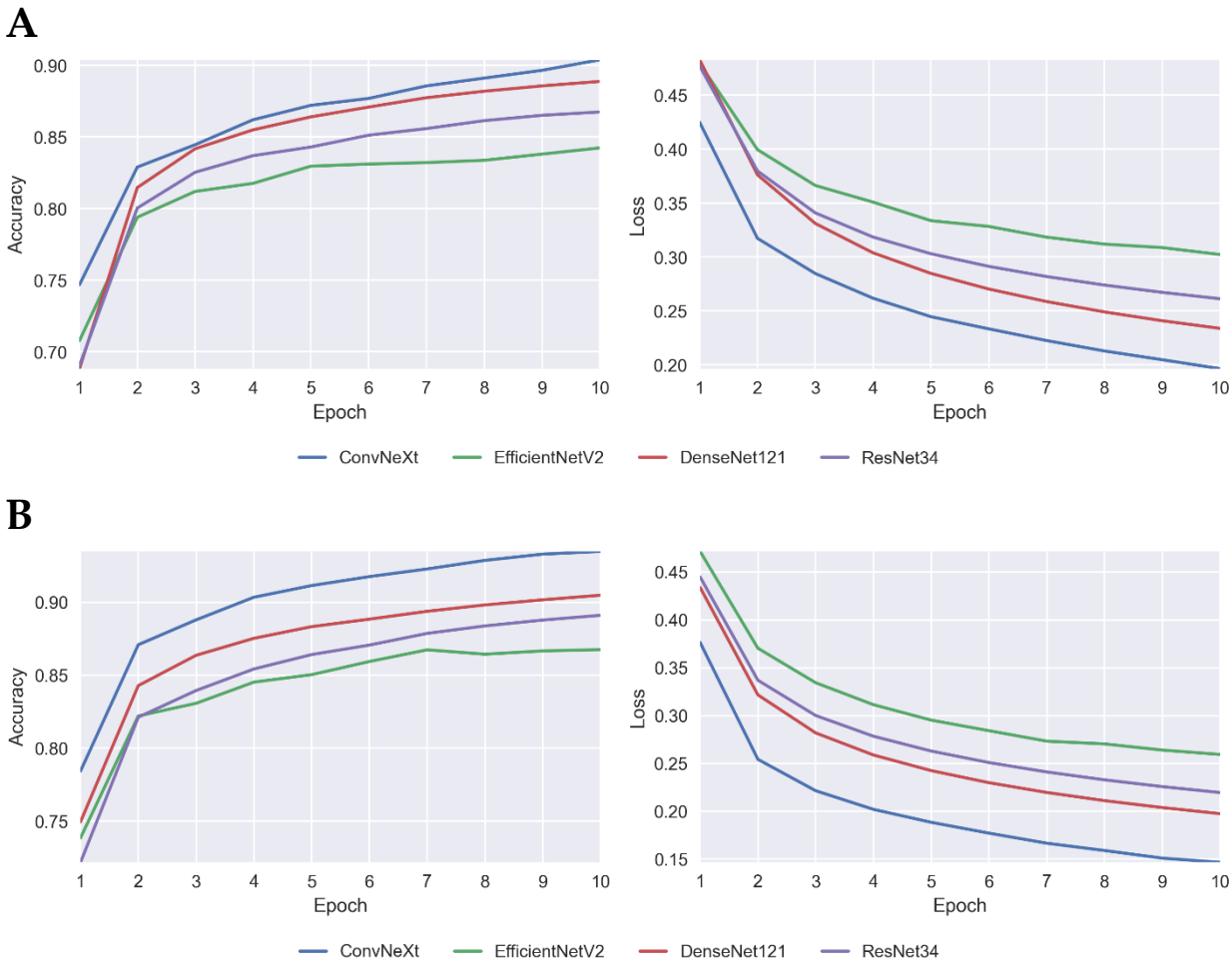

Figure 4. Average accuracy and loss for ten epochs during the running of five-fold cross-validation pertaining to (a) CT and (b) X-ray datasets.

In Figure 5a and b pertaining to CT and X-ray datasets, respectively, we illustrate the total running time for the model induction process of 4 DTL models during running five-fold cross-validation. In terms of CT dataset as shown in Figure 5a, ResNet34 is the fastest. Specifically, ResNet34 is 1.14 × faster than ConvNeXt. Moreover, the running time difference between DenseNet121 and EfficientNetV2 is marginal in which the former and latter are 1.067 and 1.062 × faster than ConvNeXt. For the X-ray dataset as illustrated in Figure 5b, ResNet34 is still the fastest method, which is 1.37 × faster than ConvNeXt. The second and third fastest methods are DeneseNet121 and EfficientNetV2, respectively. Both DeneseNet121 and EfficientNetV2 resulted in marginal running time differences in which

the former and latter are 1.242 and 1.210 × faster than ConvNeXt. These results demonstrate that although ConvNeXt yielded the highest performance results, it was the slowest among all models.

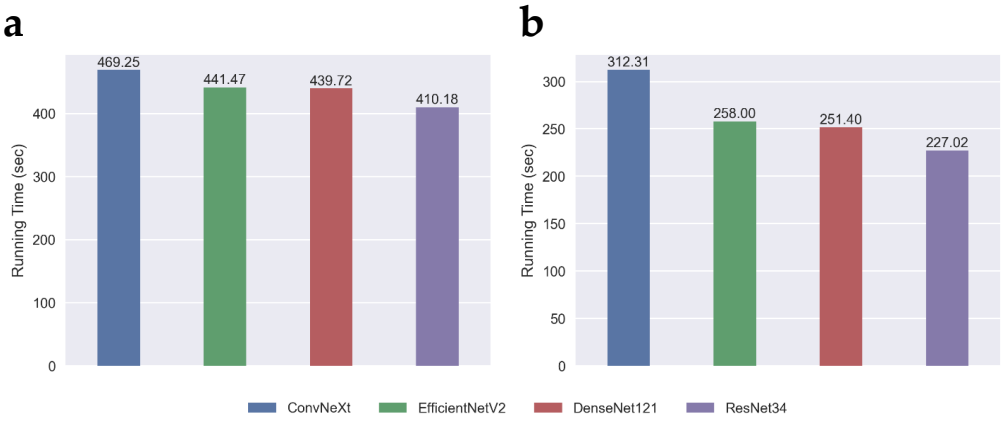

Figure 5. Total Running time for the four studied DTL models during the training phase in five-fold cross-validation pertaining to (a) CT and (b) X-ray datasets.

3.3.2. Testing Results

Boxplots in Figure 6a and b illustrate the generalization (testing) results of 4 DTL models utilizing five-fold cross-validation pertaining to CT and X-ray datasets, respectively. Regarding Figure 6a and b, the median results are indicated via the horizontal lines that are crossing each box. In terms of the CT dataset in Figure 6a, ConvNeXt yielded the best results, achieving the highest median of 0.901, 0.903, 0.874, 0.935, and 0.872 according to accuracy, F1, precision, sensitivity, and specificity, respectively. The second-best performing model is DenseNet121, achieving median accuracy 0.885, median F1 0.889, median precision of 0.857, median sensitivity of 0.930, and median specificity of 0.844. ResNet34 is the third-best performing model, generating median accuracy of 0.853, median F1 of 0.863, median precision of 0.832, median sensitivity of 0.894, and median specificity of 0.828. EfficientNetV2 is the worst performing method yielding the lowest median accuracy of 0.842, the lowest median F1 of 0.851, the lowest median precision of 0.802, the median sensitivity of 0.910, and the lowest median specificity of 0.780. Regarding the X-ray dataset in Figure 6b, ConvNeXt obtained the highest median accuracy of 0.930, the highest median F1 of 0.928, the highest median precision of 0.931, the highest median sensitivity of 0.930, and highest median specificity of 0.930. The second-best performing method is DenseNet121, which had median accuracy of 0.917, median F1 of 0.912, median precision of 0.910, median sensitivity of 0.925, and median specificity of 0.915. The third-best method is ResNet34, yielding median accuracy, F1, precision, sensitivity, and specificity of 0.885, 0.887, 0.896, 0.900, and 0.895, respectively. EfficientNetV2 was the worst-performing model, obtained a median accuracy, F1, precision, sensitivity, and specificity of 0.865, 0.862, 0.848, 0.915, and 0.835, respectively.

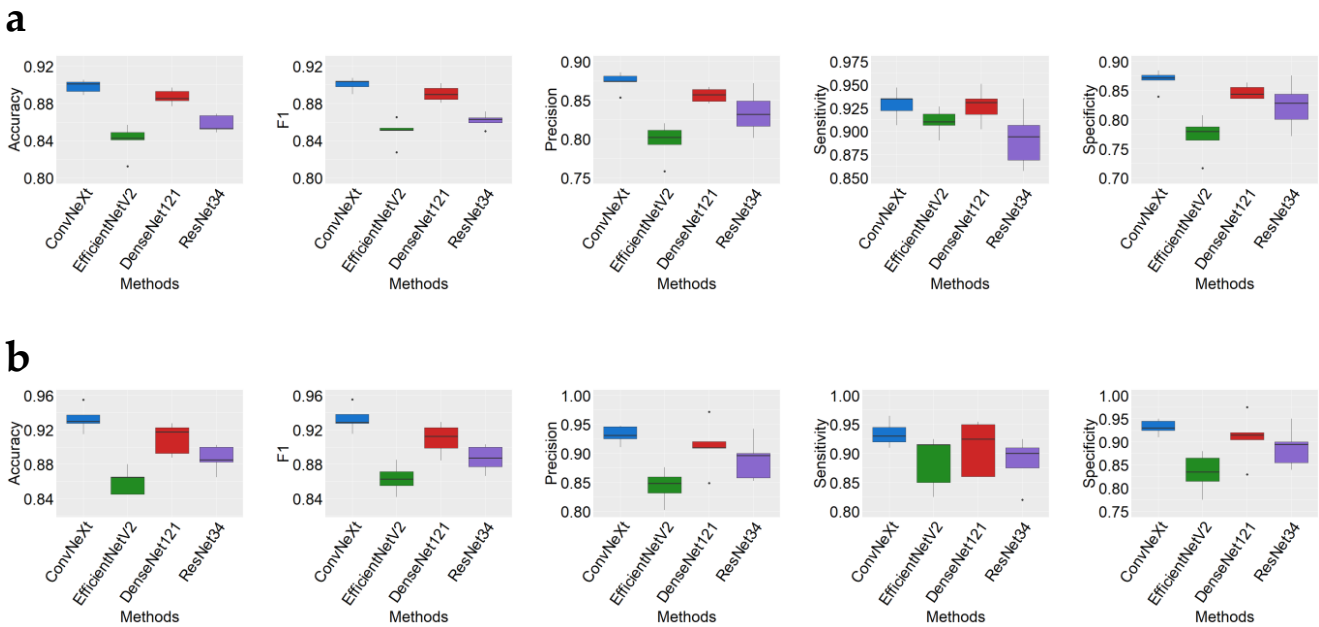

Figure 6. Boxplots illustrate several generalization (testing) performance results of four DTL models when utilizing five-fold cross-validation pertaining to (a) CT and (b) X-ray datasets.

Table 2a and b report the average performance results of the five testing folds and that constitute five-fold cross-validation results pertaining to (a) CT and (b) X-ray datasets. In terms of CT dataset (see Table 2a), ConvNeXt is the best-performing model, yielding the highest average ACC of 0.894, the highest average F1 of 0.900, the highest average PRE of 0.873, the highest average SEN of 0.929, and the highest average SPE of 0.868. The second-best model is DenseNet121, which yielded an average ACC of 0.887, average F1 of 0.890, average PRE 0.856, average SEN of 0.927, and average SPE of 0.847. EfficientNetV2 is the worst-performing model, yielding the lowest average ACC of 0.840, the lowest average F1 of 0.849, the lowest average PRE of 0.796, the average SEN of 0.910, and the lowest average SPE of 0.771. When considering the X-ray dataset as shown in Table 2b, ConvNeXt is still the best-performing model, achieving the highest average ACC, F1, PRE, SEN and SPE of 0.933, 0.933, 0.933, 0.934, and 0.932, respectively. DenseNet121, the second-best performing model, obtained an average ACC, F1, PRE, SEN, and SPE of 0.909, 0.909, 0.909, 0.910, and 0.909, respectively. The worst-performing model, EfficientNetV2, obtained the lowest average ACC, F1, PRE, and SPE of 0.860, 0.863, 0.842, and 0.834, respectively. These reported performance results illustrate the high-performance results of ConvNeXt.

Table 2. Five-fold cross-validation results calculated on test folds of four DTL models pertaining to (a) CT and (b) X-ray datasets. ACC is accuracy. PRE is precision. SEN is sensitivity. SPE is specificity. Bold indicates a model obtaining the highest performance results.

a

Model	ACC	F1	PRE	SEN	SPE
ConvNeXt	**0.894**	**0.900**	**0.873**	**0.929**	**0.868**
EfficientNetV2	0.840	0.849	0.796	0.910	0.771
DenseNet121	0.887	0.890	0.856	0.927	0.847
ResNet34	0.858	0.861	0.833	0.892	0.824

b

Model	ACC	F1	PRE	SEN	SPE
ConvNeXt	**0.933**	**0.933**	**0.933**	**0.934**	**0.932**
EfficientNetV2	0.860	0.863	0.842	0.886	0.834
DenseNet121	0.909	0.909	0.909	0.910	0.909
ResNet34	0.887	0.886	0.887	0.886	0.888

Figure 7a and b unveil computational insights for generalization capability of 4 studied DTL models, applied to testing images pertaining to CT and X-ray datasets, respectively. Prediction differences for ConvNeXt (and all other models) between COVID-19 and non-COVID-19 images were statistically significant ($P < 2.2 \times 10^{-16}$ from a *t*-test), indicating that all models are general predictors for COVID-19 or non-COVID-19 image prediction. Similarly, for the X-ray dataset in Figure 7b, prediction differences of all models were statistically significant ($P < 2.2 \times 10^{-16}$ from a *t*-test), demonstrating that these models are general, rather than specific predictors for COVID-19 prediction.

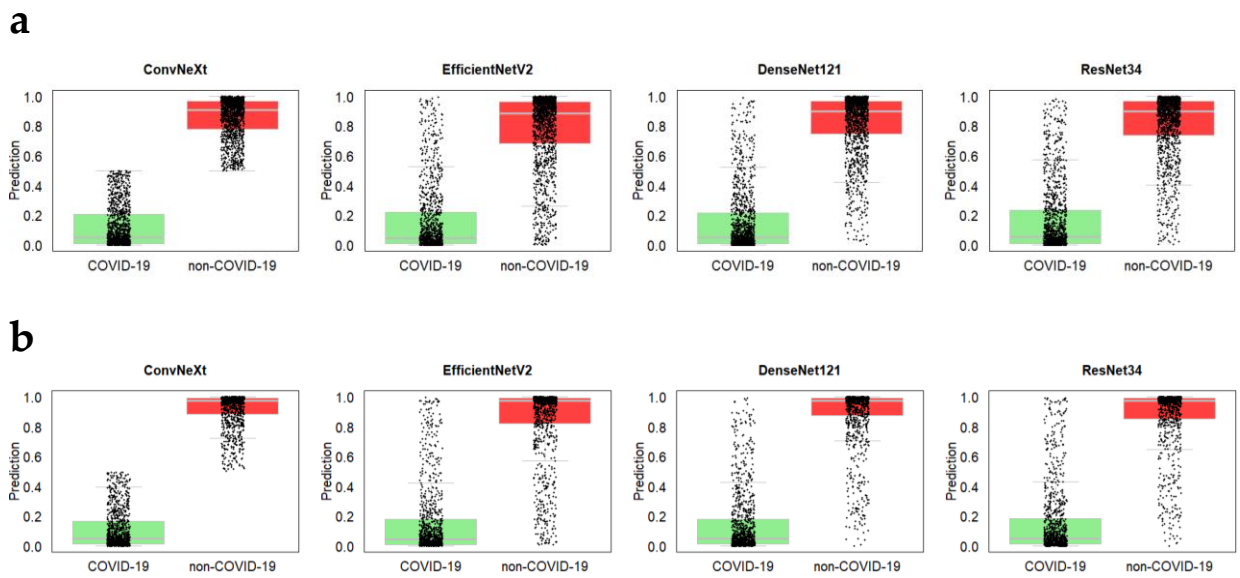

Figure 7. Boxplots and strip charts of predicted COVID-19 and non-COVID-19 images on all testing folds of four DTL models pertaining to (a) CT and (b) X-ray datasets.

After carrying out the model generation process explained in Section 2.3, we obtained 10000 models and plotted their AUC results when applied to CT and X-ray datasets as displayed in Figure 8a and b, respectively. For the CT dataset in Figure 8a, we selected the best model which we named HC, constructed as follows. Let $h_1(z)$, $h_2(z)$, $h_3(z)$, and $h_4(z)$ correspond to DTL models obtained from training ConvNext, EfficientNetV2, DenseNet121, and ResNet34, respectively, and applied to an unseen image z. In Equation 9, $h1_1(z)$ corresponds to a majority vote pertaining to two classifiers, namely ConvNext and EfficientNetV2 applied to a testing example. I is an indicator function resulting in 1 if the arguments are true and 0 otherwise. $h1_2(z)$ in Equation 10 corresponds to the majority vote of two classifiers (i.e., ConvNeXt and DenseNet121). In Equation 11, $h1_9(z)$ corresponds to a majority vote of three classifiers (i.e., ConvNeXt, DenseNet121, and ResNet34). HC in Equation 12 is constructed as a majority vote mechanism of $h1_1(z)$, $h1_2(z)$, and $h1_9(z)$.

$$h1_1(z) = \arg\max_y \sum_{i \in \{1,2\}} I(h_i(z) = y) \tag{9}$$

$$h1_2(z) = \arg\max_y \sum_{i \in \{1,3\}} I(h_i(z) = y) \tag{10}$$

$$h1_9(z) = \arg\max_y \sum_{i \in \{1,3,4\}} I(h_i(z) = y) \tag{11}$$

$$\mathrm{HC}(z) = \arg\max_y \sum_{i \in \{1,2,9\}} I(h1_i(z) = y) \tag{12}$$

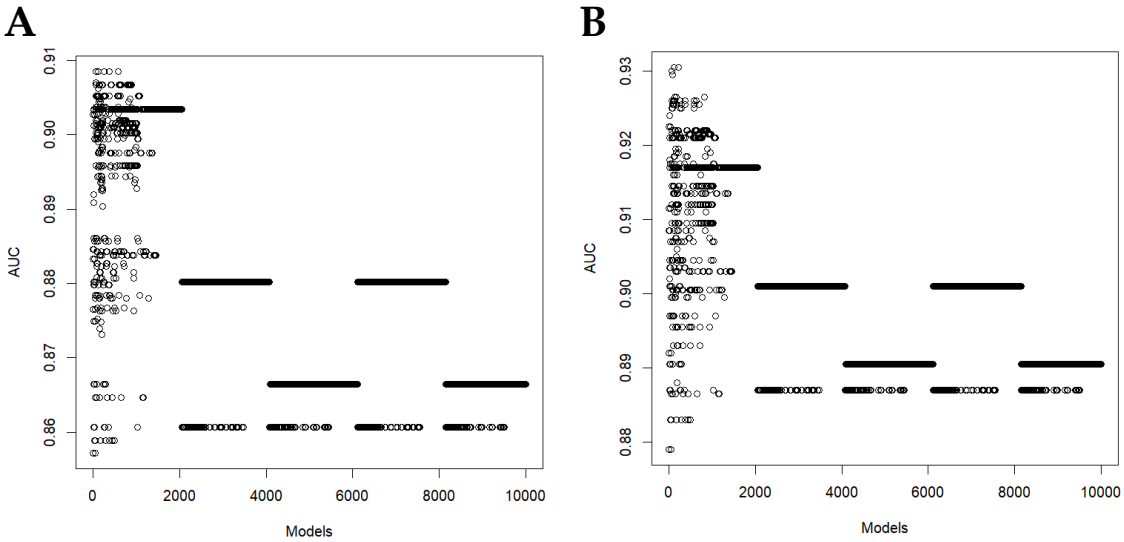

Figure 8. AUC performance results of 10000 generated models pertaining to (a) CT and (b) X-ray datasets.

The AUC for HC is 0.909 as illustrated in Figure 9a. For the X-ray dataset as illustrated in Figure 8b, we selected the best model attaining the highest AUC performance and designated as HX, working as shown in Equations 13-16 as follows. $h1_2(z)$ in Equation 13 corresponds to a majority vote pertaining to two classifiers (i.e.., ConvNeXt and DenseNet121). $h1_3(z)$ in Equation 14 is the majority vote of ConvNeXT and ResNet34. $h1_9(z)$ in Equation 15 is just the majority vote of ConveNeXt, DenseNet121, and ResNet34. In Equation 16, HX is the majority vote of $h1_2(z)$, $h1_3(z)$, and $h1_9(z)$. We report the obtained AUC results via HX in Figure 9b.

$$h1_2(z) = \arg\max_y \sum_{i \in \{1,3\}} I(h_i(z) = y) \tag{13}$$

$$h1_3(z) = \arg\max_y \sum_{i \in \{1,4\}} I(h_i(z) = y) \tag{14}$$

$$h1_9(z) = \arg\max_y \sum_{i \in \{1,3,4\}} I(h_i(z) = y) \tag{15}$$

$$\mathrm{HX}(z) = \arg\max_y \sum_{i \in \{2,3,9\}} I(h1_i(z) = y) \tag{16}$$

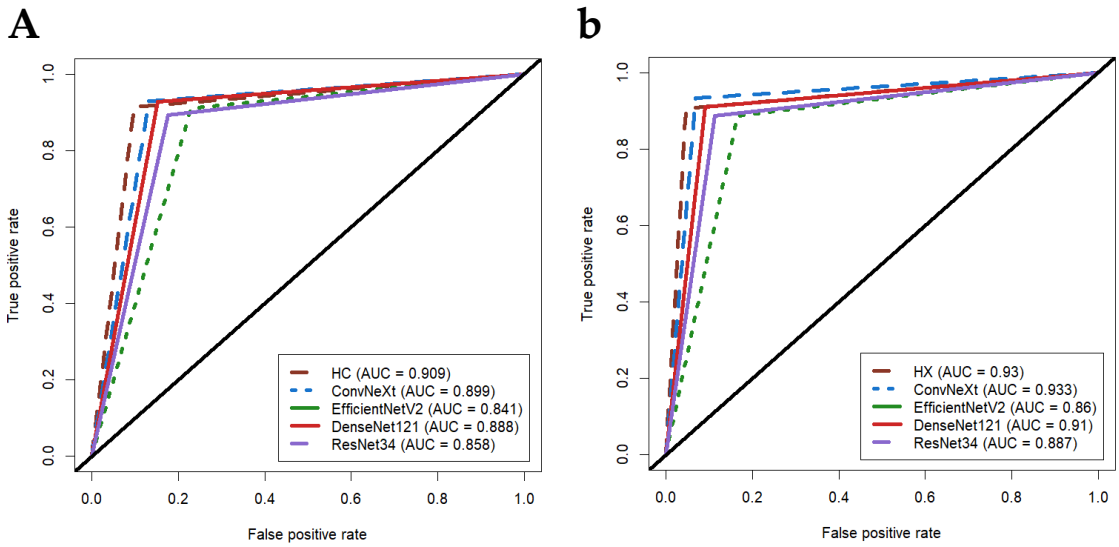

Figure 9. ROC curves and recorded AUC performance results of four DTL models against the generated models HC and HX to (a) CT and (b) X-ray datasets.

Figure 10a and b report the confusion matrices calculated based on testing for CT and X-ray datasets, respectively. When considering the CT dataset, it can be shown that HC is the best model in which accurately predicting 2254 out of 2481 CT images. One hundred and twenty-one images were predicted as non-COVID-19 while their actual label is COVID-19 and thereby counted as 121 FP. One hundred and six images were predicted as COVID-19 while their actual label is non-COVID-19. Hence, counted as 106 FN.

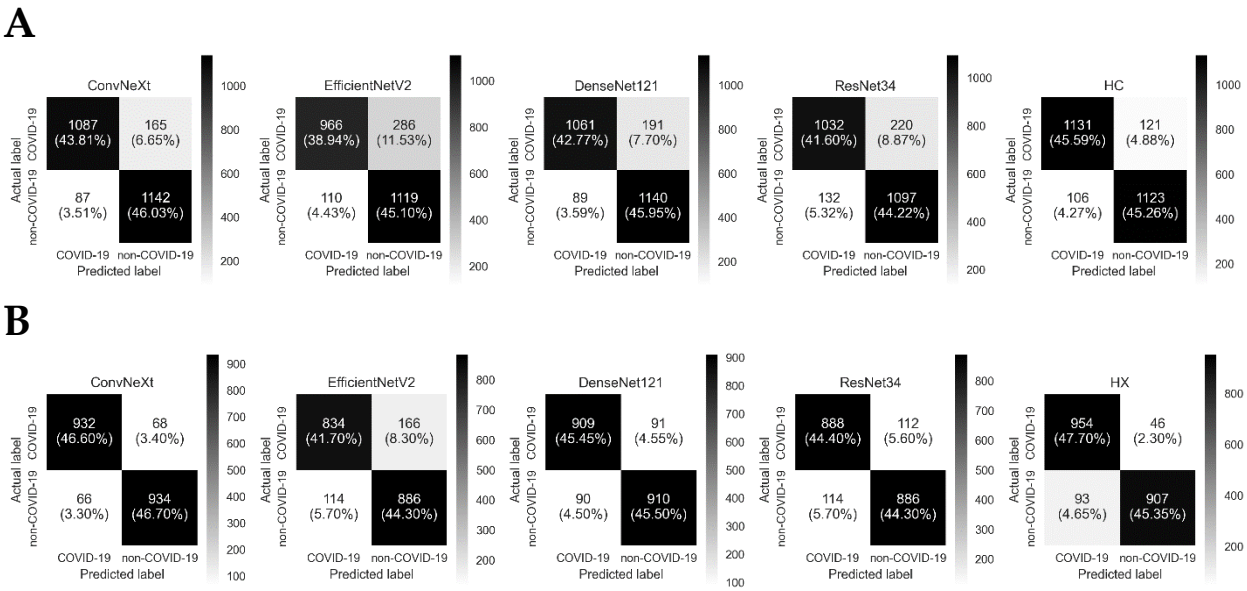

Figure 10. The combined confusion matrices of five testing folds during the running of five-fold cross-validation for ten models pertaining to (a) CT and (b) X-ray datasets.

The second-best model is ConvNeXt accurately predicting 2229 out of 2481 CT images. One hundred and sixty-five images were predicted by ConvNeXt as non-COVID-19 while their actual label is COVID-19 and thereby counted as 165 FP. Eighty-seven images were incorrectly predicted as COVID-19 while their actual label is non-COVID-19. Therefore, counted as 87 FN. EfficientNetV2 is the worst performing model, accurately predicting 2085 (1119 TP + 966 TN) CT images out of 2481 while inaccurately predicting 396 (286 FP + 110 FN) images. For the X-ray images, ConvNeXt accurately predicts 1866 (934 TP + 932 TN) images, which is marginally higher than HX (1861 calculated from 907 TP + 954 TN). Both ConvNeXT and HX inaccurately predict 134 (68 FP + 66 FN) and 139 (46 FP + 93 FN) images, respectively. ResNet34 accurately predicts 1774 (886 TP + 888 TN) images while inaccurately predicted 226 (112 TP + 114 TN) images. EfficientNetV2, the worst performing model, accurately predicted 1720 (886 TP + 834 TN) images while inaccurately predicted 280 (166 TP + 114 TN) images.

4. Discussion

To conduct a large empirical study while COVID-19 classification using CT and X-ray images, our computational approach is comprised of three parts: data preprocessing and DTL followed by model generation. In terms of data preprocessing part, the CT images composed of 2481 images pertaining to COVID-19 and non-COVID-19 patients, downloaded from Kaggle at https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset. The 2000 X-ray images regarding COVID-19 and non-COVID-19 were downloaded from Kaggle at https://www.kaggle.com/tawsifurrahman/covid19-radiography-database. We prepared all PNG images to be of size 256 × 256. For the DTL computation, we employed four pre-trained DL models, including ConvNeXT, EfficientNetV2, DenseNet121, and ResNet34. Each of these DTL models was already trained on more than a million images from the ImageNet database. We extracted features from COVID-19 images using weights obtained from pre-trained models on the ImageNet database, followed by a flattening step before training from scratch the densely connected classifier with Adam optimizer and binary cross entropy with logits loss function. The model generation part receives 4 DTL models. Then, taking the majority of 2, 3, and 4-combinations from the 4 DTL models, resulting in $\sum_{r=2}^{4}\binom{4}{r} = 11$ constructed models. After that, taking the majority of 2, 3, …, 11-combinations from the 11 models, resulting in $\sum_{r=2}^{11}\binom{11}{r} = 2036$ constructed models. Finally, selecting 7953 models from $\binom{2036}{2}$.

In our computational approach, it can be seen from Equations 1-3 that $h3_i$ depends on $h2_i$, which also depends on $h1_i$. Therefore, it can be viewed as $h3_i(h2_i(h1_i(z)))$ and that demonstrates the sequential nature for constructing these generated models. For the HC model as shown in Equations 9-12, Equations 9, 10, and 11 are exactly as calculations in Equation 1 while HC in Equation 12 is like $h2_i$ in Equation 2. Hence, it can be viewed as $HC(h1_{1,2,9}(z))$. The same computational nature holds true for HX corresponding to $HX(h1_{2,3,9}(z))$. Figure 8 reports the AUC for 10000 models applied to CT and X-ray datasets, respectively. In Supplementary ctLGM and xrayLGM, we report the 10000 constructed models obtained from CT and X-ray datasets, respectively.

To ensure fairness of experimental comparisons, we utilized five-fold cross-validation in which in all the five runs the training is used to induce models while testing is to perform prediction and evaluation. The sum of entries of each confusion matric in Figure 10 is equal to the number of images in studied datasets. The sum of each confusion matrix in Figure 10a is equal to 2481. Similarly, the sum of entries in each confusion matrix in Figure 10b is equal to 2000. This demonstrates that we evaluated the performance on the studied datasets.

In terms of reducing overfitting effect for induced DTL models, we included a dropout layer in the classification part pertaining to the densely connected classifier setting the ratio to 0.2. Furthermore, we transferred weights in the feature extraction part of pre-trained models already trained on more than one million images while training from

scratch the classification part in which weights in those layers are changed. In other words, incorporating dropout layer as well as transfer learning via freezing all layers in the feature extraction part contributed to the mitigation of overfitting in this study [53] .

Table 1 gives an overview of 14 existing studies, showing that our proposed study is unique when considering the number of models (see No. Models column). There is no large empirical DL study so far, to the best of our knowledge, reporting and investigating the feasibility of DL when tackling the COVID-19 classification task using CT and X-ray images as ours. Specifically, for each imaging technique (i.e., CT or X-ray), we reported results of 4 DTL models plus 10000 generated DTL models (see Figure 8). These results show the power of our computational approach. And we have 10004 reported results for CT images + 10004 reported results for X-ray images, resulting in a total of 20008 reported results.

5. Conclusions and Future Work

In this work, we present a novel computational approach to augment and expand empirical studies for the task of COVID-19 classification using CT and X-ray images as follows. First, we downloaded and processed 2481 CT and 2000 X-ray images from Kaggle for the COVID-19 binary classification task. Then, providing processed images of each imaging technique to four pre-trained DL models including ConvNeXt, EfficientNetV2, DenseNet121, and ResNet34. To employ transfer learning, we froze all layers including convolutional and pooling layers in the feature extraction part applied to COVID-19 images while unfreezing and updating weights during the training phase in the classification part of the densely connected classifier using Adam optimizer and binary cross entropy with logits loss function. The generation of models is done in a sequential manner, generating $\sum_{r=2}^{4}\binom{4}{r} = 11$ DTL models through taking the majority vote of 2, 3, and 4-combinations from the 4 DTL models. Then, generating $\sum_{r=2}^{11}\binom{11}{r} = 2036$ DTL models through taking majority vote of 2, 3, …, 11-combinations from 11 DTL models. Finally, generating 7953 DTL models from $\binom{2036}{2}$. For each imaging technique, we report results for 10004 (4+11+2036+7953) DTL models. Generalization (testing) results demonstrate that a generated DTL model named HC generated the highest AUC of 0.909 when applied to COVID-19 CT images. For the COVID-19 X-ray images, ConvNeXt generated the highest AUC of 0.933, which is marginally higher than the generated DTL named HX, yielding an AUC of 0.930. The results attributed to the computational approach successfully demonstrate how to scale empirical studies as well as set a unique record for COVID-19 classification.

Future work can include: (1) adopting the presented computational approach to generate and establish committee of models to be utilized in learning algorithms under the active learning scenario; (2) adjusting the computational approach to construct models via interleaved weighted majority vote; (3) expanding the model generation to reach one million constructed models ; and (4) employ the presented approach to augment and enlarge empirical studies for other problems in biology and medicine.

Data Availability

The CT dataset in this study is available at https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset, accessed on 17 March 2024. The X-ray dataset is available at https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database, accessed on 17 March 2024.

Author Contributions

T.T. conceived and designed the study. M.A. performed deep learning experiments. T.T. and M.A. contributed to the visualization of results. T.T., Y.-h.T., and M.A. evaluated the results, discussions, and outcomes and wrote and reviewed the manuscript. All authors have read and agreed to the submitted version of the manuscript.

Funding

This research received no funding.

Competing Interests

The authors declare no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

DL	Deep Learning
DTL	Deep Transfer Learning
LGM	Large Generation of Models
DenseNet	Dense Convolutional Network
ResNet	Residual Neural Network
CNN	Convolutional Neural Network
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
COVID-19	Coronavirus Disease 19
Adam	Adaptive Moment Estimation
CT	Computed Tomography
ACC	Accuracy
AUC	Area Under the ROC Curve
PRE	Precision
SEN	Sensitivity
SPE	Specificity

References

1. Wohl, D.A., et al., *COVID-19 symptoms at time of testing and association with positivity among outpatients tested for SARS-CoV-2.* PLoS One, 2021. **16**(12): p. e0260879.
2. Sosnowski, T.R., *Inhaled aerosols: Their role in COVID-19 transmission, including biophysical interactions in the lungs.* Current Opinion in Colloid & Interface Science, 2021. **54**: p. 101451.
3. Przekwas, A. and Z. Chen, *Washing hands and the face may reduce COVID-19 infection.* Medical hypotheses, 2020. **144**: p. 110261.
4. Bazaid, A.S., et al., *Knowledge and practice of personal protective measures during the COVID-19 pandemic: A cross-sectional study in Saudi Arabia.* PloS one, 2020. **15**(12): p. e0243695.
5. Sakalli, E., et al., *Ear nose throat-related symptoms with a focus on loss of smell and/or taste in COVID-19 patients.* American Journal of Otolaryngology, 2020. **41**(6): p. 102622.
6. Li, Z., et al., *Active case finding with case management: the key to tackling the COVID-19 pandemic.* The lancet, 2020. **396**(10243): p. 63-70.
7. Ozturk, T., et al., *Automated detection of COVID-19 cases using deep neural networks with X-ray images.* Computers in biology and medicine, 2020. **121**: p. 103792.
8. Apostolopoulos, I.D. and T.A. Mpesiana, *Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks.* Physical and Engineering Sciences in Medicine, 2020. **43**(2): p. 635-640.
9. Ismael, A.M. and A. Şengür, *Deep learning approaches for COVID-19 detection based on chest X-ray images.* Expert Systems with Applications, 2021. **164**: p. 114054.

10. Ravi, V., et al., *Deep learning-based meta-classifier approach for COVID-19 classification using CT scan and chest X-ray images.* Multimedia systems, 2022. **28**(4): p. 1401-1415.
11. Asif, S., et al., *A deep learning-based framework for detecting COVID-19 patients using chest X-rays.* Multimedia Systems, 2022. **28**(4): p. 1495-1513.
12. Gupta, K. and V. Bajaj, *Deep learning models-based CT-scan image classification for automated screening of COVID-19.* Biomedical Signal Processing and Control, 2023. **80**: p. 104268.
13. Ayalew, A.M., et al., *X-Ray image-based COVID-19 detection using deep learning.* Multimedia Tools and Applications, 2023. **82**(28): p. 44507-44525.
14. Constantinou, M., et al., *COVID-19 classification on chest X-ray images using deep learning methods.* International Journal of Environmental Research and Public Health, 2023. **20**(3): p. 2035.
15. Turki, T., S. Al Habib, and Y. Taguchi, *Novel Automatic Classification of Human Adult Lung Alveolar Type II Cells Infected with SARS-CoV-2 through the Deep Transfer Learning Approach.* Mathematics, 2024. **12**(10): p. 1573.
16. Cohen, J.P., et al., *Covid-19 image data collection: Prospective predictions are the future.* https://github.com/ieee8023/covid-chestxray-dataset. arXiv preprint arXiv:2006.11988, 2020.
17. M V Dadario, A., *Kaggle, 10.34740/kaggle/dsv/1019469*. 2020.
18. Kermany, D.S., et al., *Identifying medical diagnoses and treatable diseases by image-based deep learning.* cell, 2018. **172**(5): p. 1122-1131. e9.
19. Suganyadevi, S. and V. Seethalakshmi, *CVD-HNet: Classifying Pneumonia and COVID-19 in Chest X-ray Images Using Deep Network.* Wireless Personal Communications, 2022. **126**(4): p. 3279-3303.
20. *Radiology Assistant, X-ray Chest images,* https://radiologyassistant.nl/chest/lk-jg-1. 2020.
21. Ahamed, K.U., et al., *A deep learning approach using effective preprocessing techniques to detect COVID-19 from chest CT-scan and X-ray images.* Computers in biology and medicine, 2021. **139**: p. 105014.
22. Rahman, T., *Covid-19 Radiography Database,* https://www.kaggle.com/tawsifurrahman/covid19-radiography-database. 2021.
23. Mooney, P., *Chest X-Ray Images (Pneumonia).* https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia 2021.
24. Maftouni, M., *Large COVID-19 CT scan slice dataset.* https://www.kaggle.com/datasets/maedemaftouni/large-covid19-ct-slice-dataset.
25. Jain, R., et al., *Deep learning based detection and analysis of COVID-19 on chest X-ray images.* Applied Intelligence, 2021. **51**: p. 1690-1700.
26. Patel, P., *Chest X-ray (Covid-19 & Pneumonia).* https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia.
27. El-Shafai, W. and F. Abd El-Samie, *Extensive COVID-19 X-ray and CT chest images dataset.* Mendeley data, 2020. **3**(10).
28. Furtado, A., et al., *A Light Deep Learning Algorithm for CT Diagnosis of COVID-19 Pneumonia.* Diagnostics, 2022. **12**(7): p. 1527.
29. Simpson, A.L., et al., *A large annotated medical image dataset for the development and evaluation of segmentation algorithms.* arXiv preprint arXiv:1902.09063, 2019.
30. Pedrosa, J., et al., *LNDb: a lung nodule database on computed tomography.* arXiv preprint arXiv:1911.08434, 2019.
31. Yang, J., et al., *Data from lung CT segmentation challenge.* The cancer imaging archive, 2017.
32. Morozov, S.P., et al., *MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic.* Digital Diagnostics, 2020. **1**(1): p. 49-59.
33. Jun, M., et al., *COVID-19 CT lung and infection segmentation dataset.* 2020.
34. MedSeg, X., H. Jenssen, and T. Sakinis, *Medseg covid dataset 1*. 2021.
35. Vayá, M.D.L.I., et al., *BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients.* arXiv preprint arXiv:2006.01174, 2020.

36. Soares, E., et al., *SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification.* MedRxiv, 2020: p. 2020.04. 24.20078584.
37. Soares, E.A., Plamen, *SARS-COV-2 CT-Scan Dataset*. https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset. 2020.
38. Salau, A.O. and S. Jain, *Adaptive diagnostic machine learning technique for classification of cell decisions for AKT protein.* Informatics in Medicine Unlocked, 2021. **23**: p. 100511.
39. *COVID-QU-Ex Dataset*. https://www.kaggle.com/datasets/anasmohammedtahir/covidqu. 2022.
40. Echtioui, A. and Y.B. Ayed, *Automated detection of COVID-19 based on transfer learning.* Multimedia Tools and Applications, 2024. **83**(11): p. 33731-33751.
41. *Google Drive*. https://drive.google.com/drive/folders/1NLjyns6qJcQE8zZ8OZ-xfNMatcLLnMJO.
42. *SARS-CoV-2 infection in human adult lung alveolar stem cells*. https://www.ebi.ac.uk/empiar/EMPIAR-10533/. 2020.
43. Youk, J., et al., *Three-dimensional human alveolar stem cell culture models reveal infection response to SARS-CoV-2.* Cell Stem Cell, 2020. **27**(6): p. 905-919. e10.
44. Alghamdi, S. and T. Turki, *A novel interpretable deep transfer learning combining diverse learnable parameters for improved T2D prediction based on single-cell gene regulatory networks.* Scientific Reports, 2024. **14**(1): p. 4491.
45. Turki, T. and Z. Wei, *Boosting support vector machines for cancer discrimination tasks.* Computers in biology and medicine, 2018. **101**: p. 236-249.
46. Chollet, F., *Deep learning with Python*. 2021: Simon and Schuster.
47. Paszke, A., et al., *Pytorch: An imperative style, high-performance deep learning library.* Advances in neural information processing systems, 2019. **32**.
48. Paszke, A., et al., *Automatic differentiation in pytorch.* NIPS 2017 Workshop, 2017.
49. Stevens, E., L. Antiga, and T. Viehmann, *Deep learning with PyTorch*. 2020: Manning Publications.
50. Harris, C.R., et al., *Array programming with NumPy.* Nature, 2020. **585**(7825): p. 357-362.
51. McKinney, W. *Data structures for statistical computing in Python*. in *SciPy*. 2010.
52. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
53. Gao, H., J. Pei, and H. Huang. *Demystifying dropout*. in *International conference on machine learning*. 2019. PMLR.