Integrating Self-Supervised Learning with Nonlinear Classifiers in Lightweight Swin Transformer for X-Ray Image Classification

Tri-Thuc Vo[1], Thanh-Nghi Do[1,2]

[1] College of Information Technology, Can Tho University, 92000-Cantho, Vietnam

[2] UMI UMMISCO 209 (IRD/UPMC), Sorbonne University, Pierre and Marie Curie University - Paris 6, France

Correspondence: Thanh-Nghi Do, email: dtnghi@ctu.edu.vn

Received: 30/07/2024, revised: 23/09/2024, accepted: 27/10/2024

Digital Object Identifier: 10.32913/mic-ict-research-vn.v2024.n3.1313

Abstract: In this paper, we present a new approach about the integration Self-Supervised Learning with nonlinear Classifiers in Lightweight Swin Transformer (*SSLnC−LSwinT*) for improving performance of X-ray image classification. Our approach leverages unlabeled data to address the issue of labeled data scarcity in the medical field by using self-supervised learning (SSL) to extract features. One of our key contributions is the introduction of the Lightweight SwinT architecture, a more lightweight variant of SwinT, designed to enhance computational efficiency, reduce model complexity, and shorten training time. To further improve classification efficiency, we propose the integration of a nonlinear classifier instead of a linear classifier in Lightweight SwinT. The experimental results underscore our contributions, demonstrating significant reductions in model training time and notable improvements in classification performance. Our proposed method, which integrates SSL based on LSwinT with a nonlinear LightGBM classifier, achieves an accuracy of up to 87%, improving by 1.8% over the non-LightGBM SwinT version and reducing training time significantly (3:23:00 vs. 7:37:29) compared to the original SwinT architecture.

Keywords: *Self-supervised learning, X-ray image, swin transformer, multi-class classification*

Tiêu đề: **Tích hợp học tự giám sát với thuật toán phân lớp phi tuyến trong kiến trúc Lightweight Swin Transformer để phân lớp hình ảnh X-quang**

Tóm tắt: Trong bài báo này, chúng tôi trình bày một phương pháp mới về tích hợp học tự giám sát (SSL) với thuật toán phân lớp phi tuyến trong kiến trúc Lightweight Swin Transformer (*SSLnC−LSwinT*) nhằm cải thiện hiệu quả phân lớp ảnh X-quang. Phương pháp của chúng tôi nhằm khai thác dữ liệu chưa được gán nhãn để giải quyết vấn đề khan hiếm dữ liệu có nhãn trong lĩnh vực y tế bằng cách tiếp cận học tự giám sát để học và trích xuất đặc trưng. Một trong những đóng góp quan trọng của chúng tôi là giới thiệu kiến trúc Lightweight SwinT, một biến thể của SwinT với kiến trúc đơn giản hơn, được đề xuất nhằm nâng cao hiệu quả tính toán, giảm độ phức tạp của mô hình và rút ngắn thời gian huấn luyện. Để cải thiện hiệu quả phân lớp, chúng tôi đề xuất tích hợp thuật toán phân lớp phi tuyến thay vì bộ phân lớp tuyến tính trong Lightweight SwinT. Kết quả thực nghiệm nhấn mạnh các đóng góp của chúng tôi, với sự giảm đáng kể thời gian huấn luyện mô hình và cải thiện đáng kể hiệu quả phân lớp. Phương pháp đề xuất, kết hợp SSL dựa trên LSwinT với thuật toán LightGBM, đạt độ chính xác 87%, tăng 1,8% so với phiên bản SwinT không sử dụng LightGBM và giảm đáng kể thời gian huấn luyện (3:23:00 so với 7:37:29) so với kiến trúc SwinT ban đầu.

Từ khóa: *Self-supervised learning, X-ray image, swin transformer, multi-class classification*

I. INTRODUCTION

The Covid-19 pandemic has caused global devastation, with millions of cases and deaths reported by the World Health Organization. Countries worldwide have faced overwhelmed healthcare systems and limited resources. In this context, chest X-ray (CXR) images have played a crucial role in diagnosing and monitoring Covid-19 patients. CXR images are also essential in diagnosing lung diseases due to their ability to provide detailed images of lung structures. They primarily help detect abnormalities, enabling doctors to identify and assess lung conditions such as tumors, nodules, infections, or injuries. The X-ray method is preferred over other imaging methods for diagnosing lung diseases

because it is significantly more cost-effective than technologies like MRI or PET, making it a more economical choice for both patients and healthcare facilities. Additionally, radiographs are widely available and accessible at various healthcare settings, from large hospitals to small clinics. Despite their usefulness, diagnosing lung diseases with X-rays involves a risk of misdiagnosis, largely dependent on the skill and experience of the radiologist. Therefore, incorporating supportive technologies like artificial intelligence is crucial to enhance accuracy and reduce errors in X-ray image diagnosis.

The emergence of Convolutional Neural Networks (CNNs) has significantly contributed to computer vision, particularly in medical image classification and aiding disease diagnosis from medical images [1]. CNNs transformed image classification and recognition tasks by learning features through convolutional and fully connected layers, significantly improving diagnostic accuracy for conditions seen in X-rays and MRIs [2]. Vision Transformers (ViT) [3] introduced the self-attention mechanism, originally developed for natural language processing, to image processing. Swin Transformer (SwinT) [4] was designed to address the limitations of ViT, such as handling large images and improving computational efficiency. It designs a hierarchical structure to capture features at various levels, enhancing the analysis of complex image details. Both ViTs and SwinT have been extensively studied in medical image analysis [5]. However, supervised learning approaches using CNNs or ViT face significant drawbacks, primarily the need for large labeled datasets, which can be expensive and time-consuming to obtain.

Self-supervised learning is a machine learning method that leverages unlabeled data to learn features and patterns, enhancing the ability to understand and classify data without human intervention in labeling. This method has achieved significant milestones and increased efficiency in natural image classification [6]. It implements various backbones in its architecture, such as CNNs [7], ViT [8], and SwinT [9]. By effectively utilizing unlabeled data and fine-tuning it with a small labeled dataset, this approach provides notable assistance in the medical field [10], where labeled data is scarce and data labeling costs are high. In [11], Mathilde Caron et al. explored how SSL with ViT. DINO, a self-distillation method using a student-teacher framework, enables ViT-Base to achieve 80.1% top-1 accuracy on ImageNet. The authors proposed MoCo-CXR [12], an adaptation of Momentum Contrast (MoCo) for improving CXR pathology detection. Models using MoCo-CXR-pretrained representations outperform those without, enhancing AUC on the smaller Shenzhen tuberculosis dataset. The authors proposed CheSS, a SSL method for CXR models, trained on a dataset of 4.8 million X-ray images [13]. CheSS achieved a 28.5% increase in accuracy for internal 6-class disease classification and a 1.3% mean AUC improvement on the CheXpert dataset, with an 11.4% increase using 1% of the data.

Our research introduces a new method to enhance CXR image classification while reducing model complexity and training time. The proposed Lightweight SwinT architecture (LSwinT), which is based on SwinT, significantly reduces training duration due to its simpler design, making it suitable for deployment in resource- and energy-constrained environments. For the goal of improving model performance, SSL is used with unlabeled CXR images to extract features from abundant unlabeled data, addressing the scarcity of labeled data in medical fields. Additionally, nonlinear classifiers are integrated instead of linear ones during the supervised learning phase with labeled data. Nonlinear classifiers effectively handle complex data and deliver high accuracy. The obtained results show that the nonlinear fine-tuned SSL pre-trained model outperforms the linear fine-tuned ImageNet pre-trained model. The findings also highlight that our proposed method, which leverages contributions from unlabeled X-ray images with SSL on LSwinT and the integration of a nonlinear classifier, enhances classification performance. Training time is significantly reduced with the LSwinT architecture combined classifiers compared to SwinT model.

The paper is organized as follows: Section II reviews prior research on lung disease classification using X-ray images. Section III describes our proposed method. Section IV details the experimental results, and Section V summarizes the findings and suggests future research directions.

II. RELATED WORK

In analysis of CXR images, recent research has focused on developing machine learning and deep learning methods to improve diagnostic accuracy for various pathologies. In [14], the Stanford University team developed CheXNeXt, a 121-layer DenseNet CNN for detecting 14 pathologies, including pneumonia and pleural effusion, in CXR images. Their algorithm matches board-certified radiologists in detecting multiple thoracic pathologies. Chen K.C et al. presented a diagnostic system for pediatric lung diseases using CXR images [15]. A YOLOv3 model crops the lung field, and three multi-classification methods were compared. The system detects abnormalities with 92.47% accuracy and identifies specific conditions with accuracies ranging from 71.94% to 85.71%. The study in [16] proposed an approach to improve multi-label CXR classification using the Chest X-ray 14 dataset. By combining visual vectors from a fine-tuned ConvNeXt network with semantic vectors from

BioBERT and using a dual-weighted metric loss function, the model achieved an average AUC score of 0.826.

For diagnosing Covid-19 from CXR images, the authors in [17] proposed the CNN-ELM framework, which combines lightweight parallel CNN feature extraction with the classification power of the extreme learning machine algorithm. It achieved 90.92% accuracy and a 96.93% AUC for 17 classes. The research in [18] aimed to detect diseases from CXR images of Covid-19 patients, pneumonia patients, and healthy individuals using a hybrid feature extraction network named D3SENET. Feature vectors extracted by CNN models were reduced using feature selection algorithms and classified using SVM. The study [19] trained classifiers using features from a fine-tuned Momentum Contrast model with Resnet backbones. The results showed an accuracy increase of 1.5% to 4.8% compared to those achieved by MoCo with only fine-tuned the linear layer. In [20], Zahid Ullah et al. proposed a multi-task semi-supervised learning (MTSSL) framework using publicly available data from pneumonia, lung opacity, and pleural effusion tasks with the CheXpert dataset. They integrated an adversarial autoencoder within their MTSSL framework to enhance feature learning and maximize the benefits of multi-task learning. The authors in [21] proposed the classification method of X-ray images with three steps: pre-processing, feature extraction using HOG and LBP, and classification with different classifiers. This method achieved 98% classification accuracy.

The Vision Transformer has been specifically highlighted for its application in X-ray image classification [22]. In [23], the authors proposed the Input Enhanced Vision Transformer (IEViT) that improved performance on CXR images with various pathologies. Experiments on four datasets (tuberculosis, pneumonia, Covid-19) showed that IEViT outperformed ViT, achieving F1-scores between 96.39% and 100%, with up to a +5.82% improvement over ViT. The study in [24] compared various optimization methods for ViT models in predicting lung diseases from a dataset of 19,003 CXR images. The optimization method with RAdam achieved 95.87% accuracy with ViT on balanced classes, while FastViT with NAdam reached 97.63% accuracy on imbalanced classes. In [25], the authors assessed interpretation methods for ViTs in CXR classification. Layerwise Relevance Propagation outperforms other methods, offering better insights into ViT learning. Lan Huang et al. proposed a method to improve ViT for CXR image analysis by adding a sliding window approach for better lesion detection, an attention region module, and a parallel network for patient metadata integration [26]. The model achieved an average AUC of 0.831, with sensitivity of 0.863, specificity of 0.821, and accuracy of 0.834 in diagnosing 14 diseases. In [27], the authors introduced a hybrid workflow combining an ensemble five CNNs with a Transformer encoder for disease identification.

Self-supervised learning is seen as a promising approach with significant contributions to medical image classification, including X-rays, by leveraging large amounts of unlabeled data to learn features [10, 28]. In [29], CheXzero, a deep learning model via SSL developed by Stanford University's research team, automates the detection and classification of thoracic pathologies in CXR images. It outperforms fully supervised models in external validation by detecting three out of eight pathologies and generalizing well across various tasks and datasets. In [30], the authors introduced DINO-CXR, an adaptation of DINO using a vision transformer for CXR classification. It outperforms its compared methods in accuracy, AUC and F1 scores, and requires less labeled data. In [31], Jingfeng Yao et al. presented EVA-X, a SSL model that effectively captures semantic and geometric features from unlabeled X-rays. It excels in detecting over 20 chest diseases and performs well in 11 medical tasks, reducing the need for annotated data. In [32], the authors proposed a new self-supervised transfer learning approach for detecting Covid-19 from CXR images. The method outperformed six SSL methods and six pretrained DCNNs, achieving high scores: AUC 0.999 and four-class accuracy 0.953 on the Covid-19 CXR dataset.

III. METHOD OF X-RAY IMAGE CLASSIFICATION

1. MoBY Self-Supervised Learning

The MoBY method was introduced by Zhenda Xie et al. in [9] with the Swin Transformer (SwinT) as the backbone of the architecture. The SwinT architecture is hierarchically structured with varying transformer block numbers at each stage. Stage 1 uses 2 blocks to process 4×4 image patches, generating basic features. Stage 2, with 2 blocks, captures localized interactions in downsampled features. Stage 3 significantly expands in complexity with 6 blocks for high-level feature extraction from further downsampled maps. Stage 4, with 2 blocks, refines these features for final classification. MoBY combines the strengths of MoCo v2 [7] and BYOL [33]. The architecture of MoBY includes an online encoder and a target encoder. The target encoder consists of a backbone and a projector, while the online encoder is similarly designed but includes an additional predictor. The online encoder undergoes updates through gradient-based optimization, whereas the target encoder is updated by taking a moving average of the parameters from the online encoder. The architecture inherits several key elements from MoCo v2, including its momentum

design, key queue, and the use of the contrastive loss function InfoNCE, all of which contribute to robust feature representation learning. The loss function measures the similarity of image pairs, aiming to maximize the similarity between positive pairs and minimize it between negative pairs. Positive image pairs are formed by applying two data augmentation techniques on the same image, and the loss function aims to bring their representations closer together to minimize their distance. Negative image pairs comprise dissimilar images, and the loss function maximizes their distance. Additionally, MoBY incorporates aspects from BYOL, such as the use of asymmetric encoders, a variety of data augmentation techniques, and a momentum scheduler.

In this paper, we propose a new approach with SSL based on a simplified architecture of SwinT. Our goal is to leverage the strengths of SSL to extract features from unlabeled medical data, particularly X-ray images. Given the limited dataset and training resources, we propose new architectures based on the SwinT architecture by reducing the number of blocks in each stage. We introduce Lightweight SwinT (LSwinT) with two versions: LSwinT1, with 1 block per stage, and LSwinT2, with 2 blocks per stage. Reducing the model size enhances computational efficiency. This approach reduces both training and inference times, making it suitable for resource-constrained environments.

2. Self-Supervised Learning with nonlinear Classifiers in Lightweight Swin Transformer (SSLnC-LSwinT)

The algorithm *SSLnC−LSwinT* (as illustrated in Fig. 1 and Algorithm 1) unfolds in a structured sequence comprising three distinct stages. Initially, in the SSL phase, the model is set up with Enc_{online} and Enc_{target}, both based on backbone like the LSwinT architecture. During training, the model processes the unlabeled dataset $\mathcal{D}_{\text{unlabeled}}$ where data augmentation techniques such as cropping, flipping, jitter, grayscale, and GaussianBlur, are employed to generate positive pairs. The *InfoNCE* loss function guides the updates for $\text{Enc}_{\text{online}}$, while $\text{Enc}_{\text{target}}$ is refined using a momentum-based approach with a parameter m. The parameters for this step are set to 300 epochs, learning rate (0.001), weight decay (0.05), and momentum (0.99). This results in the creation of Model_1. In the subsequent fine-tuning stage, Model_1 is utilized with its backbone frozen to train a linear classifier on the labeled datasets, which include $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{valid}}$. The performance of this fine-tuned model, referred to as Model_2, is then evaluated its effectiveness through $\mathcal{D}_{\text{test}}$. Instead of freezing Backbone Model_1 and training the classifier layer, we extract its features and train a nonlinear classifier. The nonlinear classifier is integrated as a replacement for the linear classifier in the Lightweight Swin Transformer, allowing us to leverage its capabilities

Algorithm 1: Self-supervised learning with nonlinear classifiers in LSwinT

Inputs: $\mathcal{D}_{\text{unlabeled}}$, $\mathcal{D}_{\text{labeled}}$: ($\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{valid}}$, $\mathcal{D}_{\text{test}}$), Classifiers: (LightGBM, XGBoost, CatBoost), m: momentum
Outputs: Model_1, Model_2, classifier models
begin
  Phase 1: Self-Supervised Learning
  Initialize model
  $\text{Enc}_{\text{online}}$
  $\text{Enc}_{\text{target}}$
  for *each epoch* **do**
    for *each minibatch in* $\mathcal{D}_{unlabeled}$ **do**
      Update $\text{Enc}_{\text{online}}$
      Update $\text{Enc}_{\text{target}}$ using momentum:
      $\text{Enc}_{\text{target}} \leftarrow m \cdot \text{Enc}_{\text{target}} + (1-m) \cdot \text{Enc}_{\text{online}}$
    end
  end
  Output 1: Model_1
  Phase 2: Supervised Learning
  Load Model_1 from Phase 1
  Fine-tuning
    Freeze Model_1 backbone
    Train the linear classifier on $\mathcal{D}_{\text{labeled}}$
  Output 2: Model_2
  Phase 3: Integrating Nonlinear Classifiers
  Feature extraction
  $\text{Model}_2' \leftarrow \text{Model}_2$ removed the last layer
  for *each set* $\in \{\mathcal{D}_{train}, \mathcal{D}_{valid}, \mathcal{D}_{test}\}$ **do**
    $\mathcal{D}'(features, labels) \leftarrow \text{Model}_2'(set)$
  end
  Training nonlinear classifiers
  for *each classifier clf* $\in$ *Classifiers* **do**
    Train clf on $\mathcal{D}'(features, labels)$
  end
  Output 3: classifier models
end

to enhance classification efficiency. The final stage focuses on integrating a nonlinear classifier. In this step, Model_2 is used to extract features from the labeled datasets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{valid}}$, and $\mathcal{D}_{\text{test}}$ by removing its final linear layer. Various nonlinear classifiers, including *LightGBM, XGBoost, and CatBoost*, are then trained on these extracted features with learning rate (0.23) and deepth (8). The performance of each classifier is evaluated using $\mathcal{D}_{\text{test}}$.

3. Nonlinear classifiers

LightGBM, short for Light Gradient Boosting Machine, is a highly efficient gradient boosting framework developed

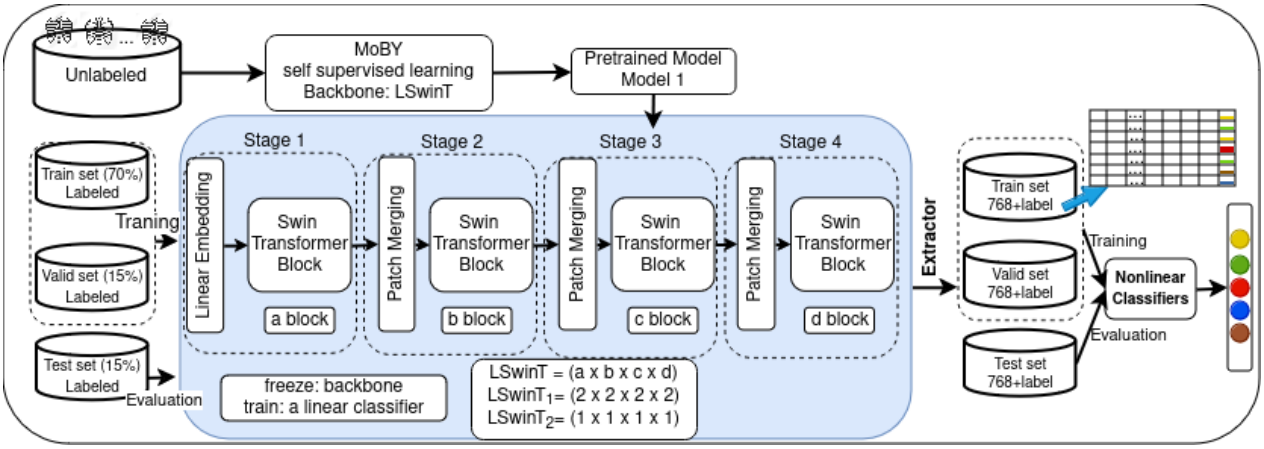

Figure 1. Diagram of *SSLnC−LSwinT* for chest X-ray image classification.

by Microsoft [34]. It is designed to be fast, distributed, and capable of handling large datasets with high accuracy. LightGBM used a novel technique called Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to significantly improve the training speed and reduce memory usage. GOSS selectively retains instances with large gradients to maintain the accuracy, while EFB bundles mutually exclusive features to reduce the number of features. This results in faster computation without compromising model performance.

XGBoost [35], which stands for Extreme Gradient Boosting, is a highly efficient machine learning algorithm optimized for speed and performance. Developed by Tianqi Chen, XGBoost enhances traditional gradient boosting by incorporating a variety of advanced features. These include regularization to prevent overfitting, parallel processing to speed up computations. XGBoost also supports tree pruning and built-in cross-validation, which contribute to its robust performance. The algorithm can handle large datasets with high dimensionality and is highly scalable, making it suitable for complex tasks.

CatBoost [36] is an effective gradient boosting algorithm developed by Yandex, optimized for efficiently handling categorical features. Unlike traditional gradient boosting methods that often require preprocessing of categorical data, CatBoost incorporates categorical variables directly into its training process. This method helps to prevent overfitting by using permutation-driven target statistics for categorical features. CatBoost also leverages symmetric trees and gradient-based optimization to enhance model accuracy and reduce training time.

4. Chest X-Ray image data set

Our study about *SSLnC−LSwinT* used two datasets: an unlabeled one from the CheXpert dataset, containing over 200,000 X-ray images, with 120,000 images used for training the MoBY model, and a labeled dataset with 98,996 images across five classes (normal, Covid-19, edema, mass nodule, pneumothorax) from various sources as in [19]. Both datasets were resized to 224x224 pixels. The labeled dataset was divided into three subsets that include training (70%), validation (15%), and test (15%). Details are in Table I, and visual examples are shown in Fig. 2.

Table I
DATASET OF CHEST X-RAY IMAGES.

Label	Trainset	Validset	Testset
Normal	18,425	3,949	3,948
Covid-19	14,252	3,054	3,054
Edema	23,240	4,980	4,980
Mass-nodule	4,077	873	874
Pneumothorax	9,303	1,993	1,994

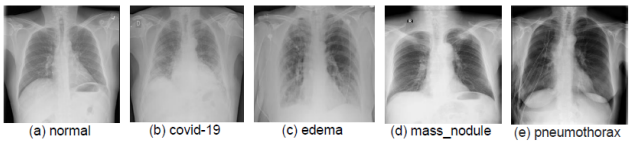

Figure 2. Sample of chest x-ray images for each class

IV. EXPERIMENTAL RESULTS

A Python program was developed for *SSLnC−LSwinT* in CXR image classification, utilizing SSL and fine-tuning with SwinT and LSwinT architectures. Scikit-learn [37] and PyTorch [38] were used for implementation. Experiments were conducted on a computer running Ubuntu 22.04.3, equipped with an AMD® Ryzen 9 5900x 12-core processor, 32 GB of RAM, and an NVIDIA GeForce GTX 1080 Ti

Table II
ACCURACY AND TRAINING TIME FOR DIFFERENT MODELS.

Model	**Accuracy (%)**	**Training Time (hh:mm:ss)**
MI-SwinT	81.8	07:31:47
MI-SwinT+LightGBM	85.3	07:34:52
MI-SwinT+CatBoost	84.2	08:59:30
MI-SwinT+XGBoost	85.0	07:45:29
MS-SwinT	85.2	07:37:29
MS-SwinT+LightGBM	87.2	07:40:35
MS-SwinT+CatBoost	86.4	09:05:08
MS-SwinT+XGBoost	87.0	07:51:09
MS-LSwinT1	82.6	03:20:00
MS-LSwinT1+LightGBM	87.0	03:23:00
MS-LSwinT1+CatBoost	86.5	04:49:04
MS-LSwinT1+XGBoost	87.1	03:34:18
MS-LSwinT2	84.1	05:59:02
MS-LSwinT2+LightGBM	87.0	06:02:03
MS-LSwinT2+CatBoost	86.3	07:13:35
MS-LSwinT2+XGBoost	86.9	06:12:58

featuring 11 GB of GDDR5X memory and 3584 CUDA cores.

The SwinT architecture comprises four stages with the following number of blocks: Stage 1 has 2 blocks, Stage 2 has 2 blocks, Stage 3 has 6 blocks, and Stage 4 has 2 blocks. Our proposed LSwinT architecture has two versions: LSwinT1, with 1 block per stage, and LSwinT2, with 2 blocks per stage. MI-SwinT is fine-tuned from the ImageNet pretrained model using the SwinT architecture, while MS-SwinT is fine-tuned from a SSL model on unlabeled X-ray images using the same architecture. MS-LSwinT1 and MS-LSwinT2 are fine-tuned from a pretrained MoBY model with different architectures; MS-LSwinT1 uses the LSwinT1 architecture, and MS-LSwinT2 uses the LSwinT2 architecture.

The experimental results of the four models (MI-SwinT, MS-SwinT, MS-LSwinT1, and MS-LSwinT2) are shown in Tab. II and Fig. 3. MS-SwinT achieves the highest performance with an accuracy of 85.2%, outperforming MI-SwinT, which has an accuracy of 81.8%. This indicates that SSL on unlabeled X-ray images provides a significant advantage over pre-training with ImageNet for fine-tuning the linear classifier. MS-LSwinT1, with an accuracy of 82.6%, performs slightly better than MI-SwinT, highlighting the effectiveness of the LSwinT architecture even with fewer blocks. MS-LSwinT2, achieving an accuracy of 84.1%, also demonstrates superior performance compared to MI-SwinT but lower than MS-SwinT.

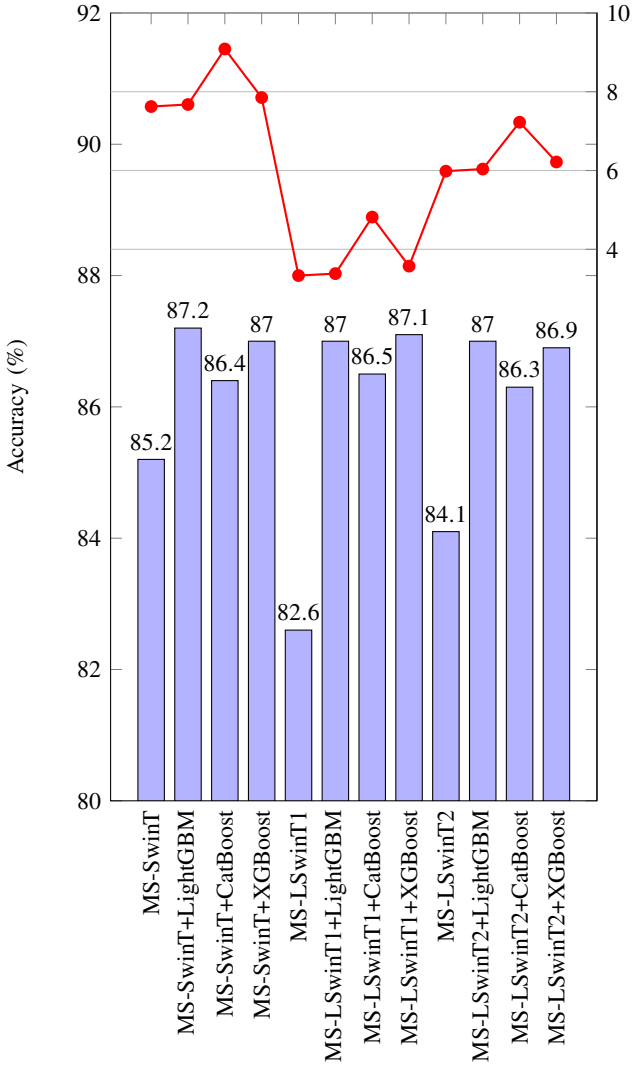

Figure 3. Accuracy and Training time for different models.

Integrating MS-LSwinT1 and MS-LSwinT2 as feature extractors with various nonlinear classifiers results in notable performance improvements in accuracy, showing no difference compared to MS-SwinT combined with nonlinear classifiers. Combining MS-SwinT with a classifier, LightGBM achieves the highest accuracy at 87.2%. XGBoost comes close with 87%, while CatBoost shows a slightly lower accuracy of 86.4%. When MS-LSwinT1 is paired with nonlinear classifiers, accuracy improves substantially. XGBoost performs well, reaching 87.1%, followed by LightGBM at 87.0% and CatBoost at 86.5%. Similarly, MS-LSwinT2's accuracy is markedly improved when combined with nonlinear classifiers, with the highest increase observed using LightGBM (+2.9%), followed by XGBoost (+2.8%), and CatBoost (+2.2%). These findings underscore that both MS-LSwinT1 and MS-LSwinT2, when paired with nonlinear classifiers, lead to enhanced feature

utilization and superior classification performance.

Training time for 100 epochs of fine-tuning a linear classifier reveals distinct differences among these models (see Tab. II and Fig. 3 with the red line). MI-SwinT and MS-SwinT, both utilizing the SwinT architecture, have nearly identical training times of 07:31:47 and 07:37:29, respectively. In contrast, the LSwinT architectures demonstrate significantly shorter training times, with MS-LSwinT1 completing in 03:20:00 and MS-LSwinT2 in 05:59:02. This difference highlights the efficiency of LSwinT architectures, likely due to their lightweight design, which reduces computational complexity and speeds up training and inference. The LSwinT models is particularly beneficial in resource-limited scenarios.

The training time reveals notable variations across models combined with boosting algorithms. For MS-SwinT, there are slight increases in training time when combined with LightGBM (07:40:35) and XGBoost (07:51:09). However, CatBoost significantly extends the training time to 09:05:08. Comparing the lightweight models, MS-LSwinT1 exhibits the shortest training time of 03:20:00. Incorporating boosting algorithms results in minimal time increases for LightGBM (03:23:00) and XGBoost (03:34:18), but a substantial increase with CatBoost, reaching 04:49:04. Similarly, MS-LSwinT2 requires 05:59:02 to train, with slight time increases for LightGBM (06:02:03) and XGBoost (06:12:58), while CatBoost again results in the longest training time at 07:13:35.

The results shows that combining MS-LSwinT1 and MS-LSwinT2 with classifiers, such as LightGBM, CatBoost, and XGBoost, leads to improved performance, aligning closely with MS-SwinT combined with the same classifiers. For example, MS-LSwinT1 with LightGBM achieves 87% accuracy, closely matching the 87.2% of MS-SwinT with LightGBM. MS-LSwinT1+XGBoost reaches 87.1%, which is nearly identical to MS-SwinT+XGBoost's 87%. Similarly, MS-LSwinT2 with these classifiers shows comparable performance, with MS-LSwinT2 using LightGBM and XGBoost achieving 87% and 86.9%, respectively. Combining MS-LSwinT1 and MS-LSwinT2 with classifiers shows no performance difference compared to MS-SwinT with the same classifiers, while significantly reducing training time. Specifically, MS-LSwinT1 using LightGBM trains in 03:23:00, over four hours faster than MS-SwinT with LightGBM, which takes 07:40:35. Similarly, MS-LSwinT1 with XGBoost finishes in 03:34:18, compared to 07:51:09 for MS-SwinT with XGBoost. Even MS-LSwinT2 shows time reductions, with training times of 06:02:03 using LightGBM and 06:12:58 with XGBoost, compared to 07:40:35 and 07:51:09 for the MS-SwinT model. These results highlight the efficiency of the lightweight architectures combined with nonlinear classifiers, delivering similar accuracy with significantly reduced training times.

The combination of MS-LSwinT1 with different classifiers demonstrates the accuracy improvement compared to MS-SwinT. Specifically, MS-LSwinT1 with LightGBM shows an accuracy increase of 1.8%, reaching 87%, compared to the 85.2% accuracy of MS-SwinT. Similarly, the combination of MS-LSwinT1 with CatBoost results in a 1.3% improvement. The most significant increase is observed with MS-LSwinT1 and XGBoost, which delivers a 1.9% increase, resulting in an accuracy of 87.1%. These improvements highlight the effectiveness of integrating lightweight models with advanced classifiers to enhance overall performance.

V. CONCLUSION AND FUTURE WORKS

In this paper, we introduce an innovative approach that combines Self-Supervised Learning with nonlinear classifiers in the Lightweight Swin Transformer (SSLnC-LSwinT) to enhance X-ray image classification performance. By leveraging unlabeled data, our method addresses labeled data scarcity in the medical field through SSL to learn features. We present the LSwinT architecture, a lightweight variant of SwinT, designed to improve computational efficiency, reduce model complexity, and shorten training time. Integrating a nonlinear classifier instead of a linear one further boosts classification efficiency. The study results demonstrate the effectiveness of our proposed approach. Fine-tuning a linear classifier for 100 epochs, MS-SwinT model takes 07:37:29, while LSwinT architectures are faster: MS-LSwinT1 at 03:20:00 and MS-LSwinT2 at 05:59:02. Results show that using MS-LSwinT1 with nonlinear classifiers enhances accuracy as MS-SwinT but with significantly reduced training time. The base MS-LSwinT1 model achieves 82.6% accuracy, which increases to 87.1% with XGBoost. The XGBoost provides the largest boost at +1.9%, followed by LightGBM at +1.8% compared to MS-SwinT.

Future work may incorporate clinical metadata to improve model performance and explore SSL techniques for on clinical metadata.

REFERENCES

[1] S. K. Zhou, H. Greenspan, and D. Shen, *Deep learning for medical image analysis*. Academic Press, 2023.

[2] K. Sailunaz, T. Özyer, J. Rokne, and R. Alhajj, "A survey of machine learning-based methods for covid-19 medical image analysis," *Medical & Biological Engineering & Computing*, vol. 61, no. 6, pp. 1257–1297, 2023.

[3] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[5] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[7] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[8] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9640–9649.

[9] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, "Self-supervised learning with swin transformers," 2021. [Online]. Available: https://arxiv.org/abs/2105.04553

[10] S.-C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: a systematic review and implementation guidelines," *NPJ Digital Medicine*, vol. 6, no. 1, p. 74, 2023.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[12] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," in *Medical Imaging with Deep Learning*. PMLR, 2021, pp. 728–744.

[13] K. Cho *et al.*, "Chess: Chest x-ray pre-trained model via self-supervised contrastive learning," *Journal of Digital Imaging*, vol. 36, no. 3, pp. 902–910, 2023.

[14] P. Rajpurkar and et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists," *PLoS medicine*, vol. 15, no. 11, p. e1002686, 2018.

[15] K.-C. Chen and et al., "Diagnosis of common pulmonary diseases in children by x-ray images and deep learning," *Scientific reports*, vol. 10, no. 1, p. 17374, 2020.

[16] Y. Jin, H. Lu, W. Zhu, and W. Huo, "Deep learning based classification of multi-label chest x-ray images via dual-weighted metric loss," *Computers in Biology and Medicine*, vol. 157, p. 106683, 2023.

[17] M. Nahiduzzaman and et al., "Parallel cnn-elm: A multiclass classification of chest x-ray images to identify seventeen lung diseases including covid-19," *Expert Systems with Applications*, vol. 229, p. 120528, 2023.

[18] M. Kaya and M. Eris, "D3senet: A hybrid deep feature extraction network for covid-19 classification using chest x-ray images," *Biomedical signal processing and control*, vol. 82, p. 104559, 2023.

[19] T.-T. Vo and T.-N. Do, "Improving chest x-ray image classification via integration of self-supervised learning and machine learning algorithms." *Journal of Information & Communication Convergence Engineering*, vol. 22, no. 2, 2024.

[20] Z. Ullah, M. Usman, and J. Gwak, "Mtss-aae: Multi-task semi-supervised adversarial autoencoding for covid-19 detection based on chest x-ray images," *Expert Systems with Applications*, vol. 216, p. 119475, 2023.

[21] S. U. Amin, S. Taj, A. Hussain, and S. Seo, "An automated chest x-ray analysis for covid-19, tuberculosis, and pneumonia employing ensemble learning approach," *Biomedical Signal Processing and Control*, vol. 87, p. 105408, 2024.

[22] S. Taslimi, S. Taslimi, N. Fathi, M. Salehi, and M. H. Rohban, "Swinchex: Multi-label classification on chest x-ray images with transformers," *arXiv preprint arXiv:2206.04246*, 2022. [Online]. Available: https://arxiv.org/abs/2206.04246

[23] G. I. Okolo, S. Katsigiannis, and N. Ramzan, "Ievit: An enhanced vision transformer architecture for chest x-ray image classification," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107141, 2022.

[24] J. Ko, S. Park, and H. G. Woo, "Optimization of vision transformer-based detection of lung diseases from chest x-ray images," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 191, 2024.

[25] P. Komorowski, H. Baniecki, and P. Biecek, "Towards evaluating explanations of vision transformers for medical imaging," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3726–3732.

[26] L. Huang, J. Ma, H. Yang, and Y. Wang, "Research and implementation of multi-disease diagnosis on chest x-ray based on vision transformer," *Quantitative Imaging in Medicine and Surgery*, vol. 14, no. 3, p. 2539, 2024.

[27] C. C. Ukwuoma *et al.*, "A hybrid explainable ensemble transformer encoder for pneumonia identification from chest x-ray images," *Journal of Advanced Research*, vol. 48, pp. 191–211, 2023.

[28] B. VanBerlo, J. Hoey, and A. Wong, "A survey of the impact of self-supervised pretraining for diagnostic tasks in medical x-ray, ct, mri, and ultrasound," *BMC Medical Imaging*, vol. 24, no. 1, p. 79, 2024.

[29] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning," *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1399–1406, 2022.

[30] M. Shakouri, F. Iranmanesh, and M. Eftekhari, "Dino-cxr: A self supervised method based on vision transformer for chest x-ray classification," in *International Symposium on Visual Computing*. Springer, 2023, pp. 320–331.

[31] J. Yao, X. Wang, Y. Song, H. Zhao, J. Ma, Y. Chen, W. Liu, and B. Wang, "Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning," 2024. [Online]. Available: https://arxiv.org/abs/2405.05237

[32] G. Li, R. Togo, T. Ogawa, and M. Haseyama, "Covid-19 detection based on self-supervised transfer learning using chest x-ray images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 4, pp. 715–722, 2023.

[33] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," 2020. [Online]. Available: https://arxiv.org/abs/2006.07733

[34] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[35] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[36] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[37] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12,

pp. 2825–2830, 2011.
[38] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," 2019. [Online]. Available: https://arxiv.org/abs/1912.01703

Tri-Thuc Vo received the B.Eng. degree in Software Engineering from the Cantho University, Vietnam, in 2011. He received his MSc. degree in informatics from the University of Brest, France, in 2018. He is currently a lecturer at the College of Information Technology, Cantho University, Vietnam. His research interests include medical data analysis and machine learning.
Email: vtthuc@ctu.edu.vn

Thanh-Nghi Do received his PhD. degree in informatics from the University of Nantes, France, in 2004. He is currently an associate professor at the College of Information Technology, Cantho University, Vietnam. He is also an associate researcher at UMI UMMISCO 209 (IRD/UPMC), Sorbonne University, and the Pierre and Marie Curie University, France. His research interests include data mining with support vector machines, kernel-based methods, decision tree algorithms, ensemble-based learning, and information visualization. He has served on the program committees of international conferences and is a reviewer for journals in his fields of expertise.
Email: dtnghi@ctu.edu.vn