ICIC Express Letters
Part B: Applications
Volume **15**, Number **1**, January **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **83–89**

EVALUATING SELF-SUPERVISED PRE-TRAINED VISION TRANSFORMER ON IMBALANCED DATA FOR LUNG DISEASE CLASSIFICATION

Elvan Selvano[1,*], Aedentrisa Yasmanda Paulindino[1]
Gregorius Natanael Elwirehardja[2,3] and Bens Pardamean[1,2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science
[2]Bioinformatics and Data Science Research Center
[3]Computer Science Department, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
aedentrisa.paulindino@binus.ac.id; { gregorius.e; bpardamean }@binus.edu
*Corresponding author: elvan.selvano@binus.ac.id

Received February 2023; accepted May 2023

Abstract. *Lung disease has been known as one of the most prevalent medical disorders globally and a leading cause of death and disability. Pneumonia, one of the most common lung diseases, accounts for over 2.4 million deaths annually, and COVID-19 has further increased deaths from pneumonia globally. Chest X-Ray (CXR) has been proven as the most prominent screening method, and deep learning techniques have been widely used for computer-aided diagnosis (CAD). This paper aims to evaluate the performance of Vision Transformer (ViT), self-supervised learning (SSL) techniques, and pre-trained convolutional neural network (CNN) models in classifying four lung conditions from publicly available dataset containing more than 20,000 CXR images. The results showed that DINO ViT-S16 performs the best with precision/recall/F1-score of 95.61%/95.75%/95.67% for the imbalanced dataset, 94.16%/94.56%/94.35% for the augmented dataset, and 93.99%/94.05%/93.86% for the undersampled dataset. The lung regions from the CXR image were correctly highlighted by the model which contributed towards the correct classification. The proposed model also offered higher performance than other previously reported approaches and provides the opportunity for an efficient evaluation with an accuracy acceptable in the medical area.*

Keywords: Lung disease, Chest X-Ray, Vision Transformer, Self-supervised learning, Imbalanced data, Deep learning techniques

1. **Introduction.** The National Institute of Environmental Health Sciences (NIEHS) states that “lung disease” refers to various illnesses or conditions that impair the ability of the lungs from functioning properly. These diseases can affect the performance of the lungs, pulmonary functions, respiratory function, or one’s breathing capacity [1]. Among all causes of death in 2015, lung disease ranks fourth globally, and according to statistics, the mortality rate for lung disease globally reaches 20%-40% [2].

The use of machine learning and AI in medicine is becoming more prevalent, specifically in the form of computer-aided diagnosis (CAD) in medical imaging [3]. This approach uses image analysis to detect disease patterns and can be beneficial for patients in remote areas [4,5]. This helps physicians to identify diseases early on and assist the pharmaceutical business in making the best medical decisions quicker. Although many machine learning solutions have been deployed for CAD, these approaches have some limitations: they may not be able to interpret complex disease patterns, cannot learn from limited data, and need enormous resources.

DOI: 10.24507/icicelb.15.01.83

Due to its robustness in handling data imbalance, SSL is often utilized for transfer learning (TL). SwAV-TL on CNN managed to help the model achieve higher performance than its supervised counterpart [6]. This study aims to verify whether the same phenomenon occurs in ViT, which was pre-trained with an SSL algorithm. This study compared the performance of DINO ViT, SwAV ResNet, and their supervised versions, to determine whether SSL pre-training can improve their accuracy in medical image analysis.

The rest of the paper is structured as follows: Section 2 reviews previous studies and highlights gaps in the literature, Section 3 describes the proposed approach, Section 4 presents the experimental results of the proposed approach, and finally, Section 5 summarizes the main contributions, discusses limitations, and suggests future research directions.

2. Related Works.

Abbas et al. [7] modified decompose, transfer, and compose (DeTraC), a CNN architecture that uses a class decomposition strategy to classify COVID-19 using CXR images. DeTraC demonstrated efficient methods for classifying COVID-19 instances with an accuracy of 98.23%.

Rahman et al. [8] proposed a method for identifying tuberculosis from CXR images using CNN. Out of 9 models that they evaluated, the CheXNet model performed better than other models without lung segmentation, but the DenseNet201 model outperforms on the segmented lungs. Without segmentation, the proposed method achieved an accuracy of 96.47%, precision of 96.62%, and recall of 96.47%. When using lung segmentation, the accuracy, precision, and recall increased to 98.60%, 98.57%, and 98.56%, respectively.

Chowdhury et al. [9] used CXR images from the COVID-19 Radiography Database to identify COVID-19 and evaluated the performance of DenseNet201 and CheXNet, in which the former outperformed the latter with 99.7% precision and 99.7% sensitivity. Muljo et al. [10] trained DenseNet121 to classify 4 lung diseases using the same dataset, and achieved 82.16% average AUC, with Viral Pneumonia having the highest AUC at 99.99%.

Pham et al. [11] suggested that the identification of COVID-19-related lesions in the lungs, based on characteristics such as location, size, and distribution, is more valuable for medical professionals to evaluate the severity of the disease, track treatment progress, and monitor patient recovery. The model accurately learned disease-related characteristics by focusing on the annotation data of lung lesion in medical images. The results showed that the method of annotating COVID-19 images improved the model's accuracy by up to 1.68 times and is comparable to commercially available options.

Following the review of the papers using the TL/SWAV/SSL method, it was discovered that most of the datasets contain limited images, making it possible that the model might not generalize well. Additionally, some datasets may be biased because they were only collected from one hospital. Some papers claimed the "COVID-19 Radiography Database" is too small, and the performance will be improved if utilizing a larger dataset. Because the dataset has several versions updated by the publisher, the current argues that the papers were created using the dataset with version < 4. However, the dataset has since been updated to version 4 with more images, and four classes may be classified compared to the 2-3 classes used by most papers.

3. Methodology.

3.1. Dataset discussion.

"COVID-19 Radiography Database" [12] was used in training and evaluating the models. This dataset includes CXR images from Italian Society of Medical Radiology (SIRM) [13], Novel Corona Virus 2019 Dataset developed by Joseph Paul Cohen, Paul Morrison, Lan Dao in GitHub [14], BIMCV-COVID19 [15], and CXR Images (pneumonia) database [16].

The dataset comprises 21,165 CXR images, which have been classified into four categories, as shown in Figure 1. These categories comprise 3,616 COVID-19, 6,012 Lung Opacity, 10,192 Normal, and 1,345 Viral Pneumonia images. Based on the number of images, the dataset can be categorized as imbalanced, and earlier approaches to this dataset were tested using the limited data without additional configurations.

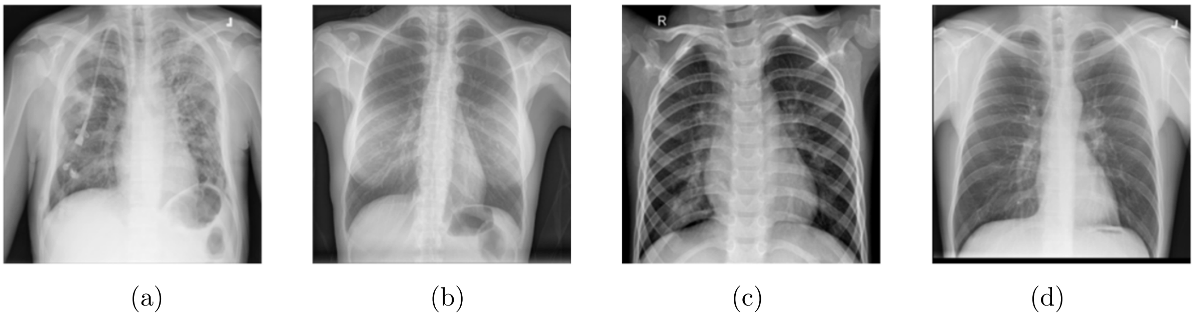


Figure 1. Sample CXR images for each class: (a) Lung Opacity; (b) COVID-19; (c) Viral Pneumonia; (d) Normal

3.2. **Data preprocessing.** The experiment aimed to test whether SSL can reduce bias and improve performance when imbalanced data is used. To thoroughly examine the differences between SSL models with supervised learning models and their performances on the same dataset with different inter-class data distributions, three datasets were generated, which are imbalanced, augmented, and undersampled. Each dataset was divided into three parts: training, validation, and testing, with the respective proportions of 70%, 10%, and 20%. The specific numbers of images for each dataset are presented in Table 1.

Table 1. Distribution of images for each class of the dataset

Class	Imbalanced/Augmented			Undersampled		
	train	val	test	train	val	test
COVID	2,531	361	724	941	134	269
Lung Opacity	4,208	601	1,203	941	134	269
Normal	7,134	1,019	2,039	941	134	269
Viral Pneumonia	941	134	270	941	134	269

The imbalanced dataset was used to evaluate the robustness of SSL models when there are large differences in the amount of data per class. The augmented dataset was used to assess the effectiveness of SSL models to improve accuracy by adding modified images through data augmentation. The undersampled dataset was used to investigate the ability of SSL models to reduce bias and improve accuracy when the number of images per class is equal. Using an undersampled dataset can be beneficial as it can decrease the amount of data required for training while still learning meaningful features.

3.3. **Experiment configuration.** To improve the generalization of the models, techniques to enhance the amount and variety of images in the dataset were used in the training process [17]. These techniques, known as data augmentation, include applying random 45-degree rotations and flipping the images horizontally or vertically. All the images were rescaled to 224 × 224 to align with the expected input size for the pre-trained models.

For the classification task, six different model architectures were trained for 50 epochs using the Adam optimizer, a learning rate of 1e-4, and a random seed of 42. Additionally, early stopping was enabled to prevent overfitting. Transfer learning was performed by adding a new fully connected layer with randomized weights that has an output dimension

Table 2. Parameter configuration of the models

Model	Parameters	Trainable parameters	Non-trainable parameters
DINO ViT-B16	85,801,732	14,180,356	71,621,376
ViT-B16	86,570,732	7,861,484	78,709,248
DINO ViT-S16	21,668,740	3,552,772	18,115,968
ViT-S16	22,053,740	3,937,772	18,115,968
ResNet50 SwAV	25,560,108	6,514,668	19,045,440
ResNet50	25,560,108	6,514,668	19,045,440

equal to the number of classes. Several pre-trained layers from the models were also unfrozen to fine-tune the model to get better performance. This influenced the number of trainable parameters which is specified in Table 2.

3.4. **Evaluated deep learning models.** Vision Transformer (ViT) performs remarkably well in various vision applications. It tokenizes an input image into patches and maps each patch to a token embedding. The final classification of the images and the aggregation of global image data are handled by an additional class token (CLS) that is added to the set of image tokens. A learnable vector (i.e., positional encoding) adds each token, and it feeds it into the feed-forward network (FFN) and multi-head self-attention (MHSA) layer of the sequentially stacked Transformer encoders [18].

DINO is a self-distillation technique; it develops a network of students and teachers and utilizes a standard cross-entropy loss to directly forecast the output of a teacher network created with a momentum encoder [19]. Swapping Assignments Between Views (SwAV) is a contrastive method that operates in batches and does not require a large memory bank. SwAV process consists of three steps. First, it computes the features of two augmented versions of images from a batch and assigns them to "prototype vectors" in a spherical feature space. Next, it clusters the data that requires the two cluster assignments for an image to match. Third, it determines "the code" from the feature by resolving a novel "swapped" prediction technique. The model concludes that similar images share information by projecting this loss to all images and their augmentations [20].

3.5. **Model evaluation.** In order to assess the performance of the model, precision (P), recall (R), and F1-score (F) were used as metrics. These can be computed by utilizing the number of True Positive (TP), False Positive (FP), and False Negative (FN), as shown in Equation (1).

$$\mathrm{F} = \frac{2\mathrm{P}\cdot\mathrm{R}}{\mathrm{P}+\mathrm{R}}, \text{ where } \mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \ \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{1}$$

Understanding the reasons behind the predictions is crucial in order to get a trustworthy result. LIME (Local Interpretable Model-agnostic Explanations) [21], a technique that can help explain individual predictions, was used to assess the model's predictions. It modifies a single image by tweaking the feature values and observes the resulting impact on the output.

4. **Result and Discussion.** Table 3 shows that DINO ViT-S16 performed the best with the highest F1-score. While the models generally performed well with F1-score of 87.00% and higher in the imbalanced dataset, ViT-S16 appeared to be underfitting in all datasets. In contrast, all models performed poorly in the undersampled dataset due to limited data.

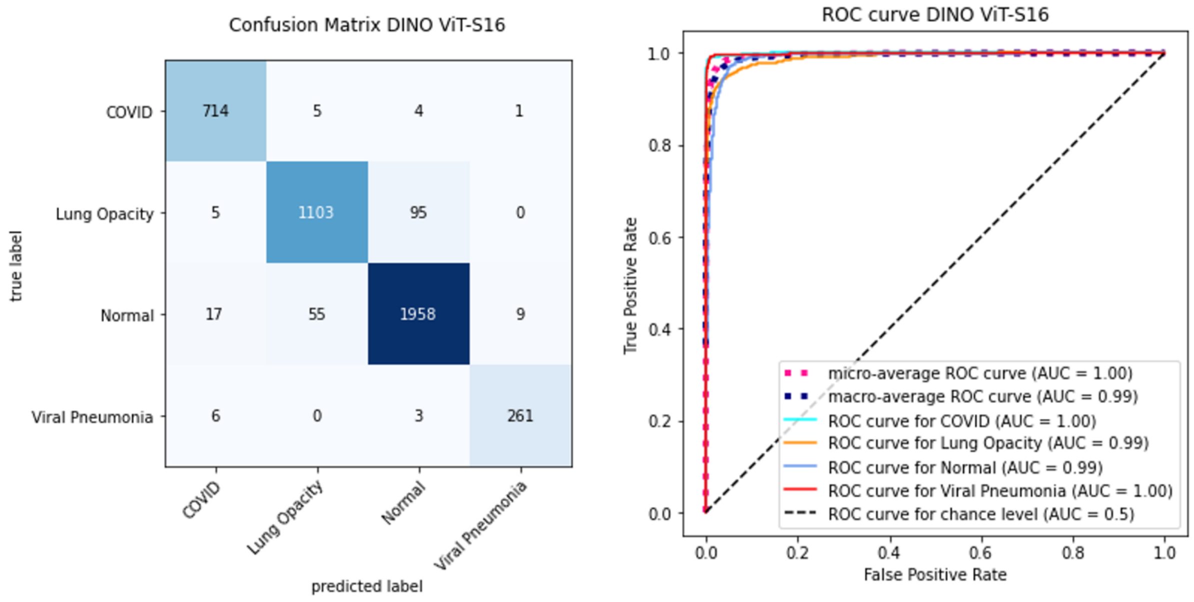
All models using SwAV or DINO also performed better than their supervised counterparts which means the methods used were effective in tackling the imbalance issue.

Table 3. Evaluation results on the test set (%)

Model	Imbalanced			Augmented			Undersampled		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
DINO ViT-B16	93.94	95.50	94.66	95.13	93.39	94.20	93.71	93.37	93.38
ViT-B16	94.58	92.61	93.54	93.76	91.57	92.55	90.49	90.21	90.24
DINO ViT-S16	**95.61**	**95.75**	**95.67**	**94.16**	**94.56**	**94.35**	**93.99**	**94.05**	**93.86**
ViT-S16	84.76	83.23	83.87	81.86	81.65	81.69	77.21	76.06	76.36
ResNet50 SwAV	90.98	90.30	90.62	89.89	89.78	89.81	89.44	88.82	89.08
ResNet50	89.90	90.39	90.07	88.08	86.33	87.10	89.11	88.23	88.28

Although DINO ViT-B16 has almost 4 times the number of parameters compared to DINO ViT-S16, it seemed that the model has overfitted and performed a bit worse than its smaller variant.

4.1. **Model evaluation plot.** Figure 2 shows the confusion matrix and the area under the receiver operating characteristic (ROC) curve for DINO ViT-S16 for the imbalanced test data. Figure 2(a) shows that the model can predict COVID, Viral Pneumonia, and Normal with less than 5% error, 714 out of 724 images for COVID, 261 out of 270 images for Viral Pneumonia, and 1,958 out of 2,039 images for Normal. This indicates the model's robustness in handling imbalanced data. Despite the similarity in characteristics between Lung Opacity and Normal, leading to mixed-up predictions, the model's overall performance appeared promising, as reflected in its F1-score.



(a) (b)

Figure 2. (color online) Confusion matrix (a) and ROC AUC curve (b) for each class

Figure 2(b) shows the area under the ROC curve for the one-vs-rest approach. The model achieved a perfect score of 1.00 for classifying COVID and Viral Pneumonia and a near-perfect score of 0.99 for Lung Opacity and Normal. However, these scores might be too optimistic, considering the imbalanced data.

4.2. **LIME visualization.** Visualization is generated where the green area (indicated by the letter "G") is the area that contributes towards the correct classification while the red area (indicated by the letter "R") indicates the opposite. The color intensity is proportional to the absolute value of the contribution. Therefore, a perfect model should highlight the parts of the lungs as dark green areas as much as possible with no red areas.

Figure 3 shows that DINO ViT-S16 performed well with a lot of green-shaded areas with the least amount of red-shaded areas, while for some models like VIT-S16 supervised, it has a lot of red areas in the lungs, which makes the predictions unreliable because the lung areas should contribute the most.

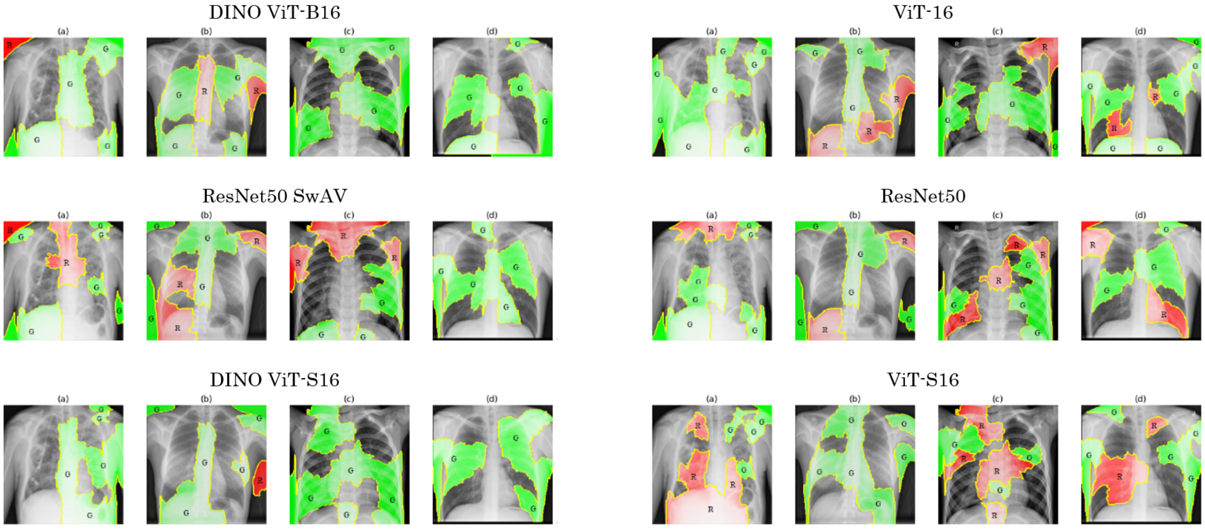


FIGURE 3. LIME visualizations for each class: (a) Lung Opacity; (b) COVID-19; (c) Viral Pneumonia; (d) Normal

5. **Conclusions.** This paper proposed a pre-trained ViT model using DINO to identify lung diseases from CXR images. The performance of the model was compared to ResNet50 models that were pre-trained with SwAV. The DINO ViT-S16 model demonstrated excellent results even with imbalanced data. The proposed model was found to be superior to other previously published methods and has the potential to provide an adequate level of accuracy suitable for the medical field. LIME was employed to demonstrate that the model trained with SSL primarily focuses on the lung region, requiring it to use relevant information only, leading to improved performance compared to its supervised counterparts. This demonstrates that SSL can handle imbalanced data and can be applied to other medical image classification tasks.

A limitation of this study is the unbalanced nature of the dataset, which may not be reliable in serving as the baseline for comparison with the state-of-the-art models. Furthermore, the study focused on using CXR images to achieve the highest classification rate for lung disease, among other types of images. To overcome this limitation, future studies should implement techniques such as data augmentation and oversampling to expand the amount of training data and explore alternative methods for evaluating the model's predictions.

Acknowledgment. This work is supported by Bina Nusantara University. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] *Lung Diseases*, National Institute of Environmental Health Sciences, https://www.niehs.nih.gov/health/topics/conditions/lung-disease/index.cfm, 2018.

[2] S. Geiger, D. Hirsch and F. G. Hermann, Cell therapy for lung disease, *European Respiratory Review*, vol.26, no.144, 2017.

[3] H. P. Chan, R. K. Samala, L. M. Hadjiiski and C. Zhou, Deep learning in medical image analysis, *Advances in Experimental Medicine and Biology*, vol.1213, pp.3-21, https://doi.org/10.1007/978-3-030-33128-3_1, 2020.

[4] Daniel, T. W. Cenggoro and B. Pardamean, A systematic literature review of machine learning application in COVID-19 medical image classification, *Procedia Computer Science*, vol.216, pp.749-756, DOI: 10.1016/j.procs.2022.12.192, 2023.

[5] B. Pardamean, T. W. Cenggoro, R. Rahutomo, A. Budiarto and E. K. Karuppiah, Transfer learning from Chest X-Ray pre-trained convolutional neural network for learning mammogram data, *Procedia Computer Science*, vol.135, pp.400-407, DOI: 10.1016/j.procs.2018.08.190, 2018.

[6] H. H. Muljo, B. Pardamean, G. N. Elwirehardja, A. A. Hidayat, D. Sudigyo, R. Rahutomo and T. W. Cenggoro, Handling severe data imbalance in Chest X-Ray image classification with transfer learning using SwAV self-supervised pre-training, *Commun. Math. Biol. Neurosci.*, Article ID 13, DOI: 10.28919/cmbn/7526, 2023.

[7] A. Abbas, M. M. Abdelsamea and M. M. Gaber, Classification of COVID-19 in Chest X-Ray images using DeTraC deep convolutional neural network, *Applied Intelligence*, vol.51, no.2, pp.854-864, DOI: 10.1007/s10489-020-01829-7, 2021.

[8] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, T. Hamid, M. T. Islam, S. Kashem, Z. Bin Mahbub, M. A. Ayari and M. E. H. Chowdhury, Reliable tuberculosis detection using Chest X-Ray with deep learning, segmentation and visualization, *IEEE Access*, vol.8, pp.191586-191601, 2020.

[9] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, A. M. Kadir, Z. Bin Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi, M. B. I. Reaz and M. T. Islam, Can AI help in screening viral and COVID-19 pneumonia?, *IEEE Access*, vol.8, pp.132665-132676, DOI: 10.1109/ACCESS.2020.3010287, 2020.

[10] H. H. Muljo, B. Pardamean, K. Purwandari and T. W. Cenggoro, Improving lung disease detection by joint learning with COVID-19 radiography database, *Commun. Math. Biol. Neurosci.*, Article ID 1, DOI: 10.28919/cmbn/6838, 2022.

[11] V. Pham, D. Dinh, E. Seo and T. M. Chung, COVID-19-associated lung lesion detection by annotating medical image with semi self-supervised technique, *Electronics (Switzerland)*, vol.11, no.18, DOI: 10.3390/electronics11182893, 2022.

[12] T. R. Muhammad E. H. Chowdhury, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi and M. Bin I. Reaz, *COVID-19 Chest X-Ray Database*, https://www.kaggle.com/tawsifurrahman/covid19-radiography-database, 2020.

[13] SIRM, *COVID-19 Database*, https://www.sirm.org/category/senza-categoria/covid-19/, 2020.

[14] J. C. Monteral, *COVID-Chestxray Database*, https://github.com/ieee8023/covid-chestxray-dataset, 2020.

[15] BIMCV-COVID19, *Datasets Related to COVID19's Pathology Course*, https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/, 2020.

[16] P. Mooney, *Chest X-Ray Images (Pneumonia)*, https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia, 2018.

[17] T. Miura, Y. Omae, Y. Saito, D. Fukamachi, K. Nagashima, Y. Okumura, Y. Kakimoto and J. Toyotani, Three-state classification of pulmonary artery wedge pressure from Chest X-Ray images using convolutional neural networks, *ICIC Express Letters, Part B: Applications*, vol.14, no.3, pp.271-277, DOI: 10.24507/icicelb.14.03.271, 2023.

[18] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang and P. Xie, Not all patches are what you need: Expediting vision transformers via token reorganizations, *arXiv Preprint*, arXiv: 2202.07800, 2022.

[19] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski and A. Joulin, Emerging properties in self-supervised vision transformers, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp.9650-9660, 2021.

[20] A. Mohamed and D. Demidov, Self-supervised visual representation learning with convolutional neural networks: A survey, *ResearchGate Preprint*, DOI: 10.13140/RG.2.2.15351.50088, 2021.

[21] L. O. Teixeira, R. M. Pereira, D. Bertolini, L. S. Oliveira, L. Nanni, G. D. Cavalcanti and Y. M. Costa, Impact of lung segmentation on the diagnosis and explanation of COVID-19 in Chest X-Ray images, *Sensors*, vol.21, no.21, 7116, 2021.