Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2024, 2024:108
https://doi.org/10.28919/cmbn/8829
ISSN: 2052-2541

SWAV TRANSFER LEARNING AND KNOWLEDGE DISTILLATION ON CHEST X-RAY CLASSIFICATION

DANIEL[1,*], ANDREA STEVENS KARNYOTO[2], GREGORIUS NATANAEL ELWIREHARDJA[2,3], TJENG WAWAN CENGGORO[2,3], BENS PARDAMEAN[1,2]

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

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

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

Abstract: COVID-19 has severely impacted human life. In response, researchers worldwide have focused on developing advanced computer systems capable of analyzing chest X-rays. Models with the best performance have been identified through numerous studies. However, these models are large and require significant computational resources. In this study, we propose combining Swapping Assignments between Views (SwAV) Transfer Learning (TL) and Knowledge Distillation (KD). By training the teacher model ResNet-50 with SwAV TL and transferring its knowledge to smaller models like ResNet-18 and ResNet-34, the performance of the smaller models improved. The ResNet-34 model's performance increased, with accuracy increasing by 5.31%, recall by 4.05%, precision by 5.62%, F1-score by 4.93%, and AUROC by 0.85%. Similarly, the ResNet-18 model's performance improved, with accuracy increasing by 1.52%, recall by 1.03%, precision by 1.87%, F1-score by 1.46%, and AUROC by 0.29%. Therefore, it has been demonstrated that the combination of SwAV TL and KD can effectively transfer knowledge from larger model to smaller ones, resulting in improved performance in the smaller models.

Keywords: chest x-ray; swapping assignments between views; knowledge distillation; transfer learning; classification.

2020 AMS Subject Classification: 68T01, 68T05, 68T07, 68T45, 68P30.

[*]Corresponding author
E-mail address: daniel017@binus.ac.id

Received August 11, 2024

1. INTRODUCTION

In December 2019, the world's attention turned to a new respiratory disease that quickly spread across the globe. Scientists figured out that it was caused by a virus called the SARS-CoV-2 virus. Common clinical signs of COVID-19 were high temperature, cough, sore throat, muscle aches, headaches, and absence of smell. In severe cases, the disease caused respiratory failure, multi-organ dysfunction, and death [1]. People with underlying conditions such as cancer [2], cardiovascular [3], and diabetes [4] were more affected to COVID-19 infection, and their mortality rate was increased. COVID-19 had affected the entire world, impacted not only the healthcare sector but also extended to other areas like the economy and education.

One way to address this condition was by conducting screening tests as efficiently as possible. The Reverse Transcription Polymerase Chain Reaction (RT-PCR) test was the primary diagnostic method for this disease in clinical settings, but it was less sensitive to certain types of COVID-19 [5]. Radiology imaging, particularly chest X-rays, could detect irregularities in the lung tissue of COVID-19 patients [6]. This method was cost effective and widely used to assess lung diseases. However, the manual process of checking these images by expert was time-consuming, which required a computer system to assist in the analysis.

Convolutional Neural Networks (CNN) were utilized as the primary method for detecting COVID-19 using Deep Learning (DL). CNN had the capability to extract important features such as medical images and hybrid support vector regression, which made it highly effective for classification tasks [7], [8]. Furthermore, this approach demonstrated potential efficacy in the detection and diagnosis of other critical medical conditions, such as lung cancer [9]. However, this approach required a substantial amount of data, resulted in large models, and demanded high computational power [10].

Within the domain of DL, producing smaller models with good performance remained a challenge. Knowledge Distillation (KD) addressed this by transferring insights from a more complex model (teacher) to a more compact model (student), thereby enabling the student to achieve enhanced performance [10]. Additionally, Swapping Assignments between Views (SwAV) Transfer Learning (TL) improved feature learning by leveraging unlabeled data and comparing various views of the same image [11]. SwAV could exceeded the performance of supervised learning as it utilized more general feature representations of data without the need for explicit labels [11]. Combining KD and SwAV TL allowed for the creation of lightweight models with good performance, aided by larger models. This is supported by the improvement of the student model after KD training [12], and training with SwAV [13] resulted in better performance compared to supervised learning.

The contributions of this research are: 1) integrating SwAV TL with KD to enhance the performance of smaller models by leveraging knowledge from larger models, and 2) optimizing model efficiency in resource-constrained environments using combination of SwAV TL and KD.

This study is organized as follows: Section 2 offers a review of prior studies on COVID-19 radiography classification. Section 3 described the methodology in detail, including data collection, model training, and assessment criteria. Section 4 presented the results of the experiments and analyzed their implications, Section 5 reviewed this study in relation to previous lessons, Section 6 addresses potential future work, and Section 7 discusses the conclusions drawn from this study.

2. Related Works

The global spread of COVID-19 encouraged researchers to develop more effective and efficient methods for its detection. At that time, many studies focused on utilizing deep learning technology to achieve this goal. One of them, Apostolopoulos and Mpesiana utilized five pre-trained deep learning models (VGG-19, MobileNetV2, Xception, Inception, and InceptionResNetV2) in their study. It was demonstrated in their study that MobileNetV2 achieved the highest performance, with an accuracy of 0.9472 [14]. Similarly, Jaiswal et al. evaluated DenseNet-201, VGG-16, InceptionResNetV2, and ResNet152V2, and found that DenseNet-201 reached an accuracy of 0.9625, with a recall and precision both measured at 0.9629.[15].

Brima et al. proposed ResNet-50 for classifying four types of lung diseases, and it reached an accuracy of 0.9385 [16]. In another study, rotation, horizontal flipping, and vertical flipping were used as augmentations, while EfficientNetB1 was utilized for feature extraction and a multilayer perceptron as the classifier, resulting in an accuracy of 0.9613 and an F_1-Score of 0.975 [17]. Additionally, Sitaula and Hossain detected COVID-19 by combining VGG16 with an attention module. The data were pre-processed and augmented using zoom range, shear range, horizontal flip, vertical flip, rotation range, and rescale. Then, the data was processed with the attention module and used to train the VGG-16 model. This combination achieved an accuracy of 0.8749 [18].

In addition to supervised learning, a significant alternative approach was Self-Supervised Learning (SSL) to improve model performance. Park et al. proposed an SSL with Attention Mechanism to detect COVID-19, utilizing a U-shaped CNN and a Convolution Block Attention

Module (CBAM) to improve the baseline model's accuracy by 0.008 [19]. In another study, Muljo et al. utilized SwAV TL to enhance ResNet-50's performance when working with imbalanced data [13]. Similarly, A pre-trained CNN model with SSL was evaluated by Gazda et al. for classifying COVID-19 and various types of pneumonia across different datasets. The study found that their method maintained consistent results regardless of variations in training data size [20].

Many studies utilized pre-trained models, which were typically large and required high computational power. Researchers sought techniques to develop lightweight models that had good performance. Shi et al proposed Densenet-169 combined with attention modules and KD. The attention modules were integrated into the teacher model to extract general features and concentrate on infected regions, while the student network focused on the irregularly shaped lesion areas. The student model reached an accuracy of 0.9344 [21]. In another study, Kabir et al. utilized the KD technique to pass insights from a bigger CNN model to a smaller, resulting in the student model attaining an accuracy of 0.9608, precision of 0.9448, and recall of 0.9776. [22]. PneuKnowNet detected COVID-19 by utilizing the KD technique and chest X-ray images that had been analyzed by radiologists. This training approach improved the performance of PneuKnowNet by 0.023 compared to the baseline model [12]. However, previous studies have not fully explored the combination of SwAV TL and KD for COVID-19 classification.

3. RESEARCH METHODOLOGY

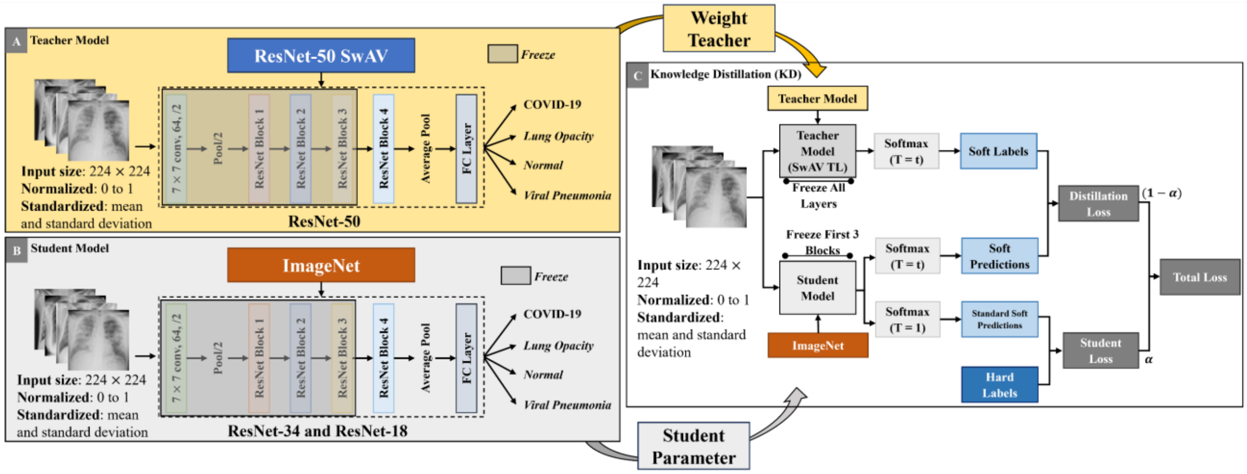

Figure 1. Illustration of Experimental Process: (A) Teacher Model, (B) Student Model, and (C) Knowledge Distillation (KD)

The workflow of this study began with acquiring and pre-processing the images for model training. The teacher model, ResNet-50, was pretrained using SwAV (see Figure 1A), whereas the student models, ResNet-34 and ResNet-18, were pretrained on the ImageNet dataset (see Figure 1B). Subsequently, the best weights from the teacher model were selected for transfer during the KD training process. Additionally, the most effective parameter configurations for the student models were identified for implementation during KD. The Student models were trained using the KD technique, as illustrated in Figure 1C. Finally, an evaluation was conducted to compare the performance of the model.

3.1. RESNET

ResNet was an improvement in deep neural networks designed to solve problems like vanishing gradients [23] and degradation problem [24] in deep models, which could prevent model optimization. To address these issues, shortcut connections [25] were used as direct links that bypassed one or more layers within the network. This bypass allowed the network to learn residual functions related to the layer inputs, rather than having to learn functions without direct references. Mathematically, it was illustrated as follows:

$$y = \mathrm{F}(x, \{W_i\}) + x \tag{1}$$

where x and y represent the input and output vectors of the layers, W_i is the weights of the layer in the residual block, and $\mathrm{F}(x, \{W_i\})$ represents the residual mapping function that needs to be learned.

Shortcut connections thus not only simplified the training of deep networks but also enabled the construction of networks with a significantly larger number of layers, leading to better performance on complex tasks. By leveraging the advantages of ResNet, we adopted ResNet-50 as the teacher model, with a size of 92.18 MB, and ResNet-18 and ResNet-34, with sizes of 43.75 MB and 83.29 MB respectively, as the student models.

3.2. TRANSFER LEARNING

Transfer learning (TL) was a model training technique that leveraged knowledge from a model previously trained on another dataset. The weights from the pre-trained model were applied to set up a new model that was then trained on the target dataset. This approach could accelerate the training process, improve model performance, and decrease the dependency on extensive training data. In mathematical terms, this process could be described as follows:

$$\begin{gathered} D_S = \{(x_{S1}, y_{S1}), (x_{S2}, y_{S2}), \dots, (x_{Sn}, y_{Sn})\}, T_S = \{y_S, f(.)\} \\ D_T = \{(x_{T1}, y_{T1}), (x_{T2}, y_{T2}), \dots, (x_{Tn}, y_{Tn})\}, T_T = \{y_T, f(.)\} \end{gathered} \tag{2}$$

where D_S is the source domain, T_S is the source task, D_T is the target domain, and T_T is target task, x is the data input, y is the target label, and $f(.)$ is the predictive function that utilizes the transferred knowledge from D_S and T_S, with the condition that $D_S \neq D_T$ or $T_S \neq T_T$ [26]. This technique has been proven in many studies [27], [28], [29], [30].

3.3. Swapping Assignments between Views (SwAV)

SwAV was a subset of SSL where the model was trained using unlabeled data to learn various representations by comparing features of images that had been altered through augmentation techniques [11]. SSL has been proven effective in lung disease classification [13], [31], [32]. As shown in Figure 2, SwAV applied various augmentation techniques, such as cropping, rotation, resizing, and color changes, to each input image (X), resulting in two different views ($X1$ and $X2$). The augmented views were mapped through the model to produce outputs $Z1$ and $Z2$. These outputs were computed using a dot product with the prototype vector C, which represents the weights of a layer with linear activation, to produce $Z1.C$ and $Z2.C$. The results were processed with the Sinkhorn-Knopp algorithm to produce $Q1$ and $Q2$. $Q1$ and $Q2$ were swapped and the loss was calculated based on the swapped results using the Cross-Entropy (CE) Loss. The CE Loss was utilized for backpropagation to adjust the model parameters and prototype vector C. The loss obtained during training was calculated using the equation below:

$$P_t^{(k)} = \frac{\exp\left(\frac{1}{\tau}(Z_t.C_k)\right)}{\sum_{k'} \exp\left(\frac{1}{\tau}(Z_t.C_{k'})\right)} \tag{3}$$

$$l(z_t, q_s) = -\sum_k q_s^{(k)} \log\left(P_t^{(k)}\right) \tag{4}$$

$$L(z_t, z_s) = l(z_t, q_s) + l(z_s, q_t) \tag{5}$$

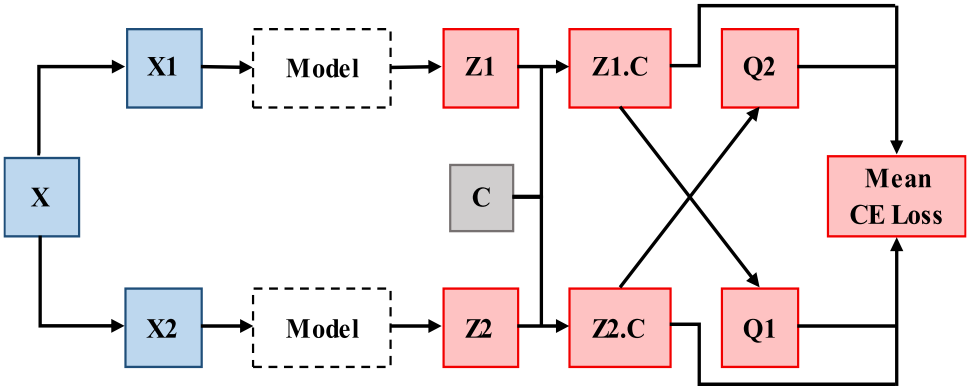

Figure 2. Process of SwAV

where k represents the value of prototype vector C and τ is a temperature parameter that controls the sharpness of the probability distribution derived from the similarity scores between the feature representations and the prototypes.

This study utilized the SwAV TL to train the teacher model due to its superior ability to recognize the same object from various views, even after transformations. With this capability, the resulting teacher model was expected to be more robust and effectively share its understanding with the student model.

3.4. Knowledge Distillation

KD was a technique for knowledge transfer that facilitated a more compact model in learning from a bigger model [10]. This technique involved the student model to replicate the behavior and predictions of the teacher model [33]. Figure 1C illustrated how this technique was performed. The teacher model, which already possessed knowledge, was trained alongside the student model on the input data. Both models were trained up to the last fully connected layer to produce raw logits. These raw logits were processed through the softmax activation function, which included a temperature parameter. The formula for the softmax temperature was shown below:

$$P(x) = \frac{\exp\left(\frac{z_c(x)}{T}\right)}{\sum_{c \epsilon C} \exp\left(\frac{z_c(x)}{T}\right)} \tag{6}$$

where $z_c(x)$ is the model that produce logits for class C and T is the temperature parameter that produced a softer probability distribution over classes. The result of this process was soft labels and soft predictions. On the other hand, the student model also performed a standard softmax activation function with a temperature of 1 to produce standard soft predictions. Then, The Kullback-Leibler Divergence between the soft labels and soft predictions was used to calculate the Distillation Loss. Additionally, the Student Loss was determined by applying standard Cross Entropy between the conventional soft labels and the hard labels. Both losses were combined with a weighting factor alpha (α). The formula of Distillation Loss, Student Loss, and Total Loss was presented below:

$$KL\ Divergence = \sum_{x \epsilon X} P_t(x) \log\left(\frac{P_s(x)}{P_t(x)}\right) \tag{7}$$

$$CE = \sum_{x \epsilon X} P_t(x) \log\big(P_t(x)\big) \tag{8}$$

$$L(P_t, P_s) = \left((1-\alpha) \times \sum_{x \epsilon X} P_t(x) \log\left(\frac{P_s(x)}{P_t(x)}\right) \right) + \left(\alpha \times \sum_{x \epsilon X} P_t(x) \log\big(P_t(x)\big) \right) \quad (9)$$

where $P_t(x)$ is the probability distribution over the classes predicted by the teacher model for a particular input x, $P_s(x)$ is the probability distribution over the classes predicted by the student model for the same input x, and α is a weighting factor that balances the contributions of the Student and Distillation Loss. This Total Loss was used to update the weights of the student model, enabling it to acquire knowledge from the teacher model. During this process, the teacher model's weights remained frozen, preventing further learning.

This study utilized the KD technique to transfer insight from the bigger model to the smaller model, resulting in an improvement in the performance of the student model. This method not only preserved the performance of the larger model but also lowered computational demands, making it ideal for use in environments with limited resources.

3.5. Dataset

The COVID-19 Radiography Database was utilized in this study [34], [35]. This database was created by researchers from Qatar, Dhaka, Pakistan, and Malaysia. The database was categorized into four classes, including 3616 COVID-19, 10192 normal, 6012 lung opacity, and 1345 viral pneumonia. Figure 3 showed an example image from each class. The COVID-19 image showed a denser fog around the lungs compared to the others, while the normal image appeared the clearest. Additionally, the dataset provided lung segmentation masks that focus the images on the lung area for all images, however this feature was not utilized.

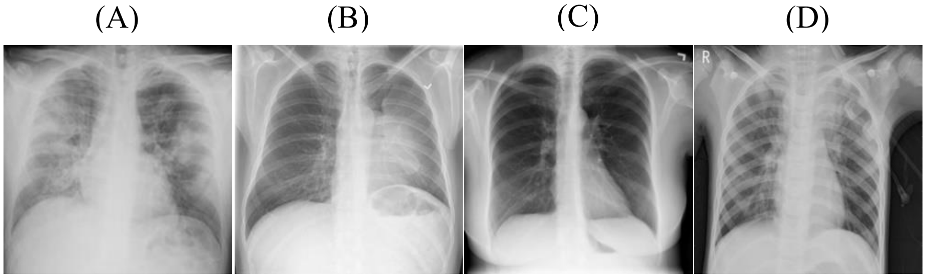

Figure 3. Examples input images for each class: (A) COVID-19, (B) Lung-Opacity, (C) Normal, (D) Viral Pneumonia

3.6. Data Preprocessing

All images in the database were adjusted to a size of 224 × 224 pixels to allow consistency. The dataset was divided into training, validation, and test sets in a 70:15:15 ratio, using seed 42 for randomization. After that, all images were scaled to float values within the range of 0 to 1 and standardized using a mean a standard deviation from ImageNet [36]. Figure 4 showed an example of a pre-processed image of each class.

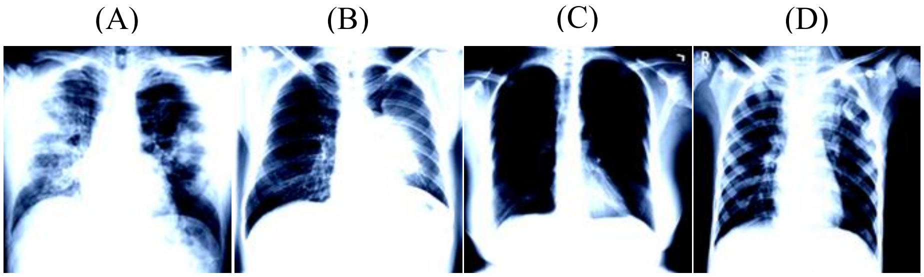

Figure 4. Examples of pre-processed input images: (A) COVID-19, (B) Lung-Opacity, (C) Normal, (D) Viral Pneumonia

3.7. Model Training

The Python programming language and the PyTorch deep learning framework were used to conduct the experiments, with the pre-processed data serving as input for model training. Following the approach from the previous study, where the ResNet architecture consisted of 4 blocks, the first 3 blocks were frozen and only the last block was trained [13]. Training for 50 epochs was carried out on all the models using a mini-batch size of 64 images, with the Cross Entropy loss and Adam optimizer. The model with the lowest validation loss was identified as the best during this training. Model training was carried out in two stages. Stage one involved training the teacher model using SwAV TL and the student model using supervised learning. Stage two was focused on training the student model using KD.

In the first stage, we initialized the teacher's weights with SwAV and the student models' weights with ImageNet. Hyperparameter tuning was performed during model training to determine the optimal performance for learning rate lr of 1e-2, 1e-3, 1e-4, 1e-5 and weight decay λ of 0, 1e-1, 1e-2, 1e-3, 1e-4. With this combination, there were 20 training session for each model. This process produced the best-performing teacher model weights, which would be applied to the teacher model in KD training, and also created the optimal student model configuration for KD training.

Moving to the next stage, the KD technique was utilized to train the student model. In this stage, the best teacher model weights from the previous training were transferred to the teacher model, while pre-trained weights from the ImageNet dataset were used to initialize the student model. This training was also conducted using the best student model parameters from the previous training, with additional hyperparameter tuning for the temperature T of 2, 4, 6, 8 and alpha α of 0.1, 0.3, 0.5, 0.7, 0.9. This stage involved 20 training sessions for each student model. Finally, the best results from the student model training in both stages were evaluated.

3.8. Evaluation Metrics

To evaluate this study, we used several important metrics in this field, such as accuracy, recall, precision, and F_1-score.

4. Results

4.1. Stage 1 Training Results

Figure 5 showed the lowest training and validation loss of the teacher and student models in Stage 1. There was a significant gap between the training and validation loss, especially in the ResNet-50 SwAV TL and ResNet-34, which experienced spikes. This condition indicated that the models were overfitting.

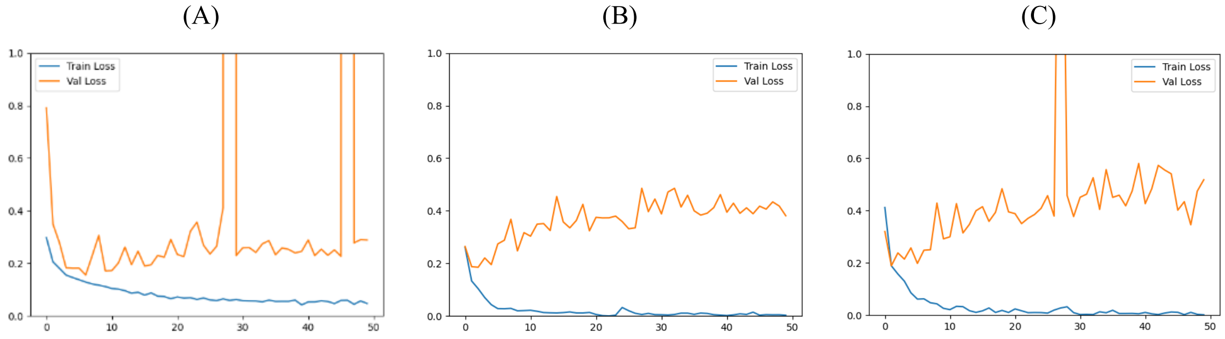

Figure 5. Training and validation loss of the best teacher and student models on stage 1: (A) ResNet-50 SwAV TL, (B) ResNet-18 Supervised, ResNet-34

In the training of the ResNet-50 SwAV TL, the training loss showed a gradual decrease from the early epoch to the final epoch. However, the validation increased significantly during certain epochs. The validation loss began at a high value and steadily decreased until epoch 8, then it fluctuated up and down until epoch 29. A significant spike occurred at epoch 30, followed by a decrease, and another spike was observed at epoch 48. The lowest validation loss was 0.15572, with a learning rate lr of 0.01 and a weight decay λ of 0.0001. The teacher model weights with this configuration were used by the teacher model in KD training.

During the training of the ResNet-18 Supervised, the validation loss increased from the first to the last epoch, while the training loss significantly decreased to near zero. The lowest validation loss was 0.1852, with a learning rate lr of 0.001 and a weight decay λ of 0. This configuration would be used in the training of the ResNet-18 student model with KD.

While training the ResNet-34 Supervised, it showed a pattern similar to the ResNet-50 SwAV TL, with the training loss decreased from the first epoch to the end of the epochs. However, the validation loss increased inconsistently, with a spike occurred at epoch 29. The lowest validation loss was 0.1889, with a learning rate lr of 0.01 and a weight decay λ of 0. This configuration would also be used in the training of the ResNet-34 student model with KD.

4.2. Stage 2 Training Results

Figure 6 showed the training and validation loss for the best-performing student models on stage 2. Different from stage 1, the gap between training and validation loss appeared much smaller and more consistent. This occurred due to the regularization effect of KD, which helped the student model learn general patterns from teacher [10]. The training loss in the ResNet-18 KD decreased consistently from the first epoch to the final epoch. The validation loss decreased from the first epoch to epoch 24, then fluctuated up and down until the end. The lowest validation loss reached 0.1280 with an alpha α configuration of 0.5 and temperature T of 2.

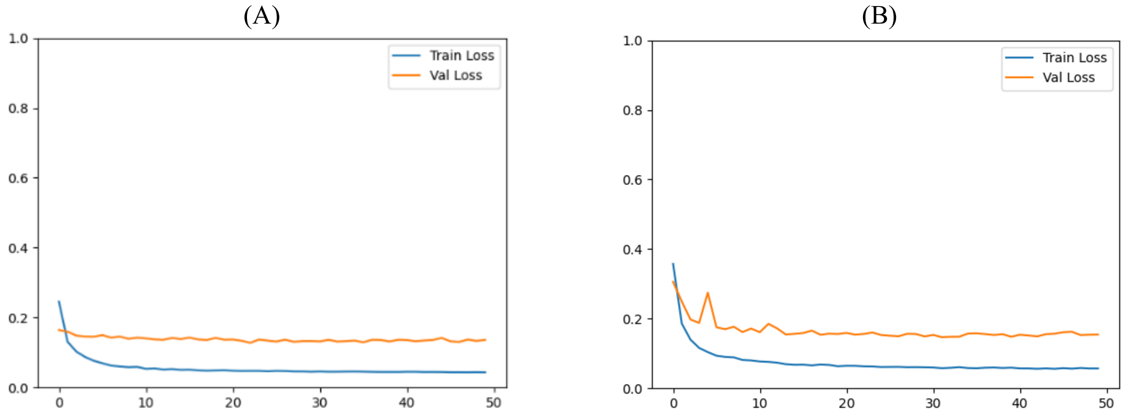

Figure 6. Training and validation loss of the best student models on stage 2: (A) ResNet-18 KD, (B) ResNet-34 KD

A similar pattern occurred during the training of the ResNet-34 KD. The difference between training and validation loss was smaller compared to previous training. Both training and validation loss decreased from the first epoch to the end. This model delivered the highest performance with an alpha α configuration of 0.3 and temperature T of 2, with a validation loss of 0.1468.

4.3. Impact of ResNet-50 SwAV TL on ResNet-18 Supervised

Table 1 showed the results of the best performance of ResNet-50 SwAV TL, ResNet-18 Supervised, and ResNet-18 KD. In general, the performance of ResNet-18 KD is higher than ResNet-18 Supervised. For accuracy, recall, precision, F_1-Score, and AUROC, the performance improved by +1.52%, +1.03%, +1.87%, +1.46%, and +0.29%, respectively. Among all metrics, precision showed the highest improvement. This indicated that the model had become better at predicting True Positive (TP), reducing the number of False Positive (FP) predictions. Furthermore, the F_1-score and AUROC values of ResNet-18 KD and ResNet-50 SwAV TL were quite close, indicating that ResNet-18 KD could effectively mimic the behavior or characteristics of ResNet-50 SwAV TL.

Table 1. Test results of ResNet-50 SwAV TL, ResNet-18, ResNet-18 KD.

Metrics	ResNet-50 SwAV TL	ResNet-18 Supervised	ResNet-18 KD
Accuracy	**0.9596**	0.9316	0.9458
Recall	**0.9604**	0.9401	0.9498
Precision	**0.9712**	0.9382	0.9558
F_1-Score	**0.9654**	0.9391	0.9528
AUROC	**0.9956**	0.9899	0.9928

Figure 7 illustrated the confusion matrices of ResNet-18 Supervised and ResNet-18 KD. An increase in TP values was observed across all classes, showing an improved of the model to accurately identify positive case. The decrease in FP values for the COVID-19 class, especially in cases where the model predicted COVID-19 but the actual condition was normal (from 13 to 5), indicated an increase in the ability of the student to distinguish COVID-19 cases. A similar reduction in FP values was also noted in the lung opacity class, especially for predictions of lung opacity that were actually normal (from 86 to 59), which indicated improved recognition of this class by the model.

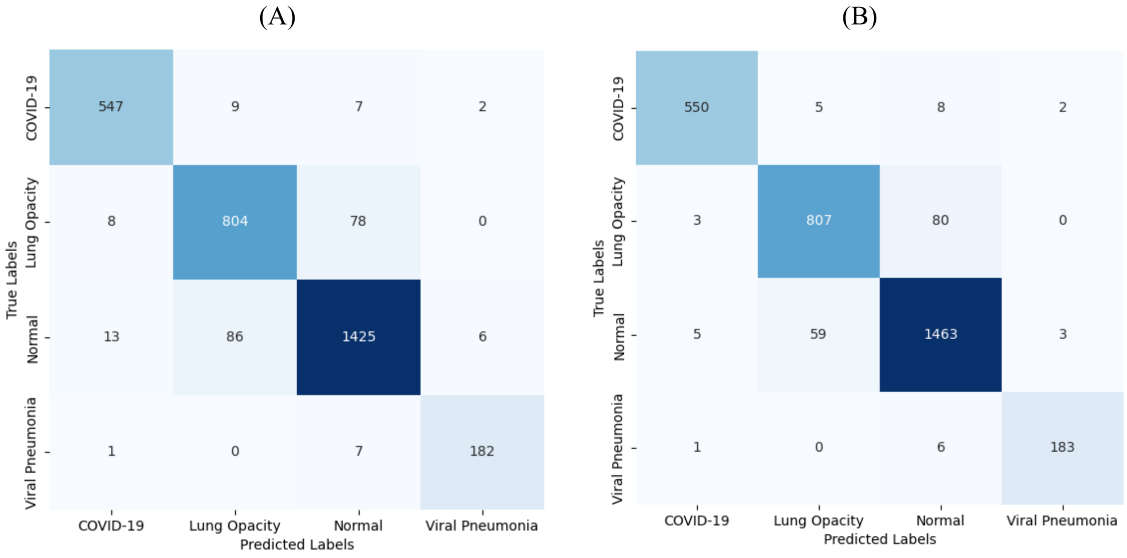

Figure 7. Confusion Matrices of: (A) ResNet-18 Supervised, (B) ResNet-18 KD

Figure 8 showed the ROC curves of ResNet-18 Supervised and ResNet-18 KD. According to the ROC curves, both models showed strong predictive ability for COVID-19 and viral pneumonia, with the ROC curve area close to 1. However, they faced some difficulty with the other two classes. The model using KD proved to be more effective in predicting lung opacity cases compared to the supervised model.

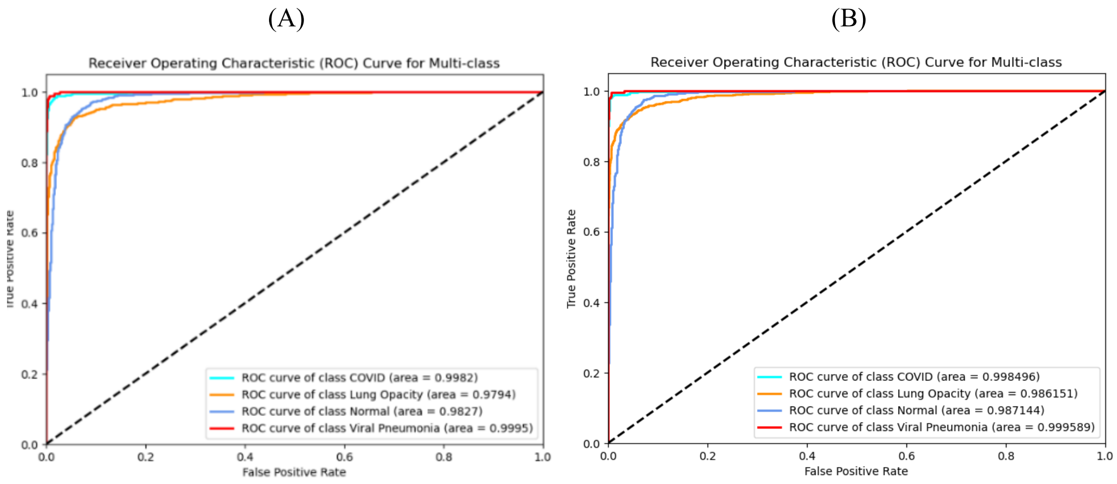

Figure 8. ROC Curves of: (A) ResNet-18 Supervised, (B) ResNet-18 KD

4.4. IMPACT OF RESNET-50 SWAV TL ON RESNET-34 SUPERVISED

Table 2 showed the results of the best performance of ResNet-50 SwAV TL, ResNet-34 Supervised, and ResNet-34 KD. The performance significantly improved with the KD technique, with accuracy increasing by +5.31%, recall by +4.05%, precision by +5.62%, F_1-score by +4.93%, and AUROC by +0.85%. ResNet-34 KD surpassed the ResNet-50 SwAV TL, with all metrics being higher, indicating that the KD technique has the potential to enhance the performance of the of the student model beyond of its teacher. This proved the student model was able to effectively acquire knowledge from the teacher model.

Table 2. Test results of ResNet-50 SwAV TL, ResNet-34, ResNet-34 KD.

Metrics	ResNet-50 SwAV TL	ResNet-34 Supervised	ResNet-34 KD
Accuracy	0.9596	0.9200	**0.9672**
Recall	0.9604	0.9299	**0.9676**
Precision	0.9712	0.9260	**0.9781**
F_1-Score	0.9564	0.9268	**0.9725**
AUROC	0.9956	0.9888	**0.9973**

Figure 9 showed the confusion matrices of ResNet-34 Supervised and ResNet-34 KD. Confusion matrices stated that TP values increased across COVID-19, lung opacity, and normal class, showing that the model got better at finding for 3 of class. The decrease FP and False Negative (FN) for the COVID-19 class indicated the model was better able to detect COVID-19 than before. The lung opacity and normal class also showed similar behavior with a reduction in both FP and FN values. On the other hand, the viral pneumonia class experienced an increase in FN value, which could have a negative impact in the healthcare field. Fortunately, the increase was not significant, and it appeared that the misdetections were originally normal class. However, the viral pneumonia class also recorded a positive impact with a decrease in FP values.

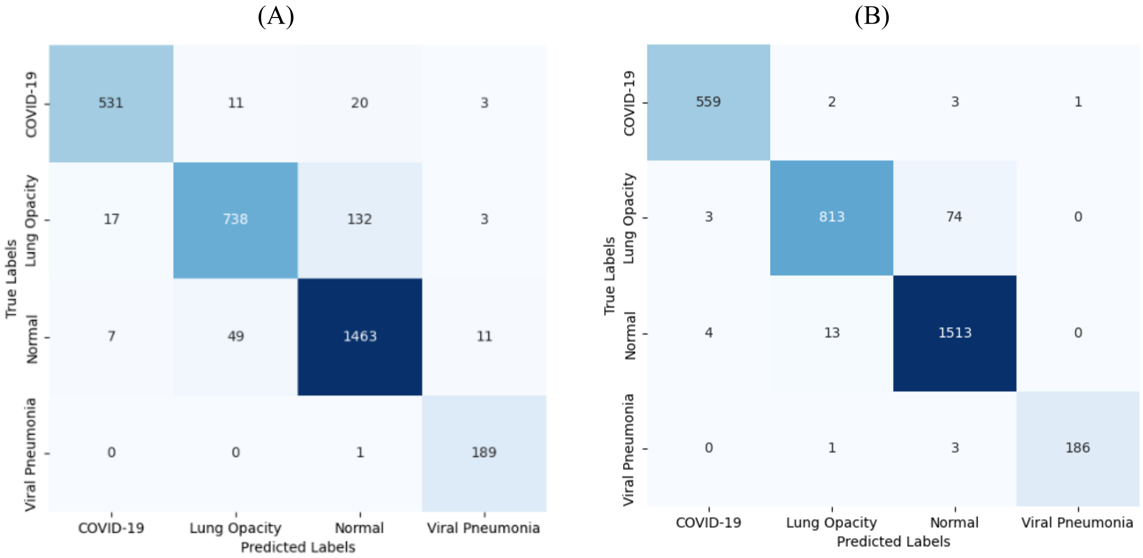

Figure 9. Confusion Matrices of: (A) ResNet-34 Supervised, (B) ResNet-34 KD.

Figure 10 showed the ROC curves of ResNet-34 Supervised and ResNet-34 KD. In general, ResNet-34 KD outperformed the supervised student model for each class. ResNet-34 Supervised struggled to accurately lung opacity and normal classes. However, this could be improved by using the KD technique, which enabled the model to predict both classes better.

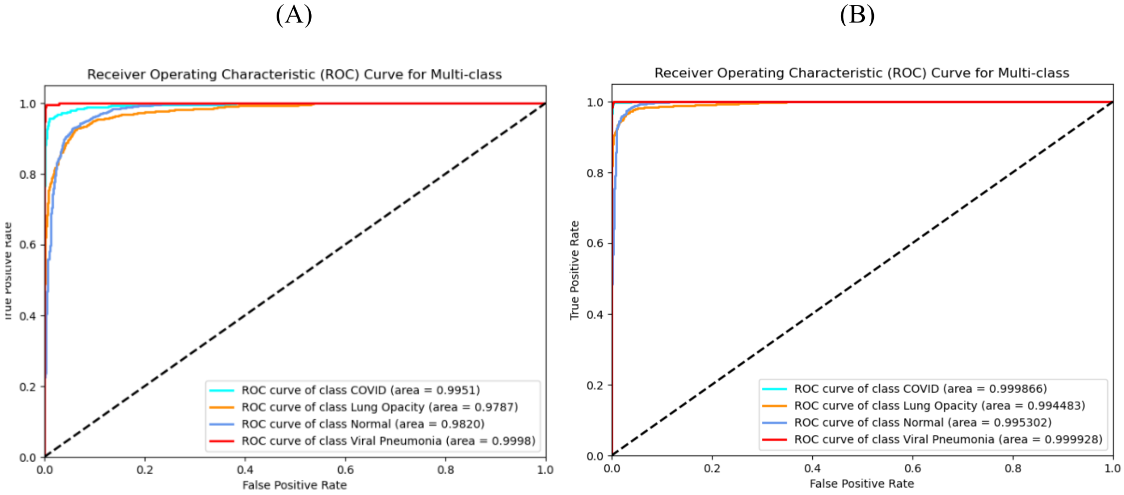

Figure 10. ROC Curves of: (A) ResNet-34 Supervised, (B) ResNet-34 KD.

5. DISCUSSION

Table 3 summarized the comparison results of ResNet-34 KD with previous study using the same data. Our result outperformed compared to all the listed studies [13], [16], [17]. However, it was important to recognize that these results were influenced by several factors. One of the factors was the distribution of proportions among the training, validation, and testing datasets, where a larger amount of training data could significantly enhance model performance. Additionally, the methods used for data processing during preparation might have impacted the results, as proper data handling could have improved the model's performance. Effective hyperparameter tuning also contributed to significant performance gains.

The EfficientNetB1 model proposed by Brima et al. was able to compete the performance of ResNet-34 KD [16]. It was noted that EfficientNetB1 had around 7.7 million parameters, which was significantly smaller compared to the 22.2 million parameters of ResNet-34. This indicated that, despite having fewer parameters, the EfficientNetB1 model still managed to deliver competitive performance, likely due to its efficient architectural design and applied optimization techniques.

Table 3. Comparison results of ResNet-34 KD with previous study using the same the data. The data used use consisted of 3616 COVID-19, 10192 normal images, 6012 lung opacity, and 1345 viral pneumonia images.

Ref.	Method	Accuracy	Recall	Precision	F1-Score	AUROC
[16]	ResNet-50	0.939	-	-	-	-
[17]	EfficientNetB1	0.961	-	0.973	0.975	-
[13]	ResNet-50 SwAV TL	-	0.808	0.843	0.821	0.925
Ours	ResNet-34 KD with the teacher ResNet-50 SwAV TL	**0.967**	**0.967**	**0.978**	**0.972**	**0.997**

6. FUTURE WORKS

For future work, this method can be applied using deeper models to explore the potential for performance improvement in image classification tasks. Further modifications, such as the addition of attention modules, can be used to guide the model in learning more accurately on specific areas of the lungs. Additionally, the use of other data sources, such as CT scans or multi-modal data, can be implemented to enhance the model's accuracy and generalization.

7. CONCLUSION

This study demonstrated that combining SwAV TL and KD effectively transfer knowledge from larger models to smaller ones, resulting in improved performance for the smaller models. SwAV TL enhanced generalization, while KD transferred knowledge from larger models to smaller ones, thereby improving the performance of the smaller models and prevented overfitting. This approach successfully developed lightweight models with strong performance by leveraging both techniques. Notably, ResNet-18 and ResNet-34 models, trained with knowledge from ResNet-50 SwAV TL, showed significant performance gains. However, the study was limited by not exploring all possible parameter combinations.

ACKNOWLEDGMENT

All models were trained using NVIDIA Tesla P100 GPU provided by NVIDIA – Bina Nusantara University Artificial Intelligence Research and Development Center.

CONFLICT OF INTERESTS

The authors affirm that there are no conflicts of interest.

REFERENCES

[1] F. Di Gennaro, D. Pizzol, C. Marotta, et al. Coronavirus diseases (COVID-19) current status and future perspectives: A narrative review, Int. J. Environ. Res. Public Health 17 (2020) 2690. https://doi.org/10.3390/ijerph17082690.

[2] O.M. Al-Quteimat, A.M. Amer, The impact of the COVID-19 pandemic on cancer patients, Amer. J. Clin. Oncol. 43 (2020), 452–455. https://doi.org/10.1097/coc.0000000000000712.

[3] J. Sabatino, S. De Rosa, G. Di Salvo, et al. Impact of cardiovascular risk profile on COVID-19 outcome. A meta-analysis, PLoS ONE 15 (2020), e0237131. https://doi.org/10.1371/journal.pone.0237131.

[4] M. Abu-Farha, F. Al-Mulla, T.A. Thanaraj, et al. Impact of diabetes in patients diagnosed with COVID-19, Front. Immunol. 11 (2020), 576818. https://doi.org/10.3389/fimmu.2020.576818.

[5] A. Tahamtan, A. Ardebili, Real-time RT-PCR in COVID-19 detection: issues affecting the results, Expert Rev. Mol. Diagnostics 20 (2020), 453–454. https://doi.org/10.1080/14737159.2020.1757437.

[6] Z. Xu, L. Shi, Y. Wang, et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome, Lancet Respir. Med. 8 (2020), 420–422. https://doi.org/10.1016/s2213-2600(20)30076-x.

[7] 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.

[8] A. Susanto, Herman, T.W. Cenggoro, et al. Transfer-learning-aware neuro-evolution for diseases detection in chest X-ray images, preprint, (2020). http://arxiv.org/abs/2004.07136.

[9] T.W. Cenggoro, B. Mahesworo, A. Budiarto, et al. Features importance in classification models for colorectal cancer cases phenotype in Indonesia, Procedia Computer Sci. 157 (2019), 313–320. https://doi.org/10.1016/j.procs.2019.08.172.

[10] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, preprint, (2015). http://arxiv.org/abs/1503.02531.

[11] M. Caron, I. Misra, J. Mairal, et al. Unsupervised learning of visual features by contrasting cluster assignments, in: Advances in Neural Information Processing Systems, 2020.

[12] D. Schaudt, R. von Schwerin, A. Hafner, et al. Leveraging human expert image annotations to improve pneumonia differentiation through human knowledge distillation, Sci. Rep. 13 (2023), 9203. https://doi.org/10.1038/s41598-023-36148-7.

[13] H.H. Muljo, B. Pardamean, G.N. Elwirehardja, et al. Handling severe data imbalance in chest X-Ray image classification with transfer learning using SwAV self-supervised pre-training, Commun. Math. Biol. Neurosci. 2023 (2023), 13. https://doi.org/10.28919/cmbn/7526.

[14] I.D. Apostolopoulos, T.A. Mpesiana, Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks, Phys. Eng. Sci. Med. 43 (2020), 635–640. https://doi.org/10.1007/s13246-020-00865-4.

[15] A. Jaiswal, N. Gianchandani, D. Singh, et al. Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning, J. Biomol. Struct. Dyn. 39 (2020), 5682–5689. https://doi.org/10.1080/07391102.2020.1788642.

[16] Y. Brima, M. Atemkeng, S.T. Djiokap, et al. Transfer learning for the detection and diagnosis of types of pneumonia including pneumonia induced by COVID-19 from chest X-ray images, Diagnostics 11 (2021), 1480. https://doi.org/10.3390/diagnostics11081480.

[17] 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.

[18] C. Sitaula, M.B. Hossain, Attention-based VGG-16 model for COVID-19 chest X-ray image classification, Appl.

Intell. 51 (2020), 2850–2863. https://doi.org/10.1007/s10489-020-02055-x.

[19] J. Park, I.Y. Kwak, C. Lim, A deep learning model with self-supervised learning and attention mechanism for COVID-19 diagnosis using chest X-ray images, Electronics 10 (2021), 1996. https://doi.org/10.3390/electronics10161996.

[20] 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.

[21] W. Shi, L. Tong, Y. Zhu, et al. COVID-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks, IEEE J. Biomed. Health Inform. 25 (2021), 2376–2387. https://doi.org/10.1109/jbhi.2021.3074893.

[22] M.M. Kabir, M.F. Mridha, A. Rahman, et al. Detection of COVID-19, pneumonia, and tuberculosis from radiographs using AI-driven knowledge distillation, Heliyon 10 (2024), e26801. https://doi.org/10.1016/j.heliyon.2024.e26801.

[23] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR 9:249-256, 2010.

[24] K. He, J. Sun, Convolutional neural networks at constrained time cost, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5353-5360, 2015.

[25] C. M. Bishop, “Neural Networks for Pattern Recognition”, Oxford university press, 1995.

[26] 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.

[27] 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.

[28] A. Mitchell, E. Edbert, G.N. Elwirehardja, et al. Offline signature verification using transfer learning and data augmentation on imbalanced dataset, ICIC Express Lett. 17 (2023), 359-366. https://doi.org/10.24507/icicel.17.03.359.

[29] K.W. Gunawan, A.A. Hidayat, T.W. Cenggoro, et al. Repurposing transfer learning strategy of computer vision for owl sound classification, Procedia Computer Sci. 216 (2023), 424–430. https://doi.org/10.1016/j.procs.2022.12.154.

[30] 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.

[31] A. Tanubrata, G.N. Elwirehardja, B. Pardamean, Focusing on correct regions: self-supervised pre-training in lung disease classification, Commun. Math. Biol. Neurosci. 2024 (2024) 68. https://doi.org/10.28919/cmbn/8577.

[32] T.W. Cenggoro, S.M. Isa, G.P. Kusuma, et al. Classification of imbalanced land-use/land-cover data using variational semi-supervised learning, in: 2017 International Conference on Innovative and Creative Information Technology (ICITech), IEEE, Salatiga, 2017: pp. 1–6. https://doi.org/10.1109/INNOCIT.2017.8319149.

[33] T.W. Cenggoro, Incorporating the knowledge distillation to improve the EfficientNet transfer learning capability, in: 2020 International Conference on Data Science and Its Applications (ICoDSA), IEEE, Bandung, Indonesia, 2020: pp. 1–5. https://doi.org/10.1109/ICoDSA50139.2020.9212994.

[34] 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.

[35] 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.

[36] 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.