Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2025, 2025:22
https://doi.org/10.28919/cmbn/9048
ISSN: 2052-2541

COMPARATIVE ANALYSIS OF SELF-SUPERVISED PRE-TRAINED VISION TRANSFORMERS AND CONVOLUTIONAL NEURAL NETWORKS WITH CHEXNET IN CLASSIFYING LUNG CONDITIONS

GREGORIUS NATANAEL ELWIREHARDJA[1,*], STEVE MARCELLO LIEM[1], MARIA LINNEKE ADJIE[1], FARREL ALEXANDER TJAN[1], JOSELYN SETIAWAN[1], MUHAMMAD EDO SYAHPUTRA[1], HERY HARJONO MULJO[2]

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

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

Abstract. Classifying lung diseases from images has been a challenging task for Deep Learning (DL) methods. Self-Supervised Learning (SSL) in particular has been widely recognized to be effective for pre-training, especially new methods such as DINOV2 ViT/S-14 and ConvNeXt-V2. In this research, Transfer Learning (TL) was conducted on the two models by using the NIH CXR-14 dataset to perform 15-class classification. Additionally, SwAV ResNet-50, DINO ViT-S/16, and CheXNet were adopted as the baselines. Evaluation results showed that DINOV2 ViT-S/14 is superior to the other three models pre-trained on ImageNet with 0.743 macro-averaged AUC, but is inferior to CheXNet which was pre-trained using the same NIH CXR-14 dataset albeit without the "No Finding" class. However, the CheXNet only obtained 0.773 AUC with 0.328 recall. Further analysis on feature separability showed that both CheXNet and DINOV2 ViT-S/14 were unable to extract meaningful features that differentiate the "No Finding" class with the other 14 lung conditions, confirming the finding from a previous study that this dataset's labels is noisy, rendering it unsuitable for downstream TL tasks. However, DINOV2 ViT-S/14 showed similar attention visualizations with CheXNet on some classes despite being pre-trained on natural images from

*Corresponding author
E-mail address: gregorius.elwirehardja@binus.ac.id
Received November 29, 2024

ImageNet. Therefore, despite the unsatisfactory performance in this dataset, DINOV2 holds great potential in similar future studies, but it may require pre-training the models on medical image datasets.

Keywords: deep learning; self-supervised learning; vision transformers; chest X-ray; convolutional neural networks.

2020 AMS Subject Classification: 68T07, 68T45, 92C55.

1. INTRODUCTION

Lung diseases have consistently remained among the top-20 leading causes of death. Chronic Obstructive Pulmonary Disease (COPD), such as emphysema, ranked third among these leading causes [1]. As such, early screening for these anomalies are imperative, such as by utilizing Machine Learning (ML) or even Deep Learning (DL). DL models had been adopted in various medical fields, but it requires massive volumes of training data. However, the issue of data scarcity persisted as a huge gap [2]. Additionally, the major imbalance between the number of healthy and sick samples introduced issues for DL models, leading to biased predictions towards the majority classes [3]. In the field of lung diseases, subtle and overlapping visual characteristics between different conditions became another hurdle to overcome for DL models [4].

In recent years, the Self-Supervised Learning (SSL) paradigm gained a lot of traction. SSL is a method to train models in learning data representations without labels, thus overcoming the hurdle of limited data availability. It had shown promising results in various fields namely Speech Recognition, Natural Language Processing, and also lung disease detection through the means of Transfer Learning (TL). Previous studies suggested that SSL pre-training in computer vision allows DL models to extract more meaningful visual features compared to supervised ones [5, 6]. Additionaly, numerous studies have demonstrated that SSL, specifically Contrastive Learning (CL), not only enhances DL model's ability to generalize but also proves effective in addressing the data imbalance problem, especially in image classification tasks where images in minority class are often under-represented in the model's predictions as well. It seems that CL, which trains models to extract more similar features invariant to transformations, allows DL models to improve their feature representations or embeddings [3]. Therefore, the method can prove promising to distinguish subtle differences in lung images such as Chest X-Rays (CXR).

Throughout the years, various SSL algorithms had been tested for CXR classification. CL methods had proven successful thus far for CXR classification cases using either Convolutional Neural Networks (CNN) or Vision Transformers (ViT), such as the Swapping Assignments between multiple Views (SwAV), self-Distillation with No labels (DINO), and SimCLR [3, 7, 8], making CL pre-training dubbed as the State-of-The-Art for CNN models [9]. Aside from CL, restorative learning such as Masked Auto-Encoders (MAE) had also received more popularity for ViT models and deemed to be more suitable [10]. Still, a lot of improvement can still be done, such as investigating the feature separability as well as explainability [9]. A previous study in particular had proven that DINO allowed ViT models to focus on more correct Regions of Interest (RoI) in lung disease classification compared to CNNs [11]. Therefore, it will be beneficial to further explore newer SSL pre-training methods for classifying lung conditions through CXR images. For ViT models, DINO's improved version, DINOV2, had shown remarkable results in various classification cases [12]. Similarly, the new ConvNeXt-V2, a CNN model claimed to be even superior to ViTs, was pre-trained using MAE and brought outstanding performance [13]. However, both methods had yet to be evaluated on CXR classification cases with considerable number of classes.

The main goal of this research is to compare and analyze the performance of newer SSL algorithms in a downstream lung condition classification task. In this research, DINOV2 ViT-S/14 and ConvNeXt-V2, which are two of the newest additions among SSL pre-trained models, were compared in classifying 15 lung conditions. DINOV2 is a CL method where ConvNeXt-V2 was pre-trained restoratively, confronting the trends of using CL on CNNs and restorative learning on ViTs. The models were compared with replicated baseline models from previous studies, namely ResNet-50 pre-trained using SwAV [3] and ViT-S/16 model pre-trained using the original DINO [11]. All four models were pre-trained on ImageNet, similar to the conditions made on the original works. In addition, CheXNet [14], which was pre-trained specifically for lung condition classification, was also replicated and re-trained. The adoption of CheXNet was done to verify whether the model's performance even when trained on natural images in ImageNet can surpass another pre-trained using medical images. However, if the models proved inadequate, CheXNet can be used for verifying whether the models' unsatisfactory performance were

caused by incorrect configurations or due to the dataset itself. The evaluation were conducted based on three key aspects: the model's discriminative confidence, feature separability, and attention visualization.

The following section describes past related studies in using DL and SSL pre-training for lung CXR image classification. Section 3 describes details of the research methodology as well as the four SSL algorithms used whereas Section 4 discusses the findings. The conclusion is presented in Section 5.

2. Related Works

Ever since the introduction of CheXNet, DL has been a rapidly growing method for Computer-Aided Diagnosis (CAD). By pre-training a DenseNet-121 model on the National Institute of Health Chest X-Ray 14 (NIH CXR-14) dataset, the model managed to outperform human radiologists on classifying 14 categories of lung conditions despite only achieving 0.435 F1 score [14]. Ever since then, pre-training a model and conducting TL on a downstream task has been a major trend among AI researchers for CXR classification. In 2022, TL on DenseNet-121 was conducted along with joint learning to train it in classifying 18 classes of lung conditions, 14 of them originating from the NIH CXR-14 dataset excluding the "No Findings" class. However, the AUC of the models only reached 0.9 for four classes and none was from the NIH CXR-14 dataset, attaining an average AUC of 0.822 [15]. In a similar study, combining ResNet-50 and DenseNet-121 brought about more than 0.93 AUC on four-class CXR classification case [16]. Other CNN backbones had also proven their capabilities on binary CXR classification, attaining around 0.98 recall for all models [17]. Other similar studies proved the dominance of ViT as well for CXR image classification, obtaining 0.95 recall for binary pneumonia classification [18] and 0.958 accuracy on four-class classification [19]. CheXNet had also been used, albeit in a binary classification context where it performed not too well due to the noisy labels [20]. As more research emerged, it also paved a way to experiments involving different pre-training schemes, including SSL.

With the growing popularity of SSL, usages of CL has grown in CXR image classification tasks. The method itself has been dubbed the State-of-The-Art for pre-training CNN models [9] and known to allow ViT models to extract informative semantic features [21]. In our previous

research, we proved that using models pre-trained through CL algorithms such as SwAV [22] allowed ResNet-50 models to maintain their performance even when trained on imbalanced data, attaining 0.952 AUC for four-class CXR classification [3]. In a recent study, even one of the simplest CL algorithm, SimCLR [23], had proven capable of allowing CNNs to reach 0.95 AUC with merely 500 images for fine-tuning in a binary classification task [7]. Moreover, CL had proven its success not only for CNNs, but also ViT models. ViT models pre-trained using DINO had proven superior to SwAV ResNet-50 when fine-tuned on four-class CXR classification, even when the dataset is imbalanced [11]. In a similar study, researchers had further verified its capability by pre-training a ViT model through DINO using the NIH CXR-14 dataset, allowing it to obtain 0.961 F1 score when fine-tuned on the Cell dataset. However, the model proved inferior compared to ResNet-50 when fine-tuned on the COVIDGR dataset [8] . Such results could be caused by the fact that ViT models require lots of training data [9] and less than 1,000 images was used in the research. In other recent studies, researchers proposed restorative learning instead of CL for pre-training ViT, such as the Masked Auto-Encoders (MAE) [10]. Through restorative learning, ViT models can obtain outstanding results after being trained on vast amounts of data, including both general objects (e.g. ImageNet) and the targeted domain samples (e.g. CXR images) [9].

Overall, CL had proven to be an effective approach in pre-training both CNNs and ViTs. Despite restorative learning growing to be more prominent in recent years for ViT, newer CL algorithms such as DINOv2 [12], the upgraded version of DINO said to be even more accurate, may be promising to test. Additionally, the advent of ConvNeXt, which was an improved version of ResNet-50, had caused a stir as it was said to be even superior to ViT for various computer vision tasks [24]. With the introduction of ConvNeXt-V2 pre-trained using MAE [13], it can be said that the streak of innovation will continue further. In this research, we aim to compare these newer models pre-trained through SSL, namely ViT pre-trained using DINOv2 and ConvNeXt-V2, in classifying CXR images from the challenging NIH CXR-14 data. As the dataset itself contain over 100,000 images, it should be an interesting test to validate whether CNNs and ViTs can maintain their performance.

3. RESEARCH METHODOLOGY

This section describes the details surrounding the experiment setup regarding the comparative analysis of ViTs and CNNs pre-trained using recently developed SSL pre-training algorithms.

3.1. Dataset. This research utilized the publicly available NIH CXR-14 dataset [25]. This dataset contains over 100,000 CXR images in non-DICOM format as well as other associated data from more than 30,000 patients. For this research, only 91,324 images with single-class labels from a total of 15 classes were selected. These 15 classes include 14 lung anomaly labels and one "No Findings" class, indicating that none of the 14 aforementioned anomalies were found. To create these labels, the authors used Natural Language Processing to text-mine disease classifications from the associated radiological reports. The labels are expected to be 90% accurate and suitable for weakly-supervised learning. It should also be noted that the dataset is severely imbalanced, where more than 66% of the images belong to the "No Findings" class. The label distribution is explained in the next section.

3.2. Data Pre-processing. For fairer comparisons with current State-of-The-Art methods, the original subset splitting provided by the NIH was used for splitting the dataset into two subsets. One was the train set and the other was the test set. As the authors of the dataset only provided the split for these two subsets, we further split the train set into two subsets, with one being used for validation. 20% images from the train set was used for the validation set, resulting in 64.368%, 14.683%, and 19.549% distribution among the train, validation, and test sets, respectively, as listed in Table 1. All images were then resized to 224×224 pixels to fit the pre-trained models' input shape. The Contrast-Limited Adaptive Histogram Equalization (CLAHE) was then applied to all subsets to enhance image contrast and the visibility of subtle features important for diagnosis while limiting enhanced contrast [26]. The method itself had proven beneficial for CNN models in CXR classification tasks [27]. The overall pre-processing steps were illustrated in Figure 1.

However, the appalling imbalance will surely affect the models' performances. The available training data for the minority class, which is Hernia, is only 52 images whereas the "No Finding" class boasts 40,400 training data. Previous studies has shown that Random Undersampling

TABLE 1. Distribution of CXR images across the three subsets.

No.	Class	Total	Test	Val	Train	Class Weight	Resampled
1	Atelectasis	4,215 (4.615%)	801	682	2,732	1.434	2,732
2	Cardiomegaly	1,093 (1.197%)	316	155	622	6.300	3,732
3	Consolidation	1,310 (1.434%)	481	165	664	5.902	3,984
4	Edema	628 (0.688%)	231	79	318	12.323	1,908
5	Effusion	3,955 (4.331%)	1,167	557	2,231	1.757	2,231
6	Emphysema	892 (0.977%)	305	117	470	8.338	2,820
7	Fibrosis	727 (0.796%)	176	110	441	8.886	2,646
8	Hernia	110 (0.120%)	45	13	52	75.363	312
9	Infiltration	9,547 (10.454%)	2,220	1,465	5,862	0.669	5,862
10	Mass	2,139 (2.342%)	443	339	1,357	2.888	8,142
11	No Finding	60,361 (66.095%)	9,861	10,100	40,400	0.097	16,160
12	Nodule	2,705 (2.962%)	457	449	1,799	2.178	10,794
13	Pleural Thickening	1,126 (1.233%)	309	163	654	5.992	3,924
14	Pneumonia	322 (0.353%)	88	46	188	20.845	1,128
15	Pneumothorax	2,194 (2.402%)	953	248	993	3.946	5,958
	Total	91,324	17,853	14,688	58,783		72,333

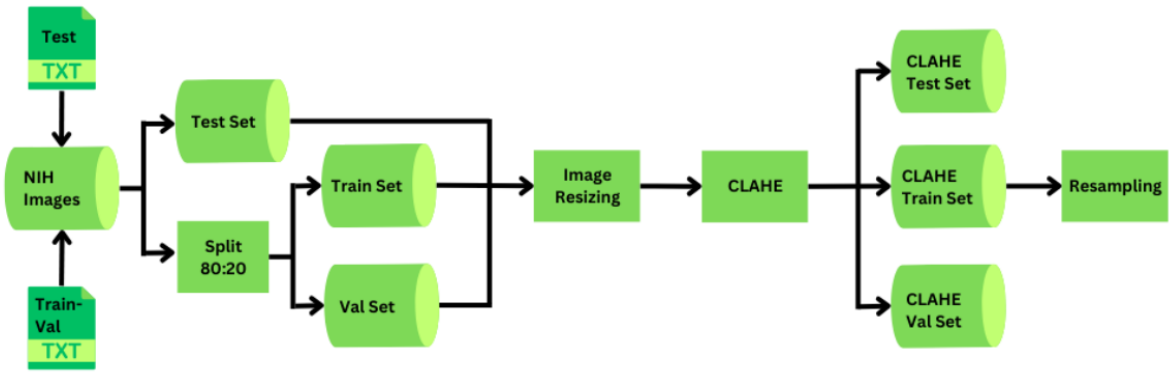

FIGURE 1. Flowchart of the data preprocessing pipeline

(RUS) did not seem to bring positive impacts in this particular task [3, 11]. Moreover, this approach will also reduce the impact of exploding gradients due to class weighting, which were used for training the models following approaches of the two aforementioned studies.

In our approach, we used the class weights of the original train set to determine which class should be oversampled and undersampled. Class weighting itself is formulized as:

$$W_C = \frac{n}{c \times n_C} \tag{1}$$

where W_C denotes the class weight for class C, n is the total number of data in the specified subset, c is the number of classes, and n_C represents the number of data in class C. The rules were set as follows:

(1) If $w_C \geq 2$, then oversample the images five times.
(2) If $1.5 \leq w_C < 2$, then keep the data as is.
(3) If $w_C < 0.1$, then undersample the data by 60%.

The thresholds were established through rigorous prior experiments based on the resulting validation loss of a baseline model. The oversampling was done using geometric augmentations, namely horizontal flipping and rotations, while RUS was used for the undersampling. After the resampling process, the values of w_C calculated using equation 1 range from 0.298 to 4.275 excluding Hernia, which had a weight of 15.456. This means that losses for misclassified Hernia images during training will be multiplied by 15.456 instead of 75.363, thereby minimizing the explosion of gradients during training. No resampling was done on the test and validation sets. For the oversampling, the rotation was set to -40, -20, 20, and 40 degrees, meaning that when added with horizontal flipping, five new images were generated for each image, yielding a total of six images per image after oversampling. Details of the data distribution after resampling are listed in Table 1. No augmentations were conducted on the validation and test sets. All images were normalized to the range of 0-1 and standardized using ImageNet's mean and standard deviation values, which are [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively.

3.3. Pre-trained models. A total of two SSL pre-trained models were evaluated along with three baselines. The former refers to a ViT-S/14 model pre-trained using DINOV2 and a ConvNeXt-V2 atto model. To the best of our knowledge, no experiments have been done on these two models for CXR classification with more than 10 classes. The baselines are CheXNet, a ResNet-50 pre-trained using SwAV, and a ViT-S/16 model pre-trained using the regular DINO

SSL algorithm, all of which were selected based on their impressive performance in CXR classification cases [14, 3, 11]. It should be noted that comparing CheXNet with the other models cannot be considered fair as the four other models were trained using ImageNet without the prior knowledge of medical image feature extraction. Hence, CheXNet is only used as a benchmark on whether the dataset is reliable as the original CheXNet was only trained on samples of the NIH CXR-14 dataset without the "No Finding" class. Details about the models and SSL pre-training algorithms are provided below.

3.3.1. *CheXNet.* CheXNet is a DenseNet-121 model pre-trained through multilabel supervised learning on 14 classes of the NIH CXR-14 dataset. The aim was enabling the model to identify these anomalies better compared to human radiologists. However, it should be noted that the training data only included anomalies without healthy CXR images, which is one of its limitations. The model still surpassed human radiologists with 0.435 F1 score while the latter obtained an average of 0.387 F1 score. Upon deeper analysis, the model still found it challenging to classify infiltration, pneumonia, consolidation, and nodules [14].

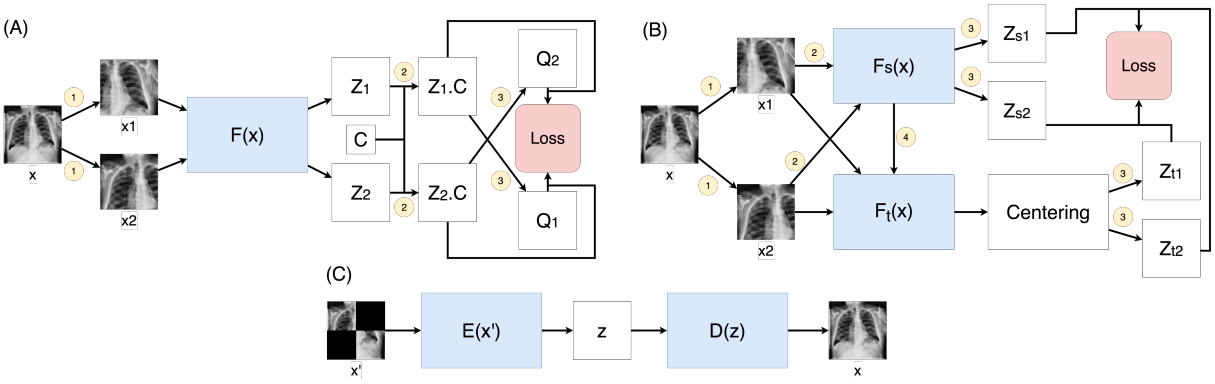

FIGURE 2. Illustrations for three of the SSL pre-training algorithms evaluated in this study: (A) SwAV, (B) DINO, and (C) ConvNeXt-V2

3.3.2. *SwAV.* SwAV is a CL algorithm that utilizes swapped prediction to train models that can produce more similar and consistent outputs for different augmentations of the same image [22], thereby allowing the models to produce robustly similar embeddings for similar or augmented images and make them more capable in finding similarities between images in the same

class. Unlike typical CL algorithms, SwAV compares the cluster assignments of each view instead of their features. From the illustration provided in Figure 2(A), SwAV comprises three main steps which are numbered in the figure as well.

(1) Perform multi-crop on an input image x to produce the augmented views, x_1 and x_2, which were further noised using color distortion and blurring. Pass both views to a DL model $F(x)$ to produce the embedding features Z_1 and Z_2.

(2) Take the learnable prototype vector C and calculate its dot product with Z_1 and Z_2 to produce $Z_1.C$ and $Z_2.C$.

(3) Assign $Z_1.C$ and $Z_2.C$ to clusters using the Sinkhorn-Knopp algorithm, producing Q_1 and Q_2 which are referred to as "codes". Perform the swapped prediction in the loss calculation using Cross-Entropy (CE) loss calculated through the following equations.

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

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

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

The loss is the sum of $L(Z_1, Q_2)$ and $L(Z_2, Q_1)$ calculated through the CE loss between Q and the softmax value of $Z.C$ where k denotes the number of prototype vectors and τ is the softening temperature.

3.3.3. *DINO.* DINO is a combination of self-distillation with CL, meaning that CL was performed using the same model architecture combined with Exponential Moving Average (EMA). The original method utilized two ViT models referred to as the student model $F_s(x)$ and the teacher model $F_t(x)$. Utilizing the multi-crop augmentation, the student model is trained to imitate the teacher model's output embedding's distribution. There are four main steps in the algorithm as illustrated in Figure 2(B).

(1) An input image x is randomly augmented, either using multi-crop or random photometric augmentation (e.g. color jitter, blurring, or solarization), generating a set of V global views. When $V = 2$, x_1 and x_2 will be produced.

(2) Create sets of local views from x_1 and x_2, which are their cropped or resized versions with lower resolutions. The local views are fed into $F_s(x)$ while the global views are processed by $F_t(x)$. The outputs of $F_s(x)$ are Z_{s1} and Z_{s2} whereas the outputs of $F_t(x)$ are Z_{t1} and Z_{t2}, respectively.

(3) Replace Z_{s1} and Z_{s2} with their softmax values and perform softmax centering on Z_{t1} and Z_{t2}, then calculate the loss of $F_s(x)$ using equation 5 where $L(a,b) = -a \times log(b)$.

$$L_s(x) = L(Z_{t1}, Z_{s2}) + L(Z_{t2}, Z_{s1}) \tag{5}$$

(4) Perform back-propagation on $F_s(x)$ and update the center value of $F_t(x)$, then update the parameters of $F_t(x)$ using EMA as shown in equation 6 where θ_t and θ_s represents the parameters of $F_t(x)$ and $F_s(x)$, respectively. λ is the momentum coefficient with values between 0.996 to 1.

$$\theta_t = \lambda\theta_t + (1-\lambda)\theta_s \tag{6}$$

Basically, the student model is trained to use local views, and generate output embeddings that mimics the global features produced by the teacher model. The method surpassed all previous CL algorithms on ImageNet downstream classification [21].

3.3.4. *DINOV2.* DINOV2 further improves DINO by combining it with the loss mechanism of Image BERT (iBOT) [28], KoLeo regularizer [29], and the cluster assignment of SwAV [22]. The processes are as follows:

(1) Train the model for image-level objective using DINO. The process is the same as regular DINO until the calculation of losses using equation 5. However, the softmax centering is replaced with the Sinkhorn-Knopp algorithm used in SwAV. This step produces L_{DINO}.

(2) Train the model for patch-level objective using iBot. This is done by masking the input image patches fed to the student model $F_s(X)$, producing the embedding Z_s. The input for the teacher model $F_t(x)$ is not masked, and its output is fed into the Sinkhorn-Knopp algorithm to produce the output embeddings Z_t. The loss is then calculated using

$$L_{iBot} = -\sum_{i=1}^{n} Z_{ti} log(Z_{si}) \tag{7}$$

where i and n denotes the index and number of patch indices of the masked tokens. This way, the student model is also trained to analyze the image's semantic features through de-masking.

(3) Regularize the model using KoLeo regularizer where $d_{n,i} = min_{i \neq j} ||x_i - x_j||$ with x_i representing a data point with the index i in a batch of input.

$$L_{KoLeo} = -\frac{1}{n} \sum_{i=1}^{n} log(d_{n,i}) \tag{8}$$

(4) Calculate the total loss as $L = L_{DINO} + L_{iBot} + L_{KoLeo}$, then perform back-propagation on $F_s(x)$ using L and update the parameters of $F_t(x)$ using EMA, as formulated in equation 6.

By combining these methods, the ViT model is trained not only to produce similar embeddings between local and global views on image-levels, but also on the patch-levels, prompting it to produce general-purpose spatial features. This led to greater accuracy on a downstream evaluation using ImageNet compared to regular DINO and MAE using ViT [12].

3.3.5. *ConvNeXt-V2.* ConvNeXt is a fully-convolutional neural network that combines the features of ViT models, specifically Swin Transformer, to ResNet-50. These features include the adoption of patching mechanism utilizing larger convolution kernels in the model's stem cells, the inverted bottleneck architecture for the stem cells, and replacing the ReLU activation function with GELU [24]. ConvNeXt-V2 takes the ConvNeXt architecture and incorporates SSL pre-training along with the usage of Global Response Normalization (GRN) layers. For the SSL pre-training, ConvNeXt-V2 adopted the Masked Auto-Encoder (MAE) approach, in which it takes a masked image x' and feeds it to the model's encoder $E(x)$ and decoder $D(x)$ to reproduce the original unmasked image x, as illustrated in Figure 2(C). Mathematically, x can be expressed as $x = D(E(x'))$. When evaluated on a downstream classification task using ImageNet, the model surpassed even standard MAE-ViT models, which was known to be remarkable in accuracy, proving once again how CNNs trained with restorative learning such as MAE can still prove superior to ViTs [13].

3.4. Experiment Setup. The aim of this research was to compare the performance of the different pre-trained models on single-label CXR classification through TL. Hence, all of the

pre-trained models got their decoders replaced with a fully-connected network, comprising a Dropout layer, a Dense layer with 128 neurons and ReLU activation function, another Dropout layer, and a Dense output layer. The models were built using PyTorch, pre-trained on ImageNet except for CheXNet, and fine-tuned using the grid search strategy. Four hyperparameters were tuned, namely the learning rate $lr \in \{1e-1, 1e-2, 1e-3, 1e-4, 1e-5\}$, the L2 regularizer lambda $\lambda \in \{0, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5\}$, the Dropout rate $r_{drop} \in \{0, 0.2, 0.4\}$, and the number of unfrozen encoder blocks for each model, effectively evaluating the best network adoption strategy for the TL as well. Only up to half of the convolution and transformer blocks for ConvNeXt-v2 and ViT models were unfrozen for the fine-tuning. Transformations using random horizontal flip and random rotation from -60 to 60 degrees were randomly applied on the input images during training to reduce the risks of overfitting. All models were trained for 50 epochs with 8 batch size. Additionally, class weighting and early stopping with a patience of 10 epochs were also applied.

3.5. Evaluation Method. Three evaluation metrics were set as the prioritized metrics, namely recall, F1 Score, and Area Under the Receiver Operating Characteristic Curve (AUROC). Recall, which is formulized in equation (9), measures the rate of correct predictions against the number of data in a class, also known as the True Positive Rate (TPR). Larger recall values show that the model can effectively classify images belonging to a specific class. F1 score is the harmonic mean between recall and precision calculated using equation (11). This particular metric can be crucial when the dataset is imbalanced as it can maintain the balanced trade-offs between recall and precision, providing insights to the proportion fo True Positives with both False Positives and False Negatives. Unlike recall and F1 which measure a classification model's correctness, AUROC, also known as Area Under the Curve (AUC), evaluates the confidence of the model's predictions to differentiate between classes, which is why it became the main metric in determining the best model in section 4. The ROC curve itself compares TPR against the False Positive Rate (FPR). Values close to 1 indicate higher confidence and better separability. In addition, the precision for the "No Finding" class is also calculated as a good model for this case should exhibit less false positives for this particular class.

$$Recall = TPR = \frac{TP}{TP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{11}$$

To measure the model's feature separation and comprehension capability, t-Distributed Stochastic Neighbor Embedding (t-SNE) and Gradient-weighted Class Activation Mapping (Grad-CAM) was used to inspect the produced features for all classes. Good classification models exhibit high degrees of separability. t-SNE works as a dimension reduction technique that changes the high dimensional data into a lower dimensional space. By analyzing these lower dimensional features from the model's embeddings, the model's ability to separate different classes can be measured. Clear t-SNE clustering indicates that the model successfully learned to use important features that differentiate the classes, especially in high dimensional data [30]. GradCAM on the other hand visualizes image regions deemed relevant by the model to make their predictions, typically shown as heatmaps. It has been commonly used for measuring the model's correctness in terms of feature importance interpretation in the form of RoI [31].

4. Results and Discusssion

4.1. Fine-tuning Results on Network Adaption for Each Model. Overall, evaluation results show that partial network adaption works best for all models. However, the models' classification performance were unsatisfactory. SwAV, which proved successful in a similar CXR classification study [3], only managed to achieve 0.730 macro-averaged AUC and 0.253 recall for this 15-class classification case using three unfrozen blocks, becoming the best SwAV model in this experiment. As shown in Table 2, despite having the best train and validation loss, the zero-adapted version with four unfrozen blocks performed slightly worse compared to the ResNet-50 model with three unfrozen blocks, whereas the fully-adapted model with zero unfrozen blocks performed the worst. 26.1% of the best model's "No Finding" was false, hence

the low recall and F1 scores. It is also implied that a lot of wrong classifications occur between the 14 anomalies based on the F1 score.

TABLE 2. Evaluation results on different numbers ofunfrozen blocks for SwAV ResNet-50. The asterisk (*) indicates the best model.

Unfrozen	Train loss	Val Loss	Test Loss	Test AUC	Recall	F1	Precision of "No Finding"
0	2.166	2.309	2.043	0.595	0.081	0.200	0.558
1	1.937	2.082	2.317	0.681	0.236	0.145	0.715
2	2.149	2.027	2.451	0.642	0.171	0.123	0.688
3*	1.798	2.093	**2.050**	**0.730**	**0.253**	**0.204**	**0.739**
4	**1.682**	**1.962**	2.108	0.727	0.250	0.186	0.754

Results for the other baseline, which was the DINO ViT-S/16, are listed in Table 3. Overall, the models performed worse compared to SwAV ResNet-50. Similar to the previous study [11], the ViT model with only one unfrozen block achieved the overall best results with 0.707 macro-averaged AUC on the test set and 0.209 recall. However, it had less false "No Finding" predictions compared to the SwAV ResNet-50 model, as implied by the precision in Table 3 that shows a 1.4% reduction in false "No Finding" classifications. However, it appears that the model was more confused in differentiating the anomalies due to the lower recall and F1 score. The results are pretty much similar when the number of unfrozen blocks increased, but the AUC plummeted when six transformer blocks were unfrozen.

ConvNeXt-V2, which was claimed to surpass numerous DL models in computer vision [13], obtained the worst overall performance in this experiment. With 0.703 AUC and 0.184 recall as shown in Table 4, the model was inferior to both of the previous baselines. Almost 30% of the model's "No Finding" predictions were false, regardless of the number of unfrozen convolution blocks. However, this also shows how consistent the model's performance is. Given how the model was previously trained on ImageNet, pre-training ConvNeXt-V2 on medical images could be promising. Overall, the best model has one unfrozen convolution block.

DINOV2 ViT-S/14 obtained the overall lowest validation loss among all models with very slight overfitting observed. As written in Table 5, the best validation loss was 1.753 despite being magnified by the class weights. In the test set, the model with five unfrozen transformer

TABLE 3. Evaluation results on different numbers of unfrozen blocks for DINO ViT-S/16. The asterisk (*) indicates the best model.

Unfrozen	Train loss	Val Loss	Test Loss	Test AUC	Recall	F1	Precision of "No Finding"
0	2.370	2.071	2.287	0.544	0.104	0.097	0.637
1*	1.959	1.912	2.124	**0.707**	0.209	**0.170**	0.753
2	1.999	**1.886**	**2.078**	0.697	0.198	0.166	0.740
3	**1.732**	1.959	2.173	0.690	**0.238**	0.161	**0.806**
4	2.039	1.927	2.112	0.706	0.218	0.167	0.762
5	2.063	1.895	2.111	0.695	0.189	0.163	0.751
6	2.227	2.008	2.167	0.606	0.121	0.139	0.693

TABLE 4. Evaluation results on different numbers of unfrozen blocks for ConvNeXt-V2. The asterisk (*) indicates the best model.

Unfrozen	Train loss	Val Loss	Test Loss	Test AUC	Recall	F1	Precision of "No Finding"
0	2.099	1.957	2.198	0.689	0.154	0.116	0.708
1*	**1.949**	1.911	2.150	**0.703**	**0.184**	0.132	**0.709**
2	2.116	1.959	2.189	0.682	0.155	0.137	0.707
3	2.070	1.940	**2.097**	0.697	0.166	**0.145**	0.707
4	2.096	**1.895**	2.165	0.688	0.156	0.131	0.682
5	2.070	1.927	2.178	0.696	0.159	0.138	0.697
6	2.115	1.940	2.167	0.694	0.156	0.131	0.696

blocks obtained the best results, with 0.743 AUC and 0.233 recall. The classification metrics are lower than SwAV ResNet-50, but the AUC is higher. This indicates comparable performance to the baseline with better confidence, as the higher AUC implied more distinct differentiations in the model's output probabilities. However, the results for all models pre-trained on ImageNet were inadequate, which can be attributed to the huge distinction between natural images in ImageNet and medical images [32]. Future studies may specifically explore pre-training on medical image datasets.

To validate whether the inadequate results were solely caused by the pre-training dataset domain gap with the downstream dataset, CheXNet was also fine-tuned in this experiment. From

TABLE 5. Evaluation results on different numbers of unfrozen blocks for DINOV2 ViT-S/14. The asterisk (*) indicates the best model.

Unfrozen	Train loss	Val Loss	Test Loss	Test AUC	Recall	F1	Precision of "No Finding"
0	2.264	2.022	2.156	0.651	0.105	0.095	0.673
1	1.699	**1.753**	**1.916**	0.738	0.189	0.166	**0.733**
2	1.596	1.789	2.007	0.729	0.221	0.185	0.709
3	1.657	1.795	1.995	0.722	0.206	0.173	0.723
4	1.608	1.789	2.008	0.734	0.211	0.177	0.713
5*	**1.565**	1.756	1.970	**0.743**	**0.233**	**0.186**	0.720
6	1.828	1.827	2.035	0.680	0.178	0.154	0.710

TABLE 6. Evaluation results on different numbers of unfrozen blocks for CheXNet. The asterisk (*) indicates the best model.

Unfrozen	Train loss	Val Loss	Test Loss	Test AUC	Recall	F1	Precision of "No Finding"
0*	2.381	2.244	**2.353**	**0.773**	0.328	**0.195**	0.749
1	2.069	1.809	2.661	0.766	**0.351**	0.183	**0.798**
2	1.763	**1.772**	3.059	0.745	0.305	0.160	0.766
3	**1.699**	1.844	2.953	0.731	0.305	0.155	0.783
4	1.974	1.920	2.930	0.740	0.294	0.151	0.793

the results shown in Table 6, it appears that even CheXNet that was pre-trained on the same NIH CXR-14 dataset performed poorly. However, it should be noted that during the pre-training of CheXNet, the "No Finding" class was not used. This means that the addition of "No Finding" shifted the model's performance greatly, resulting in only 0.773 AUC achieved with 0.328 recall. Moreover, 25.1% of the "No Finding" predictions were false. Therefore, deeper analysis will be necessary, which are described in the following subsections. Overall, zero-adaption worked best on CheXNet, which may be caused by its pre-trained parameters already having sufficient knowledge to extract noteworthy features from CXR images.

4.2. Comparison Between Models.

TABLE 7. AUC scores for each class in the test set. The asterisks (*) indicate the best model between CheXNet and those pre-trained in ImageNet.

Class (Proportion in Train Set)	SwAV	DINO	ConvNeXt-V2	DINOV2	CheXNet
Atelectasis (3.777%)	0.719	0.671	0.675	**0.735**	0.741 *
Cardiomegaly (5.159%)	**0.892**	0.823	0.845	0.863	0.897 *
Consolidation (5.508%)	**0.715**	0.701	0.697	0.714	0.732 *
Edema (2.638%)	0.831	0.828	**0.832**	0.822	0.842 *
Effusion (3.084%)	0.775	0.747	0.755	**0.790**	0.811 *
Emphysema (3.899%)	0.827	0.762	0.738	**0.874** *	0.867
Fibrosis (3.658%)	0.727	0.753	0.740	**0.813** *	0.780
Hernia (0.431%)	0.859	0.854	0.845	**0.861**	0.964 *
Infiltration (8.104%)	0.567	0.596	0.573	**0.598**	0.603 *
Mass (11.256%)	0.720	0.651	0.700	**0.740**	0.803 *
No Finding (22.341%)	0.628	0.626	0.621	**0.633** *	0.619
Nodule (14.923%)	**0.682**	0.627	0.624	0.653	0.760 *
Pleural Thickening (5.425%)	0.676	0.629	0.625	**0.680**	0.708 *
Pneumonia (1.559%)	0.552	**0.597**	0.562	0.574	0.612 *
Pneumothorax (8.237%)	0.778	0.740	0.712	**0.790**	0.850 *

To further analyze the models' performance, the class-specific AUC scores were analyzed. From the scores listed in Table 7, DINOV2 ViT-S/14 was the best model out of the four models pre-trained using ImageNet. The former obtained the best AUC scores on 10 out of 15 classes in the dataset, even outperforming CheXNet on three classes, namely "Emphysema", "Fibrosis", and "No Finding". Such results may be attributed to the fact that DINOV2 trained ViT models to also pay attention to patch-level features in addition to global image features [12], effectively combining local and global attention. Ironically, the original DINO failed to perform well compared to other models. CheXNet achieved better results overall, but it can be seen that the AUC for most classes were moderate. Only "Cardiomegaly", "Emphysema", and "Hernia" got more than 0.85 AUC on CheXNet. This indicates that the other classes' features may be more

DINOV2 ViT-S/14															
Ground Truth	Predictions														
	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Emphysema	Fibrosis	Hernia	Infiltration	Mass	No Finding	Nodule	Pleural Thickening	Pneumonia	Pneumothorax
Atelectasis	**385**	19	2	8	138	38	2	4	79	9	94	1	0	1	21
Cardiomegaly	51	**97**	0	9	21	2	4	0	80	1	50	1	0	0	0
Consolidation	81	5	**5**	14	174	17	3	0	141	5	30	0	0	0	6
Edema	23	1	1	**39**	32	4	0	0	112	3	11	3	0	0	2
Effusion	167	8	5	19	**707**	35	4	1	133	6	34	1	5	1	41
Emphysema	33	0	0	0	26	**169**	3	0	20	2	33	1	1	0	17
Fibrosis	27	4	0	1	32	8	**23**	0	21	2	48	0	2	0	8
Hernia	16	0	0	0	4	2	1	**7**	3	0	11	0	0	0	1
Infiltration	318	55	10	109	443	69	22	1	**843**	24	266	15	2	1	42
Mass	52	2	1	8	94	22	1	0	50	**79**	114	10	0	0	10
No Finding	1572	228	34	170	1630	473	95	14	2647	170	**2424**	47	15	5	337
Nodule	59	5	0	10	67	14	9	0	78	54	119	**19**	2	1	20
Pleural Thickening	48	2	3	2	122	14	12	0	40	5	47	0	**1**	0	13
Pneumonia	16	4	1	2	16	3	2	0	28	1	15	0	0	**0**	0
Pneumothorax	109	1	0	4	237	183	12	0	86	23	69	2	9	1	**217**

CheXNet															
Ground Truth	Predictions														
	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Emphysema	Fibrosis	Hernia	Infiltration	Mass	No Finding	Nodule	Pleural Thickening	Pneumonia	Pneumothorax
Atelectasis	**226**	74	29	32	66	71	42	6	23	46	81	30	15	0	60
Cardiomegaly	12	**226**	7	22	6	11	5	0	5	7	10	2	3	0	0
Consolidation	39	31	**72**	99	74	19	22	1	35	30	23	15	4	0	17
Edema	14	12	18	**105**	12	4	7	0	25	4	7	16	2	0	5
Effusion	59	76	89	123	**431**	31	46	2	25	66	39	15	70	0	95
Emphysema	3	3	4	6	3	**164**	22	1	5	16	26	13	10	0	29
Fibrosis	5	13	0	2	4	8	**46**	0	6	16	33	17	8	0	18
Hernia	1	3	0	0	0	1	1	**29**	0	3	5	0	0	0	2
Infiltration	144	247	225	436	103	131	110	9	**207**	121	227	124	30	0	106
Mass	12	21	10	18	22	26	13	1	9	**210**	33	32	14	0	22
No Finding	741	1244	395	693	561	723	722	88	615	771	**1808**	551	275	3	671
Nodule	13	26	8	17	15	29	47	0	20	96	51	**93**	10	0	32
Pleural Thickening	14	21	11	14	35	23	50	2	7	10	32	14	**38**	0	38
Pneumonia	4	11	13	7	2	6	7	0	10	4	9	7	1	**0**	7
Pneumothorax	18	22	21	29	27	169	99	1	16	45	31	32	35	0	**408**

FIGURE 3. Confusion Matrix for DINOV2 ViT-S/14 and CheXNet. Yellow-colored cells indicate which model obtained more True Positives for each class.

subtle or ambiguous. From this point onward, only DINOV2 ViT-S/14, which achieved the best performance from the models pre-trained on ImageNet, and CheXNet will be further analyzed. In addition, we also identified several problematic classes to be analyzed further based on the low AUC scores, namely ”Atelectasis”, ”Consolidation”, ”Infiltration”, ”No Finding”, ”Pleural Thickening”, and ”Pneumonia”.

Figure 3 provides the confusion matrix for the two best models. For ”Atelectasis”, ”Effusion”, ”Infiltration”, and ”No Finding”, DINOV2 ViT-S/14 obtained significantly more True Positives compared to CheXNet, despite the latter’s higher AUC scores for the classes except ”No Finding”. These three anomalies primarily show more subtle features compared to other such as ”Cardiomegaly” and ”Consolidation”, implying that DINOV2’s iBOT loss could possibly be the reason why the ViT model managed to outperform CheXNet. On the other hand, CheXNet classified remarkably more True Positives for most of the other classes. For classes

such as ”Cardiomegaly” (dilated or thicker heart), ”Nodule” (marked by lumps in the lungs), ”Hernia” (bulging lungs through the chest cavity), and ”Pneumothorax” (accumulated air in the pleural cavity), the features are sometimes visible directly on the thoracic regions, meaning that local attention could be a better fit for detecting them. Therefore, it is not strange that CheXNet would perform significantly better in classifying these classes. However, both models produced excessive misclassifications of images with ”No Findings”. DINOV2 ViT-S/14 tends to predict these images as either ”Atelectasis”, ”Effusion”, or ”Infiltration”. CheXNet seem to classify these images more evenly across other possible labels, meaning that this class specifically is very suspicious. To further verify this finding, t-SNE visualizations will be analyzed in the next subsection.

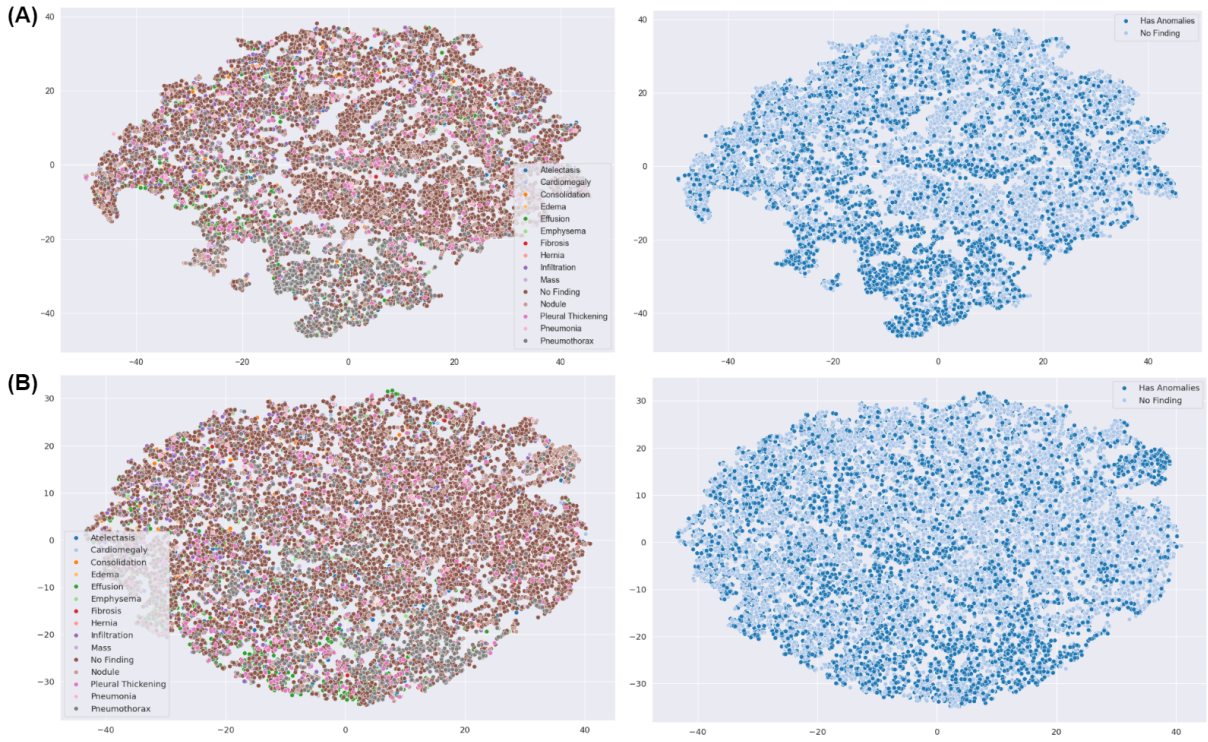

FIGURE 4. t-SNE visualizations for: (A) DINOV2 ViT-S/14 and (B) CheXNet.

4.3. Feature Separability. From the visualizations provided in Figure 4, no clear pattern can be made for any of the classes. The left side of the figure shows the dimension-reduced features for each class, while on the right side the labels are unified into merely two classes, namely ”Has Anomalies” and ”No Finding”. The ”Has Anomalies” label encompasses all data

from the 14 classes except ”No Finding”. This was done to analyze why the latter proved challenging for all models including CheXNet to classify. From the left side of the figure, it is clear why no models obtained more than 0.35 recall. Even the two best performing models failed to extract substantial information for each class, resulting in overlapping data points for each class in the scatterplot, both in Figure 4(A) and (B). However, binarizing the labels seem to show something notable. On the right side of the figure, DINO-V2 ViT-S/14 managed to group more ”No Finding” images together, implying that the features produced by the model are more similar. This explains why the model outperformed CheXNet both in AUC and True Positives for the ”No Finding” class. It is also possible that this is due to the fact that ViT models are more capable of extracting global features compared to CNNs [33].

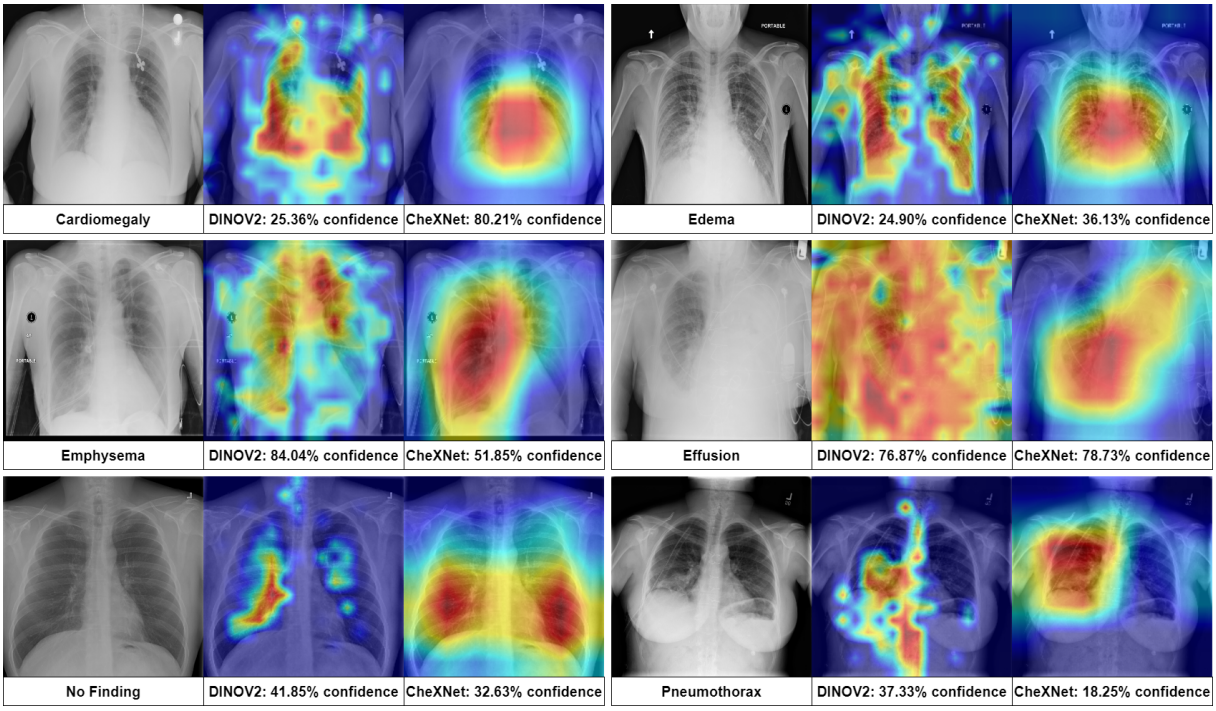

FIGURE 5. GradCAM Visualizations for DINO ViT-S/14 and CheXNet using samples from the dataset.

4.4. Visual Analysis. To further compare the two best models, GradCAM was used for visual analysis. Six samples from different classes were taken from the test set. The six samples shown in Figure 5 had correct predictions from both DINOV2-S/14 and CheXNet, hence they were chosen to visualize how well the model focused on the features deemed relevant. Cardiomegaly

is signified by the enlargement of the heart [34], so the model should focus on the regions of the heart, which was done correctly by CheXNet. However, DINOV2 ViT-S/14 focused on some regions around the lungs as well, which explains the lower confidence. Edema and Emphysema affects the lungs' internals, meaning that the heatmaps should focus on the areas inside the lungs. However, CheXNet focuses on the heart instead for Edema but correctly focuses on the right lungs in the sample with Emphysema. On the other hand, ViT focuses on the lung regions for both samples. Its confidence for the Emphysema sample is notable higher compared to CheXNet.

Effusion and Pneumothorax are signified by fluid build-ups and collapses around the lungs. In other words, the GradCAM should show more focus on the areas around the lungs. Unfortunately, DINOV2 ViT-S/14 seem to not perform too well on this aspect, as the heatmap for the Effusion sample got scattered focus nearly on the whole image, while it focused on the abdomen region as well for the Pneumothoras sample. CheXNet managed to focus on the lungs for both samples, making it superior to the ViT model despite the lower confidence on the Pneumothorax sample. One interesting finding is that CheXNet's heatmap was large for the "No Finding" sample as well while DINOV2 ViT-S/14 showed miniscule activation maps. This means that the CheXNet models found "symptoms indicating no findings", which is unusual when the goal is to detect anomalies. However, "No Finding" in the dataset means none of the 14 anomalies were found and it is possible that other unlisted anomalies exist in the "No Finding" class [25]. Future studies may further evaluate the GradCAM of the models for normal lungs as well. Overall, it can be concluded that CheXNet is still superior to DINOV2 ViT-S/14 in this experiment as the ViT model's focus still seem to be on the global features as opposed to CheXNet which focuses on local features, hence the compact attention maps for CheXNet in Figure 5.

4.5. Discussion. Table 8 displays the performance comparison between the five pre-trained models with two other studies that also utilized TL with the NIH CXR-14 dataset used in the downstream task. However, we did not find published papers that used all 15 classes of the dataset and obtained high AUC. Most studies did not use the "No Finding" class except for one [36], but only two other classes were used alongside "No Finding". Therefore, it is possible that

TABLE 8. Comparison of the five models with related previous studies that utilized NIH CXR-14 dataset.

Method	Classes	AUC	Recall	F1 Score	Precision of "No Finding"
Related Studies					
ResNet-50 [25]	8 classes excluding "No Finding"	0.696	-	-	-
MobileNetV2 for binary classification [35]	14 classes excluding "No Finding"	0.89	0.453	0.556	-
EfficientNet-V2 [36]	3 classes including "No Finding"	-	0.814	-	0.779
Baseline Replications					
SwAV ResNet-50 [3]	All 15 classes	0.730	0.253	0.204	0.739
DINO ViT-S/16 [11]		0.707	0.209	0.170	**0.753**
CheXNet [14]		**0.773**	**0.328**	**0.195**	0.749
Proposed Models					
ConvNeXt-V2	All 15 classes	0.703	0.184	0.132	0.709
DINOV2 ViT-S/14		0.743	0.233	0.186	0.720

the "No Finding" class itself is the problem. From Figures 3 and 4, even CheXNet was unable to effectively differentiate between "No Finding" and all other classes. However, DINOV2 ViT-S/14 still maintained the best AUC among all models pre-trained using the ImageNet dataset. The fact that none of the models came close to CheXNet may be caused by the discrepancies between the natural images in ImageNet and medical images [32], meaning that the models might still require another pre-training using medical image datasets.

Upon further exploration, a previous study had proven that label noises exist in the NIH CXR-14 dataset. This means that some data got wrong labels, which will obviously affect the performance of any ML models. The study involved three physicians and one radiologist, with each person labeling subsets of the dataset given to them. The study found that approximately 35% of the labels did not match the real conditions of the CXR images [20]. Hence, it is no wonder the models performed quite poorly, which is a major limitation in this study. Regardless, the GradCAM visualization still proved that simply performing TL from ImageNet to medical images did not yield adequate results. Therefore, future studies may further explore pre-training the models on medical images first. Furthermore, the NIH CXR-14 dataset can be said to be

more suitable for pre-training instead of downstream tasks. Given the 35% label noise, SSL pre-training may be the best method for pre-training models using this dataset.

5. Conclusion

Based on the experiment results, all models did not achieve remarkable results, including CheXNet. This could be caused by the noisy labels of the NIH CXR-14 dataset itself, especially in the ”No Finding” class, as found by a previous study [20]. However, DINOV2 ViT-S/14 emerged as the best model out of the other models pre-trained in ImageNet based on the AUC of most classes in the dataset. Based on the GradCAM visualizations, it can be inferred that the combination of DINO loss and iBOT loss does allow ViT models to possess the capability of local attention in addition to its global attention mechanism, which might explain its superior performance with an average AUC of 0.743. When compared to CheXNet, the DINOV2 ViT-S/14 model was still inferior in all metrics. However, it appears that using DINOV2 to pre-train ViT models directly using medical images may be a noteworthy opportunity to be explored. Due to the challenge of possible noisy labels, using the NIH CXR-14 dataset for downstream tasks is not recommended. The 0.195 overall F1 score of CheXNet had proven this as when the ”No Finding” class was not included, the same model obtained much better results. The t-SNE visualizations further supported this as CheXNet too was unable to extract meaningful representations to differentiate the ”No Finding” class with the other 14 anomalies.

Acknowledgement

The authors would like to express their utmost appreciation to BINUS University for funding this research and to NVIDIA-BINUS Artificial Intelligence Research & Development Center for supporting our research by allowing the usage of their NVIDIA Tesla P100 GPU.

Source of Funding

This study is funded by Bina Nusantara University as a part of the ”BINUS Research for Early Career Researchers” Research Grant entitled ”Pengembangan Model AI Berbasis Self-Supervised Learning Untuk Klasifikasi Penyakit Paru-paru Dengan Modul Visualisasi” with contract number 069A/VRRTT/III/2024 and contract date of March 18th, 2024.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] S.E. Vollset, H.S. Ababneh, Y.H. Abate, et al. Burden of Disease Scenarios for 204 Countries and Territories, 2022–2050: A Forecasting Analysis for the Global Burden of Disease Study 2021, The Lancet 403 (2024), 2204–2256. https://doi.org/10.1016/S0140-6736(24)00685-8.

[2] W.-C. Wang, E. Ahn, D. Feng, J. Kim, A Review of Predictive and Contrastive Self-Supervised Learning for Medical Images, Mach. Intell. Res. 20 (2023), 483–513. https://doi.org/10.1007/s11633-022-1406-4.

[3] 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.

[4] C. Zhang, H. Zheng, Y. Gu, Dive into the Details of Self-Supervised Learning for Medical Image Analysis, Med. Image Anal. 89 (2023), 102879. https://doi.org/10.1016/j.media.2023.102879.

[5] 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.

[6] H.-Y. Zhou, C. Lu, S. Yang, X. Han, Y. Yu, Preservational Learning Improves Self-Supervised Medical Image Models by Reconstructing Diverse Contexts, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, 2021: pp. 3479–3489. https://doi.org/10.1109/ICCV48922.2021.00348.

[7] K. Imagawa, K. Shiomoto, Evaluation of Effectiveness of Self-Supervised Learning in Chest X-Ray Imaging to Reduce Annotated Images, J. Imag. Inform. Med. 37 (2024), 1618–1624. https://doi.org/10.1007/s10278-024-00975-5.

[8] M. Shakouri, F. Iranmanesh, M. Eftekhari, DINO-CXR: A Self Supervised Method Based on Vision Transformer for Chest X-Ray Classification, in: G. Bebis, G. Ghiasi, Y. Fang, A. Sharf, Y. Dong, C. Weaver, Z. Leo, J.J. LaViola, L. Kohli (Eds.), Advances in Visual Computing, Springer, Cham, 2023: pp. 320–331. https://doi.org/10.1007/978-3-031-47966-3_25.

[9] J. Xiao, Y. Bai, A. Yuille, Z. Zhou, Delving into Masked Autoencoders for Multi-Label Thorax Disease Classification, in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2023: pp. 3577–3589. https://doi.org/10.1109/WACV56688.2023.00358.

[10] K. He, X. Chen, S. Xie, et al. Masked Autoencoders Are Scalable Vision Learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16000-16009.

[11] E. Selvano, A.Y. Paulindino, G.N. Elwirehardja, B. Pardamean, Evaluating Self-Supervised Pre-Trained Vision Transformer on Imbalanced Data for Lung Disease Classification, ICIC Express Lett. Part B Appl. 15 (2024), 83–89. https://doi.org/10.24507/icicelb.15.01.83.

[12] M. Oquab, T. Darcet, T. Moutakanni, et al. DINOv2: Learning Robust Visual Features without Supervision, arXiv:2304.07193 [cs.CV] (2024). https://doi.org/10.48550/arXiv.2304.07193.

[13] S. Woo, S. Debnath, R. Hu, et al. ConvNeXt V2: Co-Designing and Scaling ConvNets with Masked Autoencoders, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, Canada, 2023: pp. 16133–16142. https://doi.org/10.1109/CVPR52729.2023.01548.

[14] P. Rajpurkar, J. Irvin, K. Zhu, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, arXiv:1711.05225 [cs.CV] (2017). https://doi.org/10.48550/arXiv.1711.05225.

[15] H.H. Muljo, B. Pardamean, K. Purwandari, T.W. Cenggoro, 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.

[16] M. Mamalakis, A.J. Swift, B. Vorselaars, et al. DenResCov-19: A Deep Transfer Learning Network for Robust Automatic Classification of COVID-19, Pneumonia, and Tuberculosis from X-Rays, Comput. Med. Imag. Graph. 94 (2021), 102008. https://doi.org/10.1016/j.compmedimag.2021.102008.

[17] 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.

[18] S. Singh, M. Kumar, A. Kumar, et al. Efficient Pneumonia Detection Using Vision Transformers on Chest X-Rays, Sci. Rep. 14 (2024), 2487. https://doi.org/10.1038/s41598-024-52703-2.

[19] T. Chen, I. Philippi, Q.B. Phan, et al. A Vision Transformer Machine Learning Model for COVID-19 Diagnosis Using Chest X-Ray Images, Healthc. Anal. 5 (2024), 100332. https://doi.org/10.1016/j.health.2024.100332.

[20] R. Jang, N. Kim, M. Jang, et al. Assessment of the Robustness of Convolutional Neural Networks in Labeling Noise by Using Chest X-Ray Images From Multiple Centers, JMIR Med. Inform. 8 (2020), e18089. https://doi.org/10.2196/18089.

[21] M. Caron, H. Touvron, I. Misra, et al. Emerging Properties in Self-Supervised Vision Transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9650-9660.

[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, 2020, pp. 9912–9924.

[23] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, in: Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 1597–1607.

[24] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986.

[25] X. Wang, Y. Peng, L. Lu, et al. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, 2017: pp. 3462–3471. https://doi.org/10.1109/CVPR.2017.369.

[26] T. Rahman, A. Khandakar, Y. Qiblawey, et al. Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection Using Chest X-Ray Images, Comput. Biol. Med. 132 (2021), 104319. https://doi.org/10.1016/j.compbiomed.2021.104319.

[27] B.K. Umri, E. Utami, M.P. Kurniawan, Comparative Analysis of CLAHE and AHE on Application of CNN Algorithm in the Detection of Covid-19 Patients, in: 2021 4th International Conference on Information and Communications Technology (ICOIACT), IEEE, Yogyakarta, Indonesia, 2021: pp. 203–208. https://doi.org/10.1109/ICOIACT53268.2021.9563980.

[28] J. Zhou, C. Wei, H. Wang, et al. iBOT: Image BERT Pre-Training with Online Tokenizer, arXiv:2111.07832 [cs.CV] (2021). https://doi.org/10.48550/ARXIV.2111.07832.

[29] A. Sablayrolles, M. Douze, C. Schmid, H. Jégou, Spreading Vectors for Similarity Search, arXiv:1806.03198 [stat.ML] (2018). https://doi.org/10.48550/ARXIV.1806.03198.

[30] T.T. Cai, R. Ma, Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data, J. Mach. Learn. Res. 23 (2022), 1–54. http://jmlr.org/papers/v23/21-0524.html.

[31] S. Wang, Y. Zhang, Grad-CAM: Understanding AI Models, Comput. Mater. Contin. 76 (2023), 1321–1324. https://doi.org/10.32604/cmc.2023.041419.

[32] M. Gazda, J. Plavka, J. Gazda, P. Drotar, 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.

[33] A. Khan, Z. Rauf, A. Sohail, A.R. Khan, H. Asif, A. Asif, U. Farooq, A Survey of the Vision Transformers and Their CNN-Transformer Based Variants, Artif. Intell. Rev. 56 (2023), 2917–2970. https://doi.org/10.1007/s10462-023-10595-0.

[34] M. Gupta, A. Singh, Y. Kumar, Deep Learning for Prediction of Cardiomegaly Using Chest X-Rays, Neural Comput. Appl. 36 (2024), 19383–19391. https://doi.org/10.1007/s00521-024-10190-6.

[35] A. Souid, N. Sakli, H. Sakli, Classification and Predictions of Lung Diseases from Chest X-Rays Using MobileNet V2, Appl. Sci. 11 (2021), 2751. https://doi.org/10.3390/app11062751.

[36] S. Kim, B. Rim, S. Choi, et al. Deep Learning in Multi-Class Lung Diseases' Classification on Chest X-Ray Images, Diagnostics 12 (2022), 915. https://doi.org/10.3390/diagnostics12040915.