Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2023, 2023:13
https://doi.org/10.28919/cmbn/7526
ISSN: 2052-2541

# HANDLING SEVERE DATA IMBALANCE IN CHEST X-RAY IMAGE CLASSIFICATION WITH TRANSFER LEARNING USING SWAV SELF-SUPERVISED PRE-TRAINING

HERY HARJONO MULJO[1,2,*], BENS PARDAMEAN[1,3], GREGORIUS NATANAEL ELWIREHARDJA[1], ALAM AHMAD HIDAYAT[1], DIGDO SUDIGYO[1], REZA RAHUTOMO[1,4], TJENG WAWAN CENGGORO[1,5]

[1]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia

[2]Accounting Department, School of Accounting, Bina Nusantara University, Jakarta 11480, Indonesia

[3]Computer Science Department, BINUS Graduate Program – Master of Computer Science Program, Bina Nusantara University, Jakarta 11480, Indonesia

[4]Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta 11480, Indonesia

[5]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia



**Abstract.** Ever since the COVID-19 outbreak, numerous researchers have attempted to train accurate Deep Learning (DL) models, especially Convolutional Neural Networks (CNN), to assist medical personnel in diagnosing COVID-19 infections from Chest X-Ray (CXR) images. However, data imbalance and small dataset sizes have been an issue in training DL models for medical image classification tasks. On the other hand, most researchers focused on complex novel methods instead and few explored this problem. In this research, we demonstrated how Self-Supervised Learning (SSL) can assist DL models during pre-training, and Transfer Learning (TL) can be used in training the models, which can produce models that are more robust to data imbalance. The Swapping Assignment between Views (SwAV) algorithm in particular has been known to be outstanding in enhancing the

*Corresponding author
E-mail address: heryhm@binus.edu
Received November 14, 2022

accuracy of CNN models for classification tasks after TL. By training a ResNet-50 model pre-trained using SwAV on a severely imbalanced CXR dataset, the model managed to greatly outperform its counterpart pre-trained in a standard supervised manner. The SwAV-TL ResNet-50 model attained 0.952 AUROC with 0.821 macro-averaged F1 score when trained on the imbalanced dataset. Hence, it was proven that TL using models pre-trained through SwAV can achieve better accuracy even when the dataset is severely imbalanced, which is usually the case in medical image datasets.

**Keywords:** deep learning; self-supervised learning; SwAV; chest X-Ray; data imbalance.

**2020 AMS Subject Classification:** 93A30, 65D18, 97R40, 68T01.

## 1. INTRODUCTION

Ever since its outbreak in December 2019, the coronavirus disease 2019 (COVID-19) has become a burden for humanity due to global massive incidences of the disease that are perpetuated by its contagious and rapid spreading. The symptoms of COVID-19 are dependent on its host and variant. According to the U.S. Centers for Disease Control and Prevention (CDC) 2020, people who developed noticeable symptoms mainly experienced mild or moderate symptoms (81%), while the rest developed severe (14%) and critical symptoms (5%) [1]. The moderate symptoms were often followed by mild pneumonia, while severe and critical symptoms commonly involved hypoxia, dyspnea, respiratory failure, or multiorgan dysfunction. The chest X-Ray imaging also showed more than 50% lung involvement [2]. To detect infections, the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test has been widely considered the standard to confirm COVID-19 infection using samples from nasopharyngeal swabs [3].

Aside from RT-PCR, imaging detection can also be utilized for COVID-19 screening, such as chest X-Ray (CXR) imaging and computed tomography [4, 5]. In most cases, the CXR-based diagnosis of pneumonia caused by COVID-19 was determined based on the presence of rounded morphology with ground-glass opacities in the chest imaging, which is characterized by bilateral peripheral distribution associated with crazy-paving patterns [6]. In some peak stages of COVID-19 infection, architectural distortion marked with subpleural bands also occurred [7]. Those are typically used to differentiate cases of pneumonia caused by COVID-19 from those caused by other factors. Using those features, radiologists can manually interpret the images to annotate the presence of COVID-19 pneumonia. As manual diagnoses are not

efficient, researchers across the globe have attempted to implement computer-aided COVID-19 diagnosis using Machine Learning (ML), especially the Deep Learning (DL) approach using Convolutional Neural Networks (CNN), which can increase the efficiency and accuracy of early diagnoses to assist radiologists in rapidly finding suspicious patterns on lung images [8].

Supervised learning is the most straightforward DL approach for early detection of COVID-19 from CXR images by training DL models on labeled images [9]. The models are then evaluated on an 'unseen' test set with the assumption that both the training and test sets are obtained from the same data distribution. However, DL requires enormous volumes of data to be accurate [10, 11] and minimize the risk of overfitting, which means that it may be ineffective when the number of available data is insufficient. Moreover, training the models from scratch for a single specific task is exhaustive, which prompts researchers to deploy the Transfer Learning (TL) approach. For tasks related to medical images, the approach utilizes the learned knowledge of the models pre-trained on a large computer vision dataset such as the ImageNet dataset and repurposes the knowledge to another computer vision task [12]. This transferred knowledge can then be fine-tuned by using the zero, partial, or full network adaption [13]. TL has been extensively implemented in many research publications on CXR-based COVID-19 classification. For example, in our previous work, we developed a CXR-based classification model by using a DenseNet-121 backbone fine-tuned to classify lung diseases by combining two public CXR datasets and obtained AUC scores above 80% for different task configurations [14].

Another existing challenge in training DL models on medical datasets is the imbalanced condition of the datasets [15]. As high-risk patients may be rarer, it is common for severe imbalance to occur which will influence the performance of ML models, where the models' predictions may lean towards the class with more samples [16]. For most studies related to DL in analyzing CXR images, researchers typically preferred either augmenting the minority classes as a method of oversampling or performing Random Undersampling (RUS) [17, 18, 19]. However, both have their own weaknesses. RUS may remove valuable data randomly and cause sampling bias. In addition, the reduced number of data may be insufficient to train DL models. On the other hand, the usage of oversampling will consume more training time and

may possibly cause the models to overfit if the augmented samples were too similar [20]. In other words, when the imbalance is too severe and too many augmented images were generated, the model may focus too much on specific features of the images from the minority class and fail to capture the relevant general features, resulting in accuracy degradation when evaluated on unseen samples of the minority class.

More recent studies had delved into several pre-training algorithms for TL. The Self-Supervised Learning (SSL) algorithm, especially the Swapping Assignments between Views (SwAV) algorithm, in particular, has been conspicuous, proving that it can allow CNN models to attain better accuracy on downstream tasks [21, 22]. This means that performing TL using models that had been pre-trained using SwAV can yield better results, implying that they are more capable of determining detailed distinguishing features of objects. A previous research had proven this fact by pre-training a CNN on a CXR dataset using SwAV and performing TL to train it on another CXR dataset, which proved that this method outperformed regular TL models pre-trained in a supervised manner [23]. However, few have explored other potentials of this method, including whether it is more robust against data imbalance. In this research, we evaluated two types of CNN models, specifically the ResNet-50 CNN architecture, pre-trained using SwAV and standard supervised learning on classifying CXR images without oversampling. As the dataset is also imbalanced, the model pre-trained using SwAV may generalize better as it possesses more knowledge on distinguishing similar features without regard to the classes. This paper is organized as follows: section 2 presents some previous works related to CXR classification using TL, section 3 describes the details of how the experiments in this study were conducted, section 4 presents the obtained results and analyses, and section 5 presents the conclusion of this research.

## 2. Related Works

The emergence of research dedicated to the analysis of CXR images to diagnose COVID-19 by using TL becomes more prevalent during the peak of the COVID-19 pandemic. In 2020, Loey et al. used a Generative Adversarial Network (GAN) to augment CXR images from the only available public CXR COVID-19 dataset at the time for retraining three different pre-trained models (Alexnet, Googlenet, and Restnet18). Their finding showed that GoogleNet

achieved an accuracy of 0.806 in classifying four classes of CXR images [24]. Similarly, Rahaman et al. compared 15 pre-trained CNN architectures to classify normal, regular pneumonia, and COVID-19 CXR images, where the VGG19 model obtained an astounding accuracy of 0.893 and F1 score of 0.90 albeit augmentation was still involved [25]. In another similar research, Minaee et al. evaluated four pre-trained CNNs, namely ResNet18, ResNet50, SqueezeNet, and DenseNet-121, in detecting COVID-19 infections from collected CXR images that were analyzed by certified radiologists to determine the labels. By using TL and augmentation, the CNN models obtained around 0.9 specificity rates [26]. Overall, standard TL had been successful in classifying CXR images when trained using augmentations.

Chouhan et al. utilized the ensembling of CNN models after TL to classify pneumonia on the CXR images using five popular pre-trained CNN models. The output of each model were then assembled to produce the final output. By using augmented images from the Guangzhou Women and Children's Medical Center dataset, the model achieved 0.964 accuracy and 0.9962 recall, respectively [27]. In another research, a similar method was also performed using four public CXR datasets, where three pre-trained ResNet models were trained to perform binary classifications using three datasets (Normal-COVID, Pneumonia-COVID, Normal-Pneumonia). The models were then ensembled and further fine-tuned using the other dataset. The proposed method outperformed the individual models with a precision of 0.94 and a recall of 1.0 [28]. However, such methods require huge computation workloads as standard DL is already computationally expensive [29], not to mention the massive amount of data required to train the models and the augmented images.

On the other hand, various SSL-based approaches had been exploited in recent studies of CXR imaging classification. Liu et al. on elaborated self-supervised mean-teacher model pre-training with semi-supervised fine-tuning called $S^2MTS^2$ method that was evaluated on CXR datasets to perform multilabel classification. Using different proportions of the labeled and unlabeled data, their method produced similar results on the CXR dataset with the supervised approaches [30]. Azizi et al. proved that self-supervised pre-training on ImageNet followed by further self-supervised pre-training on unlabeled CXR images improved the model's performance on CXR classification on test sets. Combining with an alternative of contrastive learning

called Multi-Instance Contrastive Learning they showed that the method can beat the performance of supervised approaches on the testing sets [31].

More recent studies have demonstrated how SSL can be exploited for CXR-based COVID-19 classification. Abbas et al. proposed a TL approach to repurpose large-scale image classification tasks to COVID-19 detection on CXR using a self-supervised sample decomposition method. The approach called 4S-DT can deal with imbalanced class distribution in the dataset. Their method can attain high accuracies for the classification task [32]. In another study, Gazda et al. utilized self-supervised pre-training of deep CNN on CheXpert images with removed labels using contrastive learning approaches. The pre-trained models were then used to classify pneumonia types and COVID-19 recognition on different datasets. The results showed comparable results with supervised methods without using a large number of labeled datasets [33], which further proved the prowess of SSL. All in all, SSL and SwAV have yet to be fully explored for CXR image classification.

## 3. Research Methodology

**3.1. Dataset.** This research utilized the COVID-19 Radiography Database as a training dataset to run the CXR image classification task [18, 34]. The data acquired was split into four classes, containing 3616 images of positive COVID-19 cases, 10192 normal, 6012 lung opacity (non-COVID lung infections), and 1345 cases of viral pneumonia. The creator of the dataset also provides lung segmentation masks for all images, enabling segmentation to be done before training the DL models. Figure 1 shows the original image samples along with their respective segmentation masks for each class provided by the dataset. Images with irregularities were shown in the figure, such as texts and an arrow in Figure 1(A), an arrow in Figure 1(B), black-padded different-sized images in Figure 1(C), and cropped lung images in Figure 1(D). Therefore, the segmentation masks are used in segmenting the images before feeding them into the DL models and allowing the models to focus only on the lung areas of the images.

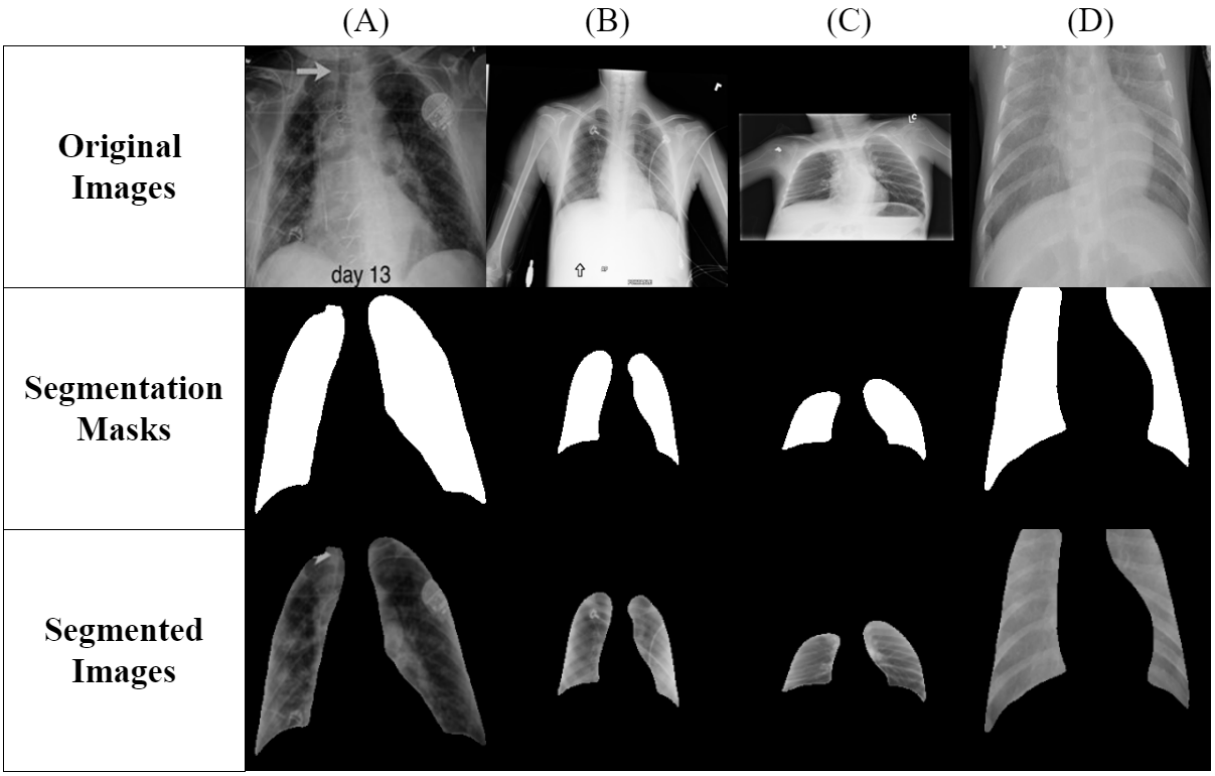


FIGURE 1. Examples of irregularities for each class of the dataset: (A) COVID, (B) Lung Opacity, (C) Normal, and (D) Viral Pneumonia

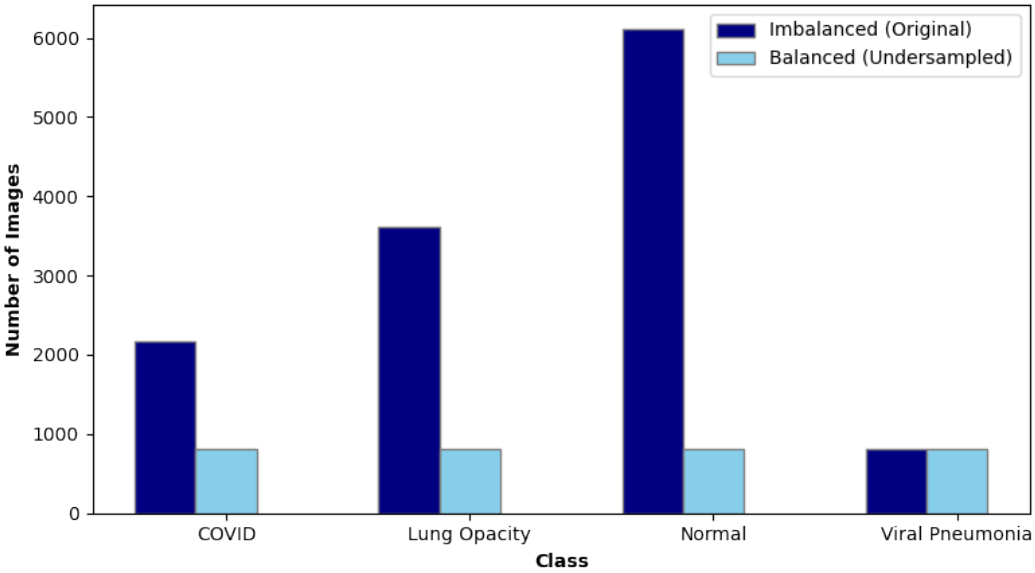


FIGURE 2. Distribution of images in each class of the train set

**3.2. Data Pre-processing.** All of the images were first segmented using the segmentation masks provided. The segmented images were then scaled down to 224 $\times$ 224 pixels. Afterwards, the images were split into train, validation, and test sets with a ratio of 60:20:20. In this research, the impact of data imbalance was also studied. Therefore, an undersampled version of the dataset was generated using Random Undersampling (RUS). It should be noted that RUS was only performed on the train set. As a result, two versions of the train set: (a) original imbalanced and (b) undersampled balanced were used and compared in training the models, meaning that two experiment scenarios were performed. The distribution of data for each version of the train set were listed in Figure 2. Following the approach from the previous research [14], the images were augmented using random horizontal flipping. However, it should be noted that the augmentation was only performed on the train set. All of the images were then normalized to floats in the range of 0 to 1 and standardized where the values of mean and standard deviation were set to [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225]. These standardized images were the ones used as inputs for the models.

**3.3. Transfer Learning from Self Supervised Learning.** In cases where the size of the dataset is insufficient to allow DL models to generalize, TL has been a broadly adopted method to assist the training of the models [35, 36]. Its main goal is to aid the models in obtaining better target predictive function $f_T(.)$ for the target task $T_T$ based on the target domain $D_T$. The models are first trained on a pretext task using the source domain $D_S$ and source task $T_S$, and the knowledge they obtained is transferred by retraining the models using $D_T$ and $T_T$ for a downstream task. Mathematically, TL can be formulized as follows:

$$\begin{aligned} D_S &= \{(x_{S1}, y_{S1}), (x_{S2}, y_{S2}), ..., (x_{Sn}, y_{Sn})\}, T_S = \{y_S, f(.)\} \\ D_T &= \{(x_{T1}, y_{T1}), (x_{T2}, y_{T2}), ..., (x_{Tn}, y_{Tn})\}, T_T = \{y_T, f(.)\} \end{aligned} \tag{1}$$

where $x$ and $y$ denotes the input data and their labels, respectively. The goal is to improve $f(.)$ by using the transferred knowledge from $D_S$ and $T_S$ [37]. In this research, the heterogenous TL method was adopted, in which $D_S \neq D_T$ and $T_S \neq T_T$. The ImageNet dataset [38] in particular, which contains more than three million images, has been widely used to pre-train proposed DL

models through supervised learning [39, 40, 41]. In general, this method has proven effective in various DL studies [36, 42, 43].

Aside from the supervised learning approach, SSL in the form of Contrastive Learning (CL) has been a prominent approach in pre-training DL models on pretext tasks [23]. CL trains DL models to cluster samples, allowing them to identify the same object from different augmented views [21]. This means that models trained using CL should be capable of distinguishing different representations of the same object, making them more robust compared to standard supervised models. The SwAV algorithm in particular is one of the most successful algorithms with the best accuracy among similar CL algorithms albeit inferior to standard supervised models. Inspired by contrastive instance learning, SwAV trains the models to differentiate various views of an image by comparing the cluster assignments produced instead of their features. This was done by utilizing the multi-crop strategy and trainable prototypes.

Figure 3 illustrates how SwAV was performed. First, the multi-crop augmentation was performed to generate various views of the input image $X$, resulting in the randomly cropped $X_1$ and $X_2$ which were further augmented using random horizontal flip. Color distortion and Gaussian blurring were then performed on $X_1$ and $X_2$, which were later fed into the model $F_\theta$. The output embeddings $Z_1$ and $Z_2$ were produced, and dot product operations were performed on $Z_1$ and $Z_2$ by using the prototype vector $C$ to produce the scores $Z_1.C$ and $Z_2.C$. The Sinkhorn-Knopp algorithm was used to assign the clusters from $Z_1.C$ and $Z_2.C$, resulting in the codes $Q_1$ and $Q_2$. In computing the loss, the assignments were then swapped as described in the equations below.

$$p_t^{(k)} = \frac{exp(\frac{1}{\tau}(Z_t \cdot C_k))}{\sum_{k'} exp(\frac{1}{\tau}(Z_t \cdot C_{k'}))} \tag{2}$$

$$L(Z_t, Q_s) = -\sum_k Q_s^{(k)} * log(p_t^{(k)}) \tag{3}$$

$$L(Z_t, Z_s) = L(Z_t, Q_s) + L(Z_s, Q_t) \tag{4}$$

where $k$ denotes the number of prototypes used and $\tau$ is a temperature variable used for softening the scores. As shown in equation 3, the loss function used is the Cross-Entropy (CE) loss, which were later averaged and the mean CE loss was used in back-propagating the model's parameters as well as $C$ [22].

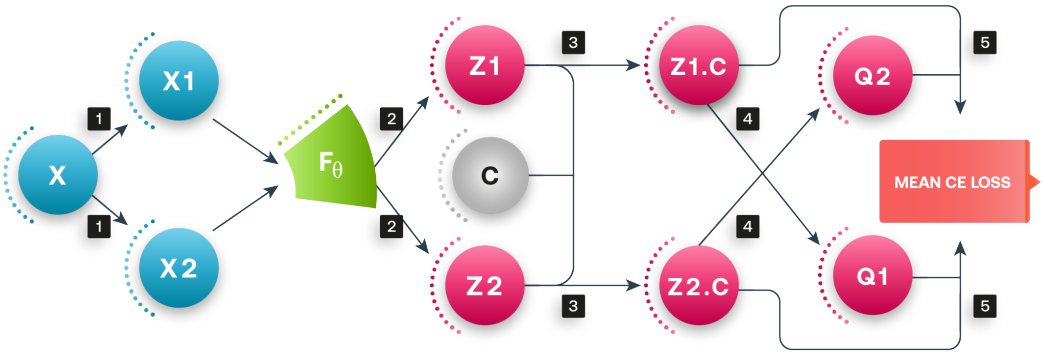


FIGURE 3. Illustration of the process of SwAV: (1) multi-crop augmentation, (2) output forward pass, (3) dot product, (4) cluster assignment, and (5) loss calculation

On most computer vision downstream tasks, performing TL on models trained using SwAV had produced better results compared to standard supervised learning [21, 22]. In a recent research, TL from SwAV had also been used for distinguishing COVID and normal lung X-Ray images and had proven to be superior to other TL models [23]. However, this implementation was still limited to binary classification and has yet to be tested on independent test sets. Therefore, deeper analyses on TL using SwAV for X-Ray image classification were conducted in this research by using four classes of chest X-Ray images. The results were also compared to that of TL using standard supervised method.

In this research, two ResNet-50 models were trained on the COVID-19 Radiography Database. Both models were pre-trained on the ImageNet dataset, one pre-trained using SwAV, which will be referred to as SwAV-TL in the following sections of this paper, and the other using standard supervised learning. First, all layers of the models were frozen and the output layer was directly attached on top of them. As the architecture of ResNet-50 consists of five convolution blocks and TL can still be performed by unfreezing some of the models' layers, the models were then trained with various number of unfrozen convolution blocks. Each model was fine-tuned by performing hyperparameter grid search. Five fine-tuned models were trained in total for the SwAV-TL and supervised versions, respectively, with 0, 1, 2, 3, and 4 unfrozen blocks counted from the output layer. The models were then evaluated on the test set and the results are presented in section 4.

**3.4. Evaluation Metrics.** For this multiclass CXR image classification task, confusion matrices were generated for each model. Several classification metrics were then derived from the matrices. In addition, the Area Under Receiver Operating Characteristic curve (AUROC) was also calculated.

**3.4.1.** *Confusion Matrix.* The confusion matrix or the error matrix is a visualization of classification tasks that presents the actual and predicted classes of data used by the classification algorithm [44]. For DL classification models, the matrix has widely been used to provide detailed information about the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for each class of the dataset. An illustration of the matrix is provided in Table 1.

TABLE 1. An illustration of a binary-class confusion matrix.

| **Actual** \ **Predicted** | **Positive** | **Negative** |
|---|---|---|
| **Positive** | $TP$ | $FN$ |
| **Negative** | $FP$ | $TN$ |

As the dataset used is imbalanced, the accuracy metric is not calculated in this experiment. It should also be noted that the models were also trained using the imbalanced dataset without resampling, meaning that it is possible for some of the models to have null precision in some classes. Therefore, only the recall/sensitivity/True Positive Rate (TPR), specificity/True Negative Rate (TNR), miss rate/False Negative Rate (FNR), and fall-out/False Positive Rate (FPR) are calculated in this research. These metrics are calculated as follows.

$$Recall = TPR = \frac{TP}{TP+FN} \tag{5}$$

$$TNR = \frac{TN}{TN+FP} \tag{6}$$

$$FNR = \frac{FN}{TP+FN} = 1 - TPR \tag{7}$$

$$FPR = \frac{FP}{TN+FP} = 1 - TNR \tag{8}$$

The precision and F1 score were only calculated for the best model of the research to compare the models to that of previous studies. These two metrics were calculated as follows:

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$F1score = 2 * \frac{Precision * TPR}{Precision * TPR} \tag{10}$$

**3.4.2.** *AUROC.* In addition to the metrics explained above, the AUROC has also been a reliable performance indicator for classification models. The ROC is a plot of TPR on the Y-axis and FPR on the X-axis for each class on various discrimination thresholds in the form of a probability curve. In other words, it describes the prowess of a classification model in distinguishing each class from the other classes of the dataset, meaning that it evaluates the binary classification ability of each class. The AUROC represents the degree of separability for each class. The higher its value is, the better the model is at distinguishing the specified class from other classes. The value of AUROC ranges from 0 to 1, where 1 indicates a better classification ability [45].

**3.5. Experiment Setup.** The experiments were conducted using the Python programming language and PyTorch DL framework. In tuning the hyperparameters, the HyperOpt library [46] was utilized to perform a grid search in determining the optimal learning rate $lr$ and L2 regularizer weight decay $\lambda$, where $lr \in \{1e-2, 1e-3, 1e-4, 1e-5\}$ and $\lambda \in \{0, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6\}$. All of the models were trained for 50 epochs using the Adam optimizer and CE loss. For each type of TL, the models were trained on the two scenarios: (a) using the original imbalanced dataset and (b) using the undersampled balanced dataset. This means that a total of 20 models were trained, 10 for each TL type. Class weighting was utilized in training the models on the imbalanced dataset. As a multiclass classification task was performed in this research, all of the reported metrics had been macro-averaged.

## 4. Results & Discussion

**4.1. Training Results.** Figure 4 summarizes the performance of each model during the training phase. It can be seen that the SwAV-TL models are able to obtain lower validation loss, both on the imbalanced and balanced dataset. On the imbalanced dataset, only the SwAV-TL model

with 4 unfrozen blocks attained validation loss $< 0.1$. Similar results were also observed on the balanced dataset, where the SwAV-TL models with 3 and 4 unfrozen blocks obtained less than 0.10625 validation losses. However, increasing the number of unfrozen blocks resulted in overfitting on all of the models despite the lower validation losses. Such results indicate that the number of training data is insufficient for the ResNet-50 models to generalize well on both datasets. The gap between the train and validation losses is even larger on the models trained using the balanced dataset, an outcome which is expected as RUS was used which reduces the number of train data as shown in Figure 2. To further validate whether TL using SwAV is more robust to overfitting, the models' performance on the test data had to be analyzed.

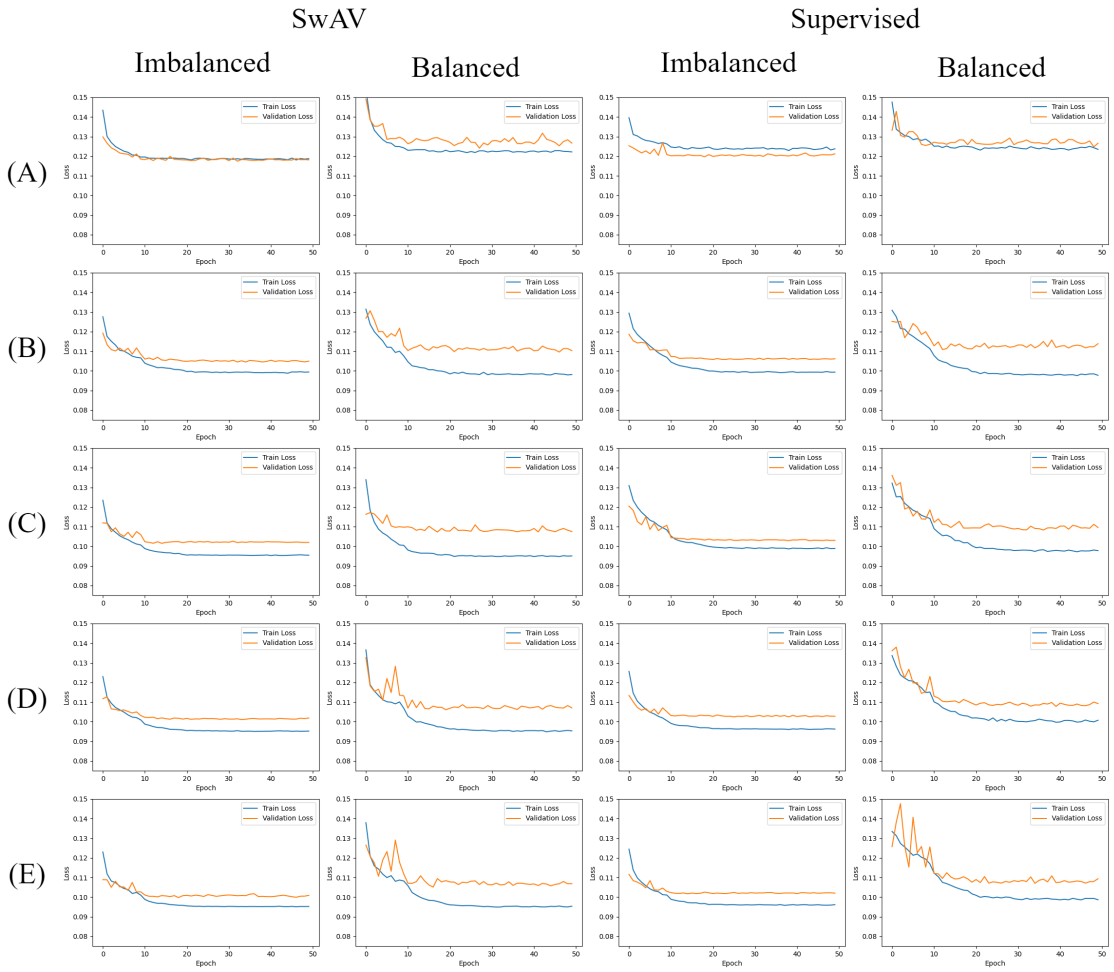


FIGURE 4. Comparison of the training and validation CE loss of each model trained using TL with: (A) 0 unfrozen blocks, (b) 1 unfrozen block, (C) 2 unfrozen blocks, (D) 3 unfrozen blocks, and (E) 4 unfrozen blocks

TABLE 2. Evaluation results of the models trained on the imbalanced dataset.

| Pre-training Method | Unfrozen Blocks | TPR | TNR | FNR | FPR | AUROC |
|---|---|---|---|---|---|---|
| **Supervised** | 0 blocks | 0.25 | 0.75 | 0.75 | 0.25 | 0.825 |
| | 1 block | 0.285 | 0.77 | 0.715 | 0.23 | 0.88 |
| | 2 blocks | 0.416 | 0.847 | 0.584 | 0.153 | 0.914 |
| | 3 blocks | 0.25 | 0.75 | 0.75 | 0.25 | 0.863 |
| | 4 blocks | 0.25 | 0.75 | 0.75 | 0.25 | 0.862 |
| **SwAV** | 0 blocks | 0.445 | 0.848 | 0.555 | 0.152 | 0.886 |
| | **1 block** | **0.808** | **0.938** | **0.192** | **0.062** | **0.952** |
| | 2 blocks | 0.472 | 0.878 | 0.528 | 0.122 | 0.934 |
| | 3 blocks | 0.487 | 0.885 | 0.513 | 0.115 | 0.944 |
| | 4 blocks | 0.46 | 0.873 | 0.54 | 0.127 | 0.933 |

**4.2. SwAV-TL Showed Exceptional Capability in Facing Data Imbalance.** Evaluation results on the models trained on the imbalanced dataset were compiled in Table 2. From the AUROC alone, it is clear that the SwAV-TL models outperform the supervised models on this downstream task, a result that parallels the findings from previous studies [23, 21, 22]. This means that SwAV-TL is confirmed to be more robust to overfitting, even when trained on the imbalanced dataset. Further observation on the TPR and TNR also implies that the supervised models were greatly affected by the class imbalance. Most of the supervised models obtained exactly 0.25 TPR and 0.75 TNR in four-classes classification, which implies that the models classify all of the test samples into a single class. Such results are expected in cases where the train data are severely imbalanced, as the ”Normal” class contains almost eight times more samples compared to the class with the least number of samples, which is the ”Viral Pneumonia” class. Even the class with the second largest number of samples, which is ”Lung Opacity”, only contains 58.98% of the number of samples in the ”Normal” class.

On the other hand, the SwAV-TL models obtained considerably better results, with the best model acquiring 0.952 AUROC. This model, which only has one unfrozen block, obtained an astounding performance with 0.808 TPR and 0.938 TNR. These results are vastly better than

the rest of the models, as the other SwAV-TL models failed to even achieve 0.5 TPR despite the high TNR. The low TPR means that the models are prone to false negatives, which can lead to a lot of undetected diseases. Such results parallel with a similar research on TL, which have proven that albeit models with more unfrozen blocks can achieve lower validation losses, they may not be the best on the independent test dataset [42].

In spite of the high AUROC, the supervised model with two unfrozen blocks still possessed lower TPR and TNR compared to the worst SwAV-TL model, which is the model with 0 unfrozen blocks which achieved 0.886 AUROC. Such results may be attributed to the fact that the SwAV-TL models were previously trained to cluster similar features of augmented images together [21], meaning that such models may have better general knowledge in grouping images with details and features that are generally similar on the downstream tasks while being robust to image transformations [22]. In simpler words, the models could have better knowledge in grouping images that are similar by highlighting detailed features that are more general compared to the supervised models which were trained to focus on extracting detailed distinguishing features of each class. Therefore, it can be inferred that TL using SwAV pre-training can result in better performance for models trained on imbalanced datasets.

**4.3. Undersampling Resulted in Generally Better Results.** To further verify whether TL using SwAV-TL is only advantageous on imbalanced datasets, experiments were also conducted on the undersampled version of the dataset. However, the number of samples used in training was significantly reduced and sampling bias may have affected the models due to the usage of RUS. The evaluation results were listed in Table 3. Although the performance of most of the models greatly improved when trained on the imbalanced dataset, the SwAV-TL models still managed to outperform all of their supervised counterparts. The best results were achieved by the SwAV-TL model with 3 unfrozen blocks, which maintained 0.948 AUROC, which is only slightly lower than that of the best SwAV-TL model on the imbalanced dataset. The difference is that on the imbalanced dataset, the other SwAV-TL models were unable to obtain at least 0.5 TPR whereas on the balanced dataset most of the models managed to obtain more than 0.76 TPR. Such results were expected as imbalanced distribution of data can severely affect the models.

TABLE 3. Evaluation results of the models trained on the undersampled dataset.

| Pre-training Method | Unfrozen Blocks | TPR | TNR | FNR | FPR | AUROC |
|---|---|---|---|---|---|---|
| **Supervised** | 0 blocks | 0.251 | 0.75 | 0.749 | 0.25 | 0.791 |
| | 1 block | 0.586 | 0.822 | 0.414 | 0.178 | 0.888 |
| | 2 blocks | 0.532 | 0.801 | 0.468 | 0.199 | 0.921 |
| | 3 blocks | 0.378 | 0.801 | 0.622 | 0.227 | 0.904 |
| | 4 blocks | 0.643 | 0.845 | 0.357 | 0.155 | 0.928 |
| **SwAV** | 0 blocks | 0.608 | 0.84 | 0.392 | 0.16 | 0.861 |
| | 1 block | 0.769 | 0.898 | 0.231 | 0.102 | 0.928 |
| | 2 blocks | 0.77 | 0.909 | 0.23 | 0.091 | 0.929 |
| | 3 blocks | **0.817** | **0.923** | **0.183** | **0.077** | **0.948** |
| | 4 blocks | 0.794 | 0.921 | 0.206 | 0.079 | 0.945 |

One intriguing fact from the results is that the performance boosts of the supervised models are much greater than the SwAV-TL models. They obtained an average of 68.31% and 3.44% improvement in TPR and TNR respectively compared to the ones trained on the imbalanced dataset. The supervised model with 4 unfrozen blocks in particular had the highest improvement with 157.09% and 12.62% higher TPR and TNR. However, the SwAV-TL models only obtained an average of 46.94% and 1.62% improvement of TPR and TNR, with the highest improvement of 72.55% and 5.55% better TPR and TNR on the SwAV-TL model with 4 unfrozen blocks. Such results also indicate that the SwAV-TL models are more robust when trained on imbalanced data. Additionally, the performances of the SwAV-TL models are still superior to the supervised models, which means that in the future, the usage of DL models trained using SSL for TL purposes can bring about significantly better results in cases where the used dataset is small or imbalanced.

**4.4. The Impact of Dataset Size and Class Imbalance.** Figure 5 presents the confusion matrices of the best models. Due to the severe data imbalance, the supervised models trained on the imbalanced dataset were unable to classify the images as COVID or Viral Pneumonia, which

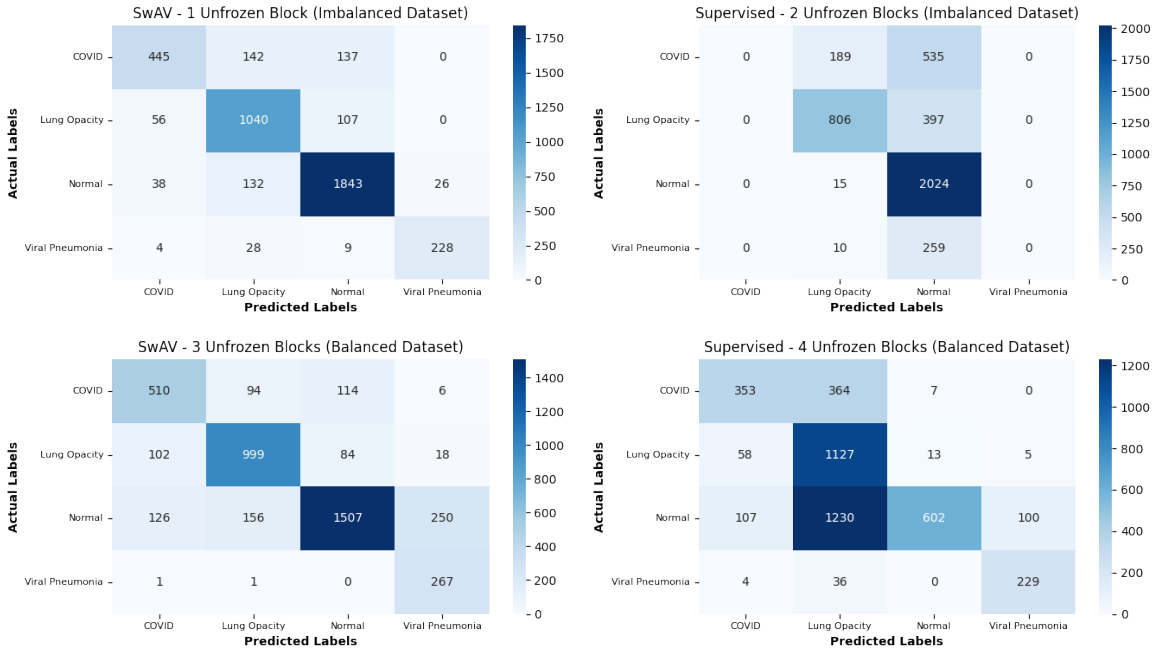


FIGURE 5. Confusion matrices of the best models for each scenario

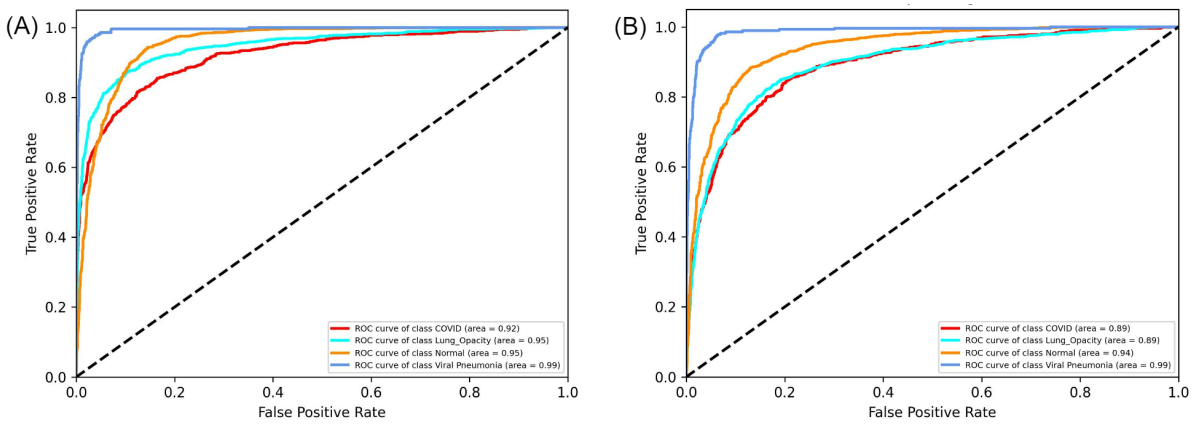


FIGURE 6. ROC curves of the best (A) SwAV-TL and (B) supervised models

explains the low TPR in Table 2 and is an expected behavior for the model as the imbalance is severe. On the contrary, the SwAV-TL model still managed to correctly classify most of the images despite the imbalanced training dataset, further proving the idea that models trained using SwAV may be more robust to data imbalance on downstream tasks. When trained on the balanced dataset, the supervised model managed to make more correct predictions for COVID, lung opacity, and viral pneumonia cases. However, its TPR plummets as a lot of false negatives

for the normal class emerged. Such occurrences may be attributed to the sampling bias and the insufficient number of data as the balanced dataset only contained 25.43% data from the original dataset. More details on the classification metrics of the models are listed in Table 4, which shows that the SwAV-TL model trained on the imbalanced dataset is the overall best model with 0.821 F1 score. Evaluation results of the supervised model trained on the imbalanced dataset were not included as the precisions were null for the COVID and Viral Pneumonia classes. Figure 6 visualizes the ROC curves for the best SwAV-TL (trained on the imbalanced dataset) and supervised (trained on the balanced dataset) models.

TABLE 4. Evaluation results of the best models.

| **Model** | **Dataset** | **Precision** | **Recall / TPR** | **F1** | **AUROC** |
|---|---|---|---|---|---|
| SwAV - 1 Unfrozen Block | Imbalanced | **0.843** | 0.808 | **0.821** | **0.952** |
| SwAV - 3 Unfrozen Blocks | Balanced | 0.717 | **0.817** | 0.744 | 0.948 |
| Supervised - 4 Unfrozen Blocks | Balanced | 0.685 | 0.643 | 0.587 | 0.928 |

Although the balanced dataset increased the number of TP for COVID and viral pneumonia classes, the TP of the other 2 classes got reduced for the SwAV-TL models. As the model still obtained good results even when trained on the imbalanced dataset, it can be inferred that the key factor of performance degradation here is the smaller number of training samples. In other words, models that are transferred after being trained using SwAV may be able to achieve better results when trained on larger datasets, even if the datasets are imbalanced. However, the limits for the imbalance severity have yet to be determined, which may be further studied in future studies. All in all, this method may bring positive impacts for other medical image classification tasks where the dataset is imbalanced and augmenting the data is not preferable due to hardware limitations, or when the imbalance is so severe that too much augmentation is required to balance the data distribution.

**4.5. Discussion.** While the best SwAV-TL model managed to obtain 0.952 AUROC, there were still some false negatives, especially in the COVID class. This means that the model can still be improved. As the images were not enhanced during the pre-processing and only

TABLE 5. Comparison of the best SwAV-TL model with previously developed CNNs using the same dataset.

| Reference | Number of Classes and Samples Used | Resampling | Method | Accuracy | F1 | AUROC |
|---|---|---|---|---|---|---|
| [23] | COVID: 3615<br>Normal: 3647 | Undersampling on the train and test sets | ResNet-50 pre-trained using SwAV on the iNat2021 Mini dataset | 0.9917, 0.8824 for the standard supervised ResNet-50 | 0.99175, 0.8232 for the standard supervised ResNet-50 | - |
| [18] | COVID: 423<br>Non-COVID Pneumonia: 423<br>Healthy: 423 | Undersampling on the train, validation, and test sets | CheXNet | 0.9774 | 0.9661 | - |
| [47] | COVID: 3616<br>Lung Opacity: 6012<br>Normal: 10192<br>Viral Pneumonia: 1345 | - | Keras InceptionV3 with 2 hidden layers | 0.9972 | - | - |
| [48] | COVID: 3616<br>Lung Opacity: 6012<br>Normal: 10192<br>Viral Pneumonia: 1345 | Augmentation to oversample | EfficientNetB1 with 1 hidden dense, batch normalization, and dropout layers, respectively | 0.9613 | 0.975 | - |
| Ours | COVID: 3616<br>Lung Opacity: 6012<br>Normal: 10192<br>Viral Pneumonia: 1345 | - | ResNet-50 TL using SwAV | - | 0.821 | 0.952 |

horizontal flipping was used to augment the data, such methods may further be explored along with the SwAV-TL models or other CNN architectures pre-trained using SwAV in future studies. However, it should be noted that most similar studies utilizing this dataset did not use the four available classes [23, 18]. The ”Lung Opacity” class was rarely used, and a lot of previous

studies that utilized the COVID-19 Radiography Database combined it with X-Ray images from other datasets to allow DL models to learn from more data and enhance their accuracy. In studies that include usages of lung opacity classes, some performance degradations may be noticeable, which are signified by the reduction in F1 scores as shown in Table 5.

Table 5 summarizes the comparison of the SwAV-TL model with models deployed in previous studies using the COVID-19 radiography database. It can be seen that the results obtained by the SwAV-TL model are inferior to all of the listed studies. However, it should be noted that the datasets were configured differently, where some even undersampled the test set [23, 18] which may be affected by sampling bias. Additionally, the different proportions of the dataset subsets will produce different results due to the larger number of training data. In one of the cited study, 70% of the images in the dataset was used for training the models and the former was further augmented to handle the data imbalance, which yield an astounding F1 score of 0.9275 [48]. In one study that is similar to this research, no further details were provided regarding the proportions or model architectures albeit the authors stated that no augmentation was performed in one of the experiment scenarios, which results were included in Table 5. To summarize, the proposed SwAV-TL model is inferior to the ones in previous studies, but further experiments are still required as slightly different training configurations can greatly affect the model. In future studies, TL using SwAV can be further tested by training the models on oversampled train sets and more modifications on the model may be conducted as no modifications nor hidden layers were implemented in this experiment. Specifically, model compression methods can also be considered to be adopted in future studies to improve the efficiency of the models in deployment stages.

## 5. CONCLUSION

Overall, TL using models pre-trained through SwAV had brought positive impacts for the task of classifying chest X-Ray images even when the available datasets are either imbalanced or small. The ResNet-50 models used in this study had proven to be more robust to the severe data imbalance and attained a great AUROC value when pre-trained using SwAV, proving that it is superior to standard supervised pre-training. Even though TL using SwAV pre-training allowed the models to perform better, further experiments are required to discover to what extent it can

improve the accuracy. The experiments conducted in this research are limited to the training of ResNet-50 without resampling and with undersampling. In the future, oversampling with augmentations can be tested on TL using SwAV, and smaller models may be deployed as large models such as the ResNet-50 require massive volumes of data.

## ACKNOWLEDGEMENT

The authors express their utmost appreciation and gratitude to Faiz Ayyas Munawwar for assisting in making the illustrations presented in this paper as well as Faisal Asadi, Rudi Nirwantono, and Gokma Sahat Tua Sinaga for assisting in the research.

## SOURCE OF FUNDING

This study is funded by Directorate General of Higher Education, Ministry of Education, Culture, Research, and Technology, Indonesia as a part of 2021 Applied Excellent Research in Higher Education Grant Number 163/E4.1/AK.04.PT/2021.

## CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

## REFERENCES

[1] N.S. Elbarbary, T.J. Santos, C. Beaufort, et al. COVID-19 outbreak and pediatric diabetes: Perceptions of health care professionals worldwide, Pediatr. Diabetes. 21 (2020), 1083–1092. https://doi.org/10.1111/pedi.13084.

[2] M. Zantah, E. Dominguez Castillo, R. Townsend, et al. Pneumothorax in COVID-19 disease-incidence and clinical characteristics, Respir. Res. 21 (2020), 236. https://doi.org/10.1186/s12931-020-01504-y.

[3] F.C. Fang, S.N. Naccache, A.L. Greninger, The laboratory diagnosis of coronavirus disease 2019-frequently asked questions, Clinic. Infect. Dis. 71 (2020), 2996–3001. https://doi.org/10.1093/cid/ciaa742.

[4] M. Elgendi, M.U. Nasir, Q. Tang, et al. The performance of deep neural networks in differentiating chest X-rays of COVID-19 patients from other bacterial and viral pneumonias, Front. Med. 7 (2020), 550. https://doi.org/10.3389/fmed.2020.00550.

[5] R. Aljondi, S. Alghamdi, Diagnostic value of imaging modalities for COVID-19: Scoping review, J. Med. Internet Res. 22 (2020), e19673. https://doi.org/10.2196/19673.

[6] A.H. Elmokadem, D. Bayoumi, S.A. Abo-Hedibah, et al. Diagnostic performance of chest CT in differentiating COVID-19 from other causes of ground-glass opacities, Egypt. J. Radiol. Nuclear Med. 52 (2021), 12. https://doi.org/10.1186/s43055-020-00398-6.

[7] H.X. Bai, B. Hsieh, Z. Xiong, et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT, Radiology. 296 (2020), E46–E54. https://doi.org/10.1148/radiol.2020200823.

[8] J.P. Kanne, B.P. Little, J.H. Chung, et al. Essentials for radiologists on COVID-19: An update—radiology scientific expert panel, Radiology. 296 (2020), E113–E114. https://doi.org/10.1148/radiol.2020200527.

[9] E.F. Ohata, G.M. Bezerra, J.V.S. das Chagas, et al. Automatic detection of COVID-19 infection using chest X-ray images through transfer learning, IEEE/CAA J. Autom. Sinica. 8 (2021), 239–248. https://doi.org/10.1109/JAS.2020.1003393.

[10] R. Rahutomo, A.S. Perbangsa, Y. Lie, et al. Artificial intelligence model implementation in web-based application for pineapple object counting, in: 2019 International Conference on Information Management and Technology (ICIMTech), IEEE, Jakarta/Bali, Indonesia, 2019: pp. 525–530. https://doi.org/10.1109/ICIMTech.2019.8843741.

[11] B. Pardamean, H.H. Muljo, T.W. Cenggoro, et al. Using transfer learning for smart building management system, J. Big Data. 6 (2019), 110. https://doi.org/10.1186/s40537-019-0272-6.

[12] J. Zhu, B. Shen, A. Abbasi, et al. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs, PLoS ONE. 15 (2020), e0236621. https://doi.org/10.1371/journal.pone.0236621.

[13] M.H. Hesamian, W. Jia, X. He, et al. Deep learning techniques for medical image segmentation: Achievements and challenges, J. Digit. Imaging. 32 (2019), 582–596. https://doi.org/10.1007/s10278-019-00227-x.

[14] H.H. Muljo, B. Pardamean, K. Purwandari, et al. Improving lung disease detection by joint learning with COVID-19 radiography database, Commun. Math. Biol. Neurosci. 2022 (2022), 1. https://doi.org/10.28919/cmbn/6838.

[15] M. Khushi, K. Shaukat, T.M. Alam, et al. A comparative performance analysis of data resampling methods on imbalance medical data, IEEE Access. 9 (2021), 109960–109975. https://doi.org/10.1109/access.2021.3102399.

[16] M.M. Rahman, D.N. Davis, Addressing the Class Imbalance Problem in Medical Datasets, Int. J. Mach. Learn. 3 (2013), 224–228. https://doi.org/10.7763/ijmlc.2013.v3.307.

[17] M.F. Aslan, M.F. Unlersen, K. Sabanci, et al. CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection, Appl. Soft Comput. 98 (2021), 106912. https://doi.org/10.1016/j.asoc.2020.106912.

[18] M.E.H. Chowdhury, T. Rahman, A. Khandakar, et al. Can AI help in screening viral and COVID-19 pneumonia?, IEEE Access. 8 (2020), 132665–132676. https://doi.org/10.1109/access.2020.3010287.

[19] M.Z. Islam, M.M. Islam, A. Asraf, A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, Inform. Med. Unlocked. 20 (2020), 100412. https://doi.org/10.1016/j.imu.2020.100412.

[20] R. Mohammed, J. Rawashdeh, M. Abdullah, Machine learning with oversampling and undersampling techniques: Overview study and experimental results, in: 2020 11th International Conference on Information and Communication Systems (ICICS), IEEE, Irbid, Jordan, 2020: pp. 243–248. https://doi.org/10.1109/ICICS49469.2020.239556.

[21] A. Jaiswal, A.R. Babu, M.Z. Zadeh, et al. A survey on contrastive self-supervised learning, Technologies. 9 (2020), 2. https://doi.org/10.3390/technologies9010002.

[22] M. Caron, I. Misra, J. Mairal, et al. Unsupervised learning of visual features by contrasting cluster assignments, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 9912–9924.

[23] M.B. Hossain, S.M.H.S. Iqbal, M.M. Islam, et al. Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images, Inform. Med. Unlocked. 30 (2022), 100916. https://doi.org/10.1016/j.imu.2022.100916.

[24] M. Loey, F. Smarandache, N.E. M. Khalifa, Within the lack of chest COVID-19 X-ray dataset: A novel detection model based on GAN and deep transfer learning, Symmetry. 12 (2020), 651. https://doi.org/10.3390/sym12040651.

[25] M.M. Rahaman, C. Li, Y. Yao, et al. Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches, J. X-Ray Sci. Technol. 28 (2020), 821–839. https://doi.org/10.3233/xst-200715.

[26] S. Minaee, R. Kafieh, M. Sonka, et al. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning, Med. Image Anal. 65 (2020), 101794. https://doi.org/10.1016/j.media.2020.101794.

[27] V. Chouhan, S.K. Singh, A. Khamparia, et al. A novel transfer learning based approach for pneumonia detection in chest X-ray images, Appl. Sci. 10 (2020), 559. https://doi.org/10.3390/app10020559.

[28] S. Misra, S. Jeon, S. Lee, et al. Multi-channel transfer learning of chest X-ray images for screening of COVID-19, Electronics. 9 (2020), 1388. https://doi.org/10.3390/electronics9091388.

[29] B.A. Sangeroki, T.W. Cenggoro, A fast and accurate model of thoracic disease detection by integrating attention mechanism to a lightweight convolutional neural network, Procedia Computer Sci. 179 (2021), 112–118. https://doi.org/10.1016/j.procs.2020.12.015.

[30] F. Liu, Y. Tian, F.R. Cordeiro, et al. Self-supervised Mean Teacher for Semi-supervised Chest X-Ray Classification, in: C. Lian, X. Cao, I. Rekik, X. Xu, P. Yan (Eds.), Machine Learning in Medical Imaging, Springer International Publishing, Cham, 2021: pp. 426–436. https://doi.org/10.1007/978-3-030-87589-3_44.

[31] S. Azizi, B. Mustafa, F. Ryan, et al. Big self-supervised models advance medical image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3478–3488.

[32] A. Abbas, M.M. Abdelsamea, M.M. Gaber, 4S-DT: Self-supervised super sample decomposition for transfer learning with application to COVID-19 detection, IEEE Trans. Neural Netw. Learn. Syst. 32 (2021), 2798–2808. https://doi.org/10.1109/tnnls.2021.3082015.

[33] M. Gazda, J. Plavka, J. Gazda, et al. Self-supervised deep convolutional neural network for chest X-ray classification, IEEE Access. 9 (2021), 151972–151982. https://doi.org/10.1109/access.2021.3125324.

[34] T. Rahman, A. Khandakar, Y. Qiblawey, et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images, Computers Biol. Med. 132 (2021), 104319. https://doi.org/10.1016/j.compbiomed.2021.104319.

[35] B. Pardamean, T.W. Cenggoro, R. Rahutomo, et al. Transfer learning from chest X-ray pre-trained convolutional neural network for learning mammogram data, Procedia Computer Sci. 135 (2018), 400–407. https://doi.org/10.1016/j.procs.2018.08.190.

[36] N. Dominic, Daniel, T.W. Cenggoro, et al. Transfer learning using inception-ResNet-v2 model to the augmented neuroimages data for autism spectrum disorder classification, Commun. Math. Biol. Neurosci. 2021 (2021), 39. https://doi.org/10.28919/cmbn/5565.

[37] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data. 3 (2016), 9. https://doi.org/10.1186/s40537-016-0043-6.

[38] J. Deng, W. Dong, R. Socher, et al. ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, 2009: pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

[39] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[40] G. Huang, Z. Liu, L. Van Der Maaten, et al. Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, 2017: pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

[41] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, Vol. 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 6105–6114.

[42] Suharjito, G.N. Elwirehardja, J.S. Prayoga, Oil palm fresh fruit bunch ripeness classification on mobile devices using deep learning approaches, Computers Electron. Agric. 188 (2021), 106359. https://doi.org/10.1016/j.compag.2021.106359.

[43] I.W. Harsono, S. Liawatimena, T.W. Cenggoro, Lung nodule detection and classification from Thorax CT-scan using RetinaNet with transfer learning, J. King Saud Univ. - Computer Inform. Sci. 34 (2022), 567–577. https://doi.org/10.1016/j.jksuci.2020.03.013.

[44] S.V. Stehman, Selecting and interpreting measures of thematic classification accuracy, Remote Sensing Environ. 62 (1997), 77–89. https://doi.org/10.1016/S0034-4257(97)00083-7.

[45] O.S. Sushkova, A.A. Morozov, A.V. Gabova, et al. A statistical method for exploratory data analysis based on 2D and 3D area under curve diagrams: Parkinson's disease investigation, Sensors. 21 (2021), 4700. https://doi.org/10.3390/s21144700.

[46] J. Bergstra, D. Yamins, D.D. Cox, et al. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms, in: Proceedings of the 12th Python in science conference, Vol. 13, Citeseer, 2013, p. 20.

[47] Z. Saeed, M.U. Khan, A. Raza, et al. Classification of pulmonary viruses X-ray and detection of COVID-19 based on invariant of inception-V 3 deep learning model, in: 2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), IEEE, Quetta, Pakistan, 2021: pp. 1–6. https://doi.org/10.1109/ICECube53880.2021.9628338.

[48] E. Khan, M.Z.U. Rehman, F. Ahmed, et al. Chest X-ray classification for the detection of COVID-19 using deep learning techniques, Sensors. 22 (2022), 1211. https://doi.org/10.3390/s22031211.