IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 June 2024, accepted 22 June 2024, date of publication 24 June 2024, date of current version 1 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3418985

RESEARCH ARTICLE

Dual-Masked Autoencoders: Application to Multi-Labeled Pediatric Thoracic Diseases

TAEYOUNG YOON[1] **AND DAESUNG KANG**[2]

[1]Department of Healthcare Information Technology, Inje University, Gimhae-si 50834, Republic of Korea
[2]Department of Medical Information Technology, Inje University, Gimhae-si 50834, Republic of Korea

Corresponding author: Daesung Kang (danniskang@gmail.com)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by Korean Government [Ministry of Science and ICT (MSIT)] under Grant RS-2023-00249104.

ABSTRACT Pediatric thoracic diseases present significant health risks to children. While chest X-rays are commonly used for diagnosing thoracic diseases, interpreting pediatric images comes with unique challenges such as anatomical variations, developmental differences, and potential artifacts. Deep learning offers promise in addressing these challenges, yet its effectiveness is hindered by the limited availability of pediatric chest X-ray data. To overcome this limitation, we introduce the dual-masked autoencoders (dual-MAE) algorithm, consisting of online and target networks with encoder and decoder modules. These networks are optimized by minimizing three losses: between the reconstructed image of the online network and the target network, between the input image and the reconstructed image of the online network, and between the input image and the reconstructed image of the target network. To learn efficiently from pediatric chest X-rays, we employ a two-step training strategy: pretraining the dual-MAE model on adult chest X-rays, then fine-tuning it on pediatric X-rays for diagnosing multi-labeled pediatric thoracic diseases. The proposed model exhibited superior performance with the highest mean AUC score (0.752), surpassing the ResNet-34 (0.669) and ViT-S (0.645) trained from scratch. Additionally, the dual-MAE model outperformed the ResNet-34 (0.697) and ViT-S (0.638), both pretrained on the ImageNet dataset and then fine-tuned on pediatric chest X-rays. Despite being pretrained on a significantly smaller number of X-rays compared to the ImageNet dataset, our model demonstrated better performance. Furthermore, it outperformed the ResNet-34 (0.712), ViT-S (0.673), and vanilla MAE method (0.735), all pretrained on adult chest X-rays and fine-tuned on pediatric chest X-rays. Even with only 50% of labeled pediatric chest X-ray images, dual-MAE demonstrated comparable performance to that of the vanilla MAE method and outperformed ResNet-34 and ViT-S fine-tuned with 100% labeled pediatric chest X-ray images.

INDEX TERMS Adult chest X-rays, dual-masked autoencoders, masked autoencoders, pediatric chest X-rays, pediatric thoracic diseases.

I. INTRODUCTION

Pediatric thoracic diseases present substantial risks to children's health, affecting various aspects of their welfare. These conditions, ranging from respiratory issues such as pneumonia and bronchitis to congenital heart defects, can lead to compromised lung and cardiovascular function [1]. Children with thoracic diseases may face a higher susceptibility to infections, with an increasing potential for complications. Thoracic diseases during this critical developmental period may consequently have adverse effects on lung structure and function, increasing the risk of subsequent chronic lung disease [2]. Addressing these risks necessitates early diagnosis, comprehensive medical management, and a holistic approach to care to enhance the overall health and quality of life for affected children. Among diverse diagnosis meth-

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

ods, chest X-rays are commonly used to diagnose pediatric thoracic diseases by providing detailed images of the lungs. They are effective in visualizing abnormalities in lung structures, aiding in the identification of infections and congenital anomalies [3]. The non-invasive and rapid characteristics of X-rays render them suitable for pediatric patients, while their cost-effectiveness and widespread availability contribute to their frequent use in initial screenings.

However, interpreting pediatric chest X-rays presents unique challenges compared to adult chest X-rays due to various factors [4]. First, the smaller size of pediatric anatomy requires higher resolution imaging to discern intricate details, making abnormalities potentially harder to detect. Second, developing skeletal structures in children may introduce variations that could be mistaken for abnormalities. Additionally, children are more prone to respiratory motion artifacts during imaging, impacting image clarity. The dynamic characteristics of pediatric diseases, coupled with rapid changes in lung tissue, further complicates differentiation between normal and pathological findings. Furthermore, limited cooperation from pediatric patients may result in suboptimal positioning during imaging, further complicating interpretation. The combination of anatomical differences, developmental variations, respiratory motion artifacts, and the dynamic characteristics of pediatric diseases contributes to the increased complexity of interpreting pediatric chest X-rays compared to those of adults.

Deep learning has significantly contributed to the field of medical imaging [5], [6], [7], [8]. With the ability to process large amounts of medical imaging data, deep learning models demonstrate proficiency in recognizing intricate patterns and abnormalities in chest X-rays associated with thoracic diseases [6], [7], [8]. Therefore, employing deep learning algorithms for the interpretation of pediatric chest X-rays in diagnosing thoracic diseases could be a promising approach. However, the limited availability of pediatric chest X-ray data presents a significant obstacle to achieving optimal results. The scarcity of diverse and comprehensive datasets hampers the ability of deep learning models to generalize effectively across various pediatric cases. This limitation can lead to suboptimal performance and hinder the algorithm's capacity to accurately identify and differentiate between normal and pathological findings in pediatric chest X-rays.

This study proposes a novel solution to address challenges associated with limited pediatric chest X-ray data: the dual-masked autoencoders (dual-MAE) algorithm. The dual-MAE aims to enhance the performance of deep learning models in pediatric chest X-rays by addressing the constraints of limited datasets. By leveraging the proposed algorithm, it seeks to offer a more robust and accurate tool for interpreting pediatric chest X-rays, contributing to the improvement of generalization capabilities in the pediatric thoracic imaging domain. We summarize the contributions of this study as follows:

- This study introduces a novel dual-MAE architecture, which consists of online and target networks with encoder and decoder modules. The architecture aims to enhance feature learning, image reconstruction, and overall network optimization.
- To address the limited availability of pediatric chest X-ray images, this study proposes a two-step training strategy: initially pretraining the network with a large number of adult chest X-ray images, followed by fine-tuning on pediatric chest X-ray images.
    - Through experiments, this study shows that the online network is effective at extracting global features. Furthermore, by increasing the proportion of labeled data used for fine-tuning, it demonstrates that the target network can efficiently learn features even with a limited amount of labeled data.
- The experimental results demonstrate the superior performance of dual-MAE in multi-label classification of pediatric thoracic diseases. The dual-MAE achieves the highest mean AUC values compared to other methods, emphasizing its effectiveness in pediatric thoracic disease classification.

II. RELATED WORKS

A. CLASSIFICATION OF MULTI-LABELED THORACIC DISEASES FROM CHEST X-RAYS

Identifying thoracic diseases in chest X-rays is challenging task due to the intricate and overlapping visual patterns associated with various conditions. This becomes even more difficult because of the multi-label characteristic, where a single X-ray may exhibit multiple diseases simultaneously. To address these difficulties, deep learning techniques are employed for the classification of thoracic diseases from chest X-rays [8], [9], [10], [11]. Wang et al. demonstrated the detection and localization of common thoracic diseases using a unified weakly-supervised multi-label image classification and disease localization framework on the ChestX-ray8 dataset. They achieved higher accuracy in detecting larger abnormalities compared to smaller ones, with an average AUC of 80.30% [8]. Rajpurkar et al. introduced CheXNet, a modified DenseNet model with 121 convolutional layers, designed for the detection of 14 chest abnormalities. CheXNet, trained and evaluated on the ChestX-ray14 dataset, exhibited impressive performance using binary relevance classification for the 14 diseases in the dataset. It outperformed radiologists' performance with an average AUC of 84.11% and an F1-score of 43.50% on a test set comprising 420 images [9]. Bhusal and Panday introduced a multi-label disease diagnosis model for chest X-rays using DenseNet, incorporating model interpretability with Grad-CAM. The model achieved the highest AUC of 0.896 for Cardiomegaly and an accuracy of 0.826, while the lowest AUC was for Nodule at 0.655 with an accuracy of 0.660. Heatmaps and confidence intervals were used for model interpretability and uncertainty estimation, demonstrating high performance in multi-label disease diagnosis tasks [10].

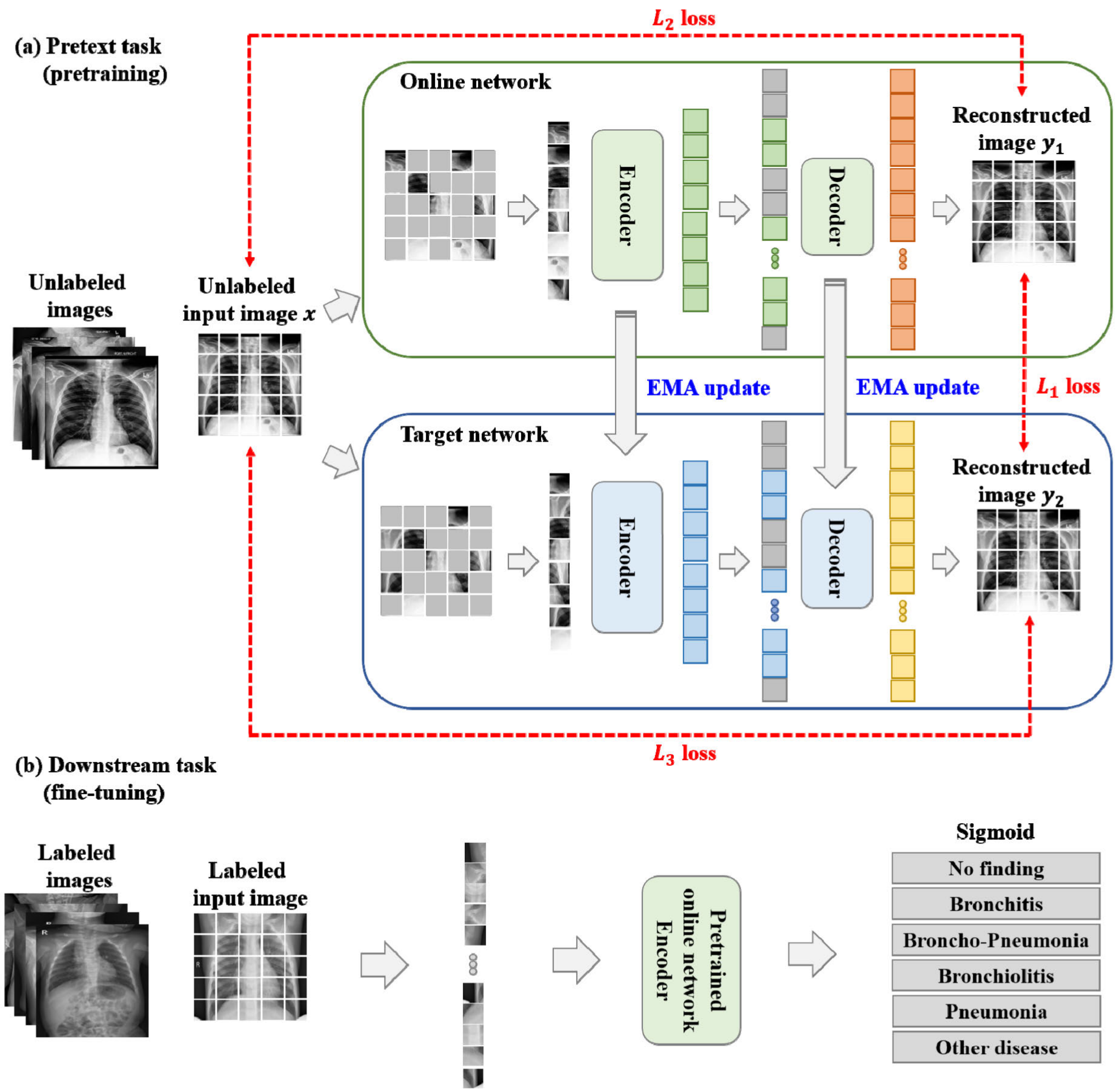

FIGURE 1. Architecture of dual-MAE used in the pretraining and fine-tuning stages: (a) Pretext task stage for backbone model pretraining with large amount of unlabeled data (b) Downstream task stage for fine-tuning the pretrained backbone model with labeled data. (EMA stands for exponential moving average).

Hasanah et al. introduced a fusion architecture, combining CheXNet and Feature Pyramid Network (FPN). This architecture employs a pyramid of feature maps with varying spatial resolutions to capture both low-level and high-level semantic information, enhancing the model's capability to detect diverse features. The evaluation on the ChestX-ray14 dataset demonstrated superior performance, with an average AUC of 0.846 and an accuracy of 0.914. Furthermore, the proposed model exhibited faster diagnostic inference (0.013s) compared to recent approaches [11].

B. MASKED AUTOENCODERS IN MEDICAL IMAGE ANALYSIS

To address the challenges of employing deep learning algorithms with a limited set of labeled medical images, transfer learning is commonly utilized. However, this approach may introduce domain discrepancy issues between medical and natural images since most of the transferring models are pretrained using natural images, such as the ImageNet dataset [12]. A promising alternative solution is the application of masked autoencoders (MAE) as a self-supervised pretraining technique, enhancing neural network representation learning without the need for labeled data [13]. Zhou et al. introduced a self-pretraining paradigm using MAE on the target data to overcome the absence of an ImageNet-scale medical image dataset. Their study demonstrated substantial improvements in diverse medical image tasks, including chest X-ray disease classification, abdominal CT multi-organ segmentation, and MRI brain tumor segmentation [14]. Xing et al. employed MAE for COVID-19 diagnosis from chest X-ray images, comparing it with two other ViT models. The first ViT model was trained from scratch using COVID-19 data, while the second ViT model was pretrained on the ImageNet dataset and then fine-tuned with COVID-19 data. The MAE model demonstrated superior performance with an accuracy of 0.985 and an AUC of 0.996. The study emphasizes the efficiency of MAE on labeled data for fine-tuning, achieving comparable results with only 30% of the labeled training dataset, highlighting its potential for enhanced disease diagnosis in scenarios with limited imaging information [15]. Yoon and Kang applied the MAE to improve pneumonia diagnosis in the context of limited labeled pediatric chest X-rays. They pretrained the MAE model on adult chest X-ray images and then fine-tuned it on a pediatric pneumonia dataset, demonstrating competitive diagnostic performance with an AUC of 0.996 and 95.89% accuracy in distinguishing normal and pneumonia. Furthermore, the approach attained high AUC values (normal: 0.997, bacterial pneumonia: 0.983, viral pneumonia: 0.956) and 93.86% accuracy in classifying normal, bacterial pneumonia, and viral pneumonia [16].

III. MODEL

A. ARCHITECTURE

In this study, we introduce a novel architecture called dual-MAE, comprising an online network and a target network, each with encoder and decoder modules. As outlined in Fig. 1 (a), the input image is divided into non-overlapping 16 × 16 patches and each patch is then transformed into a token through linear projection with an additional positional embedding. Subsequently, a subset of tokens is randomly sampled based on a masking ratio ranging from 65% to 95%, and this selected subset is then masked. The visible and unmasked tokens employed as input for the online network are denoted as x_1, while those for the target network are denoted as x_2. Dual-MAE achieves efficient pretraining with reduced computational demands, as its encoders process only visible and unmasked tokens, similar to vanilla MAE. These encoders are designed to extract a global representation from partial observations, with their output tokens facilitating the reconstruction of learnable masked tokens in the decoders. The decoders process the full set of tokens by combining encoded visible tokens with learnable mask tokens, utilizing positional embeddings in all input tokens. As a result, the decoders reconstruct patches at specific masked positions, reshaping the output into a reconstructed image. The online network decoder reconstructs a full image, denoted as y_1, using the randomly selected unmasked tokens x_1 and learnable mask tokens, while the target network decoder simultaneously reconstructs another full image, denoted y_2, using the randomly selected unmasked tokens x_2 and another learnable mask tokens. The decoders, intentionally smaller than the encoder, are exclusively employed in pretraining to improve efficiency. In medical imaging tasks, we emphasize the vital role of contextual information in reconstructing masked image patches, considering the intrinsic dependence and connection between the region of interest (ROI) and its physiological environment and surroundings.

The loss function, denoted as L_1, quantifies the mean squared errors between the reconstructed images of online network y_1 and those of target network y_2 in pixel space. Additionally, in alignment with the vanilla MAE approach, the mean squared errors between the original images (x) and reconstructed images of online / target networks (y_1 and y_2) in pixel spaces are computed, respectively. Specifically, the L_2 loss term quantifies the mean squared errors between x and y_1 while L_3 loss term measures the mean squared errors between x and y_2 in pixel space. The total loss, expressed as L_{total}, serves as a key optimization metric during training and is computed at each training step using following equation:

$$L_{total} = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3. \tag{1}$$

In this study, the values of λ_1, λ_2 and λ_3 are set to 1. The proposed dual-MAE algorithm includes a stochastic optimization step during each training iteration, aiming to minimize the total loss, L_{total}. The online network is defined by a specific set of weights denoted as θ, while the target network employs a distinct set of weights denoted as ξ. Notably, the target network actively contributes to training the online network by providing reconstructed images, y_2, and its parameters ξ are updated using an exponential moving average (EMA) of the online parameters θ. This update, governed by a target decay rate $\tau \in [0, 1]$, ensures continuous improvement in the performance of the target network by involving a weighted combination of the current target parameters and the historical online parameters, as expressed in the following equation [17]:

$$\xi \leftarrow \tau\xi + (1 - \tau)\,\theta. \tag{2}$$

The target decay rate τ is set to 0.99 in this study. This sophisticated dual-MAE approach is designed to improve feature learning, image reconstruction, and network optimization, providing a comprehensive framework for effective image representation and reconstruction tasks.

B. PRETRAINING DUAL-MAE MODEL AND FINE-TUNING VIT MODEL

The objective of this study is to categorize multi-labeled pediatric thoracic diseases. However, due to the limited number of pediatric chest X-rays, it is impractical to pretrain a dual-MAE network using only pediatric chest X-rays. To overcome this limitation, we aim to pretrain the dual-MAE network using adult chest X-rays, which share substantial similarities with pediatric chest X-rays in structural and textural aspects and exhibit minimal domain discrepancy. The adult chest X-rays in this context are obtained from the CheXpert and ChestX-ray14 datasets [7], [8].

The adult chest X-rays were resized to 256 × 256 pixels and standardized using mean and standard deviation of ImageNet dataset. We then performed random resizing cropping (scale range: 0.5∼1.0) to 224 × 224 pixels and applied horizontal flipping for dataset augmentation. However, to avoid

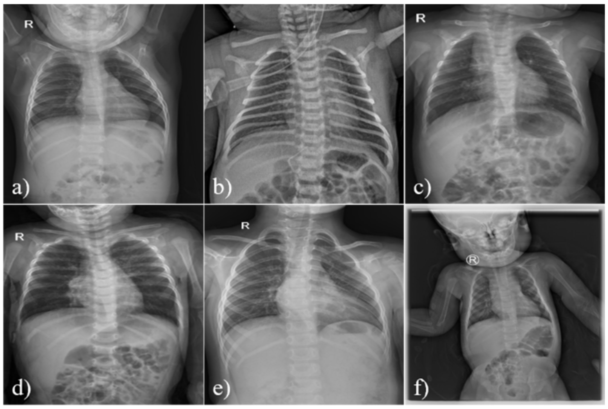

FIGURE 2. Multi-labeled pediatric chest X-ray sample images. (Each image can belong to multiple disease categories): a) No finding category, b) Bronchitis & Pneumonia categories, c) Broncho-Pneumonia & Bronchiolitis categories, d) Bronchiolitis & Pneumonia categories, e) Pneumonia & Other disease categories, f) Other disease category.

TABLE 1. Details on datasets used in this study.

	Dataset Name	Data Category	Number of training data	Number of test data
Pretraining data	ChestX-ray14	-	112,120	-
	CheXpert	-	191,229	-
Fine-tuning data	PediCXR	No finding	5,143	907
		Bronchitis	842	174
		Broncho-pneumonia	545	84
		Bronchiolitis	497	90
		Pneumonia	392	89
		Other diseases	485	85

TABLE 2. Mean AUC comparison of MAE(X-ray) and dual-MAE(X-ray) across various masking ratio (Highest values are in bold).

	65%	70%	75%	80%	85%	90%	95%
MAE (X-ray)	0.698	0.681	0.733	0.731	**0.735**	0.734	0.726
dual-MAE (X-ray)	0.699	0.747	**0.752**	0.741	0.746	0.740	0.697

potential risks such as cropping or introducing bias to informative lesions or organs, we refrained from using additional augmentation methods. The dual-MAE was optimized using the AdamW optimizer with parameters ($\beta_1 = 0.9$, $\beta_2 = 0.95$) and a weight decay of 0.05. The ViT transformer blocks in the dual-MAE were initialized using Xavier uniform initialization. The initial learning rate and batch size were set to 1.5e-4 and 256, respectively. The learning rate was warmed up for the initial 20 epochs and adjusted using a cosine annealing schedule. The pretraining process of dual-MAE ran for 800 epochs.

Fine-tuning was performed using an end-to-end approach on pediatric chest X-ray images as shown in Fig. 1 (b). During fine-tuning, we utilized the pretrained ViT encoder from the online network of dual-MAE. Following [18], a linear classifier is added after the class token output from the ViT encoder. In this study, the task of predicting pediatric thoracic diseases from chest X-rays was treated as a multi-label classification problem, where each input example may be associated with multiple disease labels as shown in Fig. 2. To address this multi-label classification problem, we replaced the final fully connected layer in ViT network with a fully connected layer that generates a 6-dimensional output. Subsequently, we applied an elementwise sigmoid nonlinearity. The resulting output represents the predicted probability of each thoracic disease. Next, we modified the loss function to optimize the mean of binary cross-entropy (BCE) losses and fine-tuned the ViT model with the BCE loss [9], [19]. The ViT model was optimized using the AdamW optimizer, incorporating parameters ($\beta_1 = 0.9$, $\beta_2 = 0.95$) and a weight decay of 0.05. The initial learning rate and batch size were set to 2.5e-3 and 128, respectively, following a cosine annealing schedule. Adhering to the recommendations in [20], layer-wise LR decay of 0.55, RandAug magnitude of 6, and a DropPath rate of 0.2 were implemented. The fine-tuning process ran for 75 epochs, including a warm-up period of 5 epochs. Each set of experiments was repeated three times, applying different random seeds for weight initialization.

After fine-tuning, the fine-tuned model was used to evaluate test images.

IV. EXPERIMENTS

A. PRETRAINING AND FINE-TUNING DATASETS

We utilized two datasets, CheXpert and ChestX-ray14, to pretrain the proposed dual-MAE model. The CheXpert dataset contains 224,316 chest X-rays with both frontal and lateral views. It is annotated for 14 observations (12 pathologies, support devices, and observations with no findings). We employed only frontal view images from the CheXpert dataset, amounting to 191,229 images [7]. The ChestX-ray14 dataset consists of 112,120 frontal view chest X-ray images, with 51,708 exhibiting one or more pathologies across 14 classes, and the remaining 60,412 images indicating no signs of disease [8]. Although both datasets have labels, we do not use the label information during pretraining. The total number of images for dual-MAE pretraining is the sum of both datasets, reaching 303,349 images. The pediatric chest X-ray images for this study was obtained from the PediCXR dataset [21]. This dataset comprises 9,125 pediatric chest X-ray images, officially categorized into training

TABLE 3. AUC comparison of different methods across different diseases. (Highest values are in bold and values in parenthesis stands for standard deviation).

Methods	**No Finding**	**Bronchitis**	**Broncho-Pneumonia**	**Bronchiolitis**	**Pneumonia**	**Other Disease**	**mean AUC**
ResNet-34 (random)	0.681 (0.005)	0.647 (0.005)	0.732 (0.011)	0.669 (0.010)	0.692 (0.005)	0.596 (0.005)	0.669 0.003)
ViT-S (random)	0.650 (0.006)	0.641 (0.003)	0.702 (0.01)	0.663 (0.014)	0.639 (0.01)	0.576 (0.003)	0.645 (0.005)
ResNet-34 (IN)	0.702 (0.045)	0.661 (0.042)	0.770 (0.034)	0.686 (0.035)	0.742 (0.03)	0.624 (0.026)	0.697 (0.032)
ViT-S (IN)	0.646 (0.002)	0.642 (0.003)	0.680 (0.017)	0.663 (0.011)	0.632 (0.022)	0.568 (0.005)	0.638 (0.006)
ResNet-34 (X-ray)	0.722 (0.014)	0.674 (0.023)	0.781 (0.029)	0.703 (0.022)	0.760 (0.031)	0.570 (0.112)	0.712 (0.023)
ViT-S (X-ray)	0.673 (0.005)	0.660 (0.011)	0.718 (0.001)	0.679 (0.007)	0.662 (0.002)	0.588 (0.007)	0.669 (0.005)
MAE (X-ray)	0.747 (0.010)	0.710 (0.003)	0.805 (0.005)	0.692 (0.012)	0.813 (0.017)	0.643 (0.017)	0.735 (0.010)
dual-MAE (X-ray)	**0.765** **(0.001)**	**0.712** **(0.003)**	**0.828** **(0.000)**	**0.709** **(0.007)**	**0.833** **(0.008)**	**0.665** **(0.006)**	**0.752** **(0.003)**

and test datasets. The training set contains 7,728 images, while the test set consists of the remaining 1,397 images. In the training set, every chest X-ray image is annotated for 15 diseases, while the official test set comprises 11 diseases. To ensure a fair and robust assessment, rare diseases (with positive samples fewer than 5 in test set) were aggregated into a category labeled "other diseases" [22]. As a result, the PediCXR dataset includes 6 classes: no finding, bronchitis, broncho-pneumonia, bronchiolitis, pneumonia, and other disease as shown in Table 1 and Fig. 2. Table 1 provides detailed information on pretraining and fine-tuning data. Eighty percent of the PediCXR training data was used for fine-tuning the pretrained networks or training networks from scratch for comparison. The remaining twenty percent of the PediCXR training data was served as validation data.

The data used in this study is publicly available. The ChestX-ray14 dataset can be accessed from [23], the CheXpert dataset from [24], and the PediCXR data from [25]. Ethical review and approval were waived for this study as it involved the analysis of anonymous clinical open data.

B. COMPARATIVE METHODS

In this study, we conducted a comparative analysis to demonstrate the effectiveness of pretraining the dual-MAE model on adult chest X-ray images for the diagnosis of pediatric thoracic diseases. The vanilla MAE and dual-MAE models, pretrained on adult chest X-rays, will be denoted as MAE(X-ray) and dual-MAE(X-ray), respectively. The first comparative approach involves training pediatric thoracic disease data from scratch, utilizing ResNet-34 and ViT-S models as backbone models, denoted as ResNet-34(random) and ViT-S(random), respectively [18], [26]. ViT-S was chosen as the backbone model due to the architecture of dual-MAE being based on ViT-S model. The reason for choosing ResNet-34 is that the ResNet model is a representative algorithm in convolutional neural networks (CNNs), and the selection of the ResNet-34 model is based on its parameter size, which is similar to ViT-S, both having 22 million parameters. The second comparative approach involves ResNet-34(IN) and ViT-S(IN), pretrained on the ImageNet dataset, which are then fine-tuned using pediatric thoracic disease data. The third comparative approach involves using ResNet-34(X-ray) and ViT-S(X-ray) models, pretrained on adult chest X-rays, which are then fine-tuned with pediatric chest X-rays. The forth comparative approach entails fine-tuning MAE(X-ray) with pediatric thoracic disease data. These methods are compared with the proposed dual-MAE(X-ray) in this study.

C. EXPERIMENTAL RESULTS

The performance of MAE(X-ray) and dual-MAE(X-ray) models varies with the masking ratio (the proportion of masked patches during pretraining). Optimizing the masking ratio involves addressing information redundancy in the data. Notably, BERT used a 15% masking ratio for language tasks, while vanilla MAE opted for a 75% masking ratio in natural image-related tasks [13], [27]. Due to the substantial similarity in chest anatomy, chest X-rays inherently exhibit higher information redundancy than natural images. Consequently, vanilla MAE has favored a 90% masking ratio specifically for chest X-rays [19]. In this study, experiments were conducted by incrementally increasing the masking ratio from 65% to 95% in 5% increments to determine the optimal masking ratio for MAE(X-ray) and dual-MAE(X-ray) models.

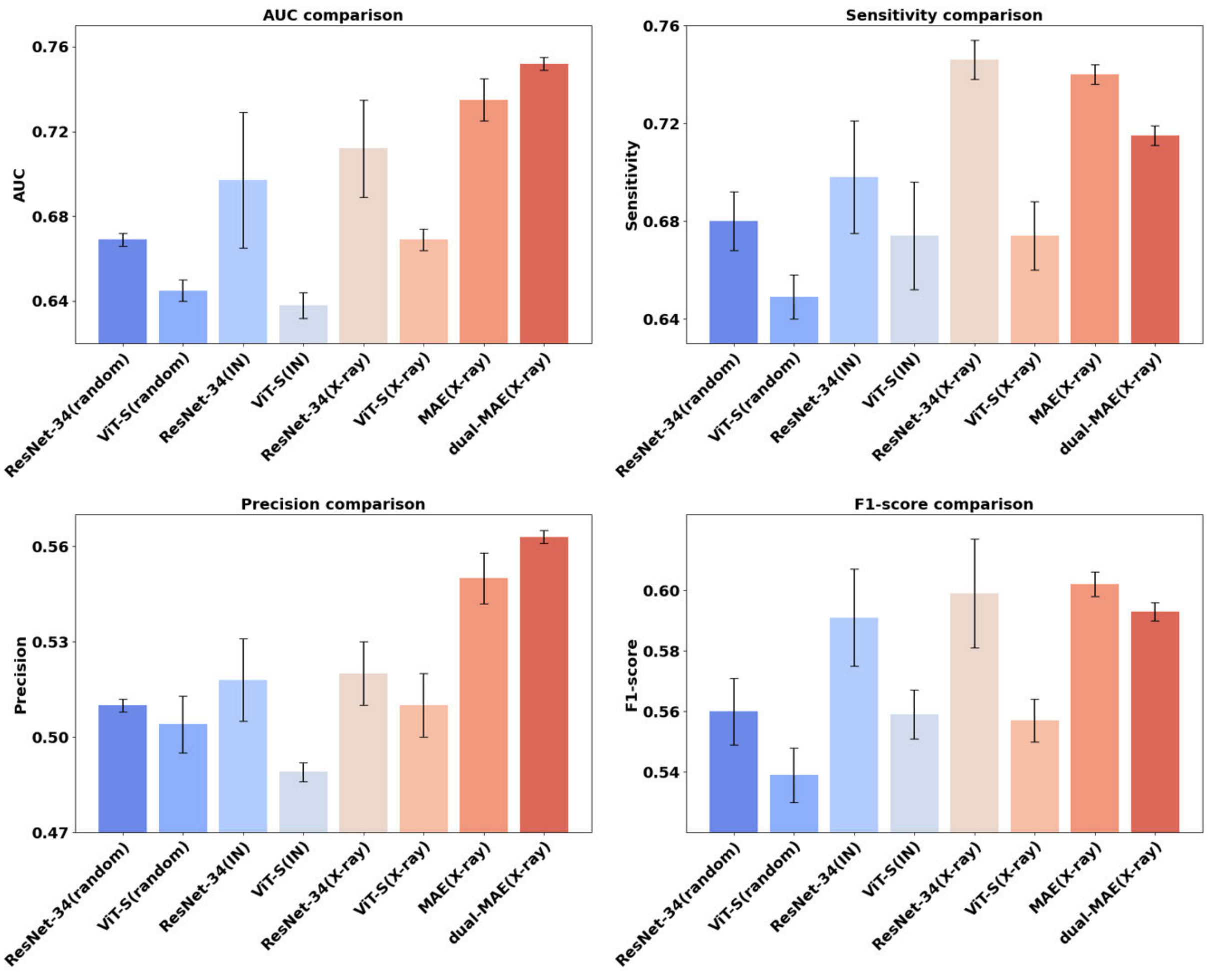

FIGURE 3. Performance metrics comparison of different methods for multi-labeled pediatric thoracic disease classification.

Table 2 presents the mean AUC values of the MAE(X-ray) and dual-MAE(X-ray) models across various masking ratios. These AUC values reflect the models' ability to classify pediatric thoracic diseases from X-rays, with higher scores indicating better performance. Remarkably, the dual-MAE(X-ray) consistently demonstrates competitive performance compared to MAE(X-ray) across different masking ratios, emphasizing its robustness in multi-label classification tasks. Specifically, for MAE(X-ray), the mean AUC peaked at 0.735 with a masking ratio of 85%, while dual-MAE(X-ray) achieved its highest macro AUC of 0.752 at masking ratios of 75%. In subsequent experiments, the masking ratio was fixed at 85% for MAE(X-ray) and 75% for dual-MAE(X-ray). Compared to other researches in [13] and [20], our study similarly demonstrated an optimal masking ratio of 75% for dual-MAE(X-ray) and 85% for MAE(X-ray). Those masking ratios imply a significant reduction in the computational and memory demands of the model, indicating the potential for efficient pretraining.

Table 3 presents the AUC values for various pediatric thoracic disease categories across different deep learning methods. The dual-MAE(X-ray) consistently achieves high AUC scores across all categories, emphasizing its robust disease classification performance. Specifically, the dual-MAE(X-ray) model demonstrates superior performance in classifying broncho-pneumonia class and pneumonia class, yielding AUC values of 0.828 and 0.833, respectively. However, for other diseases class, although the dual-MAE(X-ray) model outperforms other methods, its AUC value of 0.665 is relatively lower compared to the rest of the disease categories. Other diseases class comprises images of ten different minority diseases grouped together, indicating challenges in their classification.

Fig. 3 and Table 4 illustrate the performance metrics such as AUC, sensitivity, precision, and F1-score of different methods for classifying multi-labeled pediatric thoracic diseases. Both the dual-MAE(X-ray) and MAE(X-ray) models exhibited superior performance across all performance metrics compared to other models. Specifically, dual-MAE(X-ray) and MAE(X-ray) models exhibited higher AUC values (0.735 and 0.752, respectively) than ResNet-34 and ViT-S models, demonstrating the efficacy of the MAE-based pretraining approach for classifying pediatric thoracic diseases from X-ray images. In terms of sensitivity, the

TABLE 4. Performance metrics comparison of different methods for multi-labeled pediatric thoracic disease classification. (Highest values are in bold and values in parenthesis represent standard deviation).

Methods	AUC	Sensitivity	precision	F1-score
ResNet-34 (random)	0.669 (0.003)	0.680 (0.012)	0.510 (0.002)	0.560 (0.011)
ViT-S (random)	0.645 (0.005)	0.649 (0.009)	0.504 (0.009)	0.539 (0.009)
ResNet-34 (IN)	0.697 (0.032)	0.698 (0.023)	0.518 (0.013)	0.591 (0.016)
ViT-S (IN)	0.638 (0.006)	0.674 (0.022)	0.489 (0.003)	0.559 (0.008)
ResNet-34 (X-ray)	0.712 (0.023)	**0.746 (0.008)**	0.520 (0.010)	0.599 (0.018)
ViT-S (X-ray)	0.669 (0.005)	0.674 (0.014)	0.510 (0.010)	0.557 (0.007)
MAE (X-ray)	0.735 (0.010)	0.740 (0.004)	0.550 (0.008)	**0.602 (0.004)**
dual-MAE (X-ray)	**0.752 (0.003)**	0.715 (0.004)	**0.563 (0.002)**	0.593 (0.003)

ResNet-34(X-ray) model achieved the highest value (0.746), followed by the MAE(X-ray) model (0.740), indicating their effectiveness in identifying positive cases. Moreover, focusing on precision, the dual-MAE(X-ray) model outperformed other models with a value of 0.563, followed by the MAE(X-ray) model with 0.550, highlighting their ability to minimize false positives. The F1-score emphasizes the ability to balance false positives and false negatives. However, the F1-score of dual-MAE(X-ray) at 0.593 is slightly lower than that of the MAE(X-ray) and ResNet-34(X-ray) models, which stand at 0.602 and 0.599, respectively. Overall, the dual-MAE(X-ray) model exhibited the highest performance in terms of AUC and precision. Meanwhile, the MAE(X-ray) model demonstrated the best performance in F1-score, and the ResNet-34(X-ray) model showcased the highest sensitivity. One notable observation is that, except for the ViT-S(X-ray) model, the models pretrained on adult chest X-rays and fine-tuned on pediatric chest X-rays outperformed the performance of other models. From a transfer learning perspective, this suggests that the similarity in domain between the data used for pretraining and fine-tuning makes transfer learning more effective. On the other hand, fine-tuning with models pretrained on ImageNet and pediatric chest X-rays faces challenges due to the domain discrepancy between the two datasets, indicating potential issues with effective fine-tuning. For the ViT-S(X-ray) model, it seems to exhibit an inductive bias issue because the adult chest X-rays used for pretraining are relatively limited compared to the ImageNet data.

Table 5 represents the performance of the dual-MAE(X-ray) model based on the percentage of labeled data used for fine-tuning. As the percentage of labeled data increases, there is a noticeable improvement in AUC values. For instance, with only 1% of labeled data, the AUC is 0.573, indicating relatively lower performance. However, as the percentage of labeled data increases to 10%, 50%, and finally 100%, the AUC values improve to 0.620, 0.733, and 0.752, respectively. Similarly, precision exhibits an upward trend with increasing labeled data percentages, reaching performance levels of 0.475, 0.479, 0.554, and 0.563. However, sensitivity and F1-score show an ascending trend with labeled data proportions until they peak at 50%, after which they slightly decrease at 100%. An interesting observation is the similarity in performance between the dual-MAE(X-ray) model fine-tuned with 50% labeled data and the MAE(X-ray) model fine-tuned with 100% labeled data. For instance, the dual-MAE(X-ray) model achieves an AUC of 0.733, sensitivity of 0.730, precision of 0.554, and F1-score of 0.599, while the MAE(X-ray) model exhibits an AUC of 0.735, sensitivity of 0.740, precision of 0.550, and F1-score of 0.601.

TABLE 5. Performance of dual-MAE(X-ray) model based on labeled data percentage for fine-tuning. (Highest values are in bold and values in parenthesis represent standard deviation).

Percentage of labeled data	AUC	Sensitivity	precision	F1-score
1%	0.573 (0.016)	0.680 (0.012)	0.475 (0.006)	0.540 (0.006)
10%	0.620 (0.038)	0.696 (0.053)	0.479 (0.020)	0.543 (0.014)
50%	0.733 (0.008)	**0.730 (0.017)**	0.554 (0.009)	**0.599 (0.010)**
100%	**0.752 (0.003)**	0.715 (0.004)	**0.563 (0.002)**	0.593 (0.003)

Table 6 compares the AUC of each model at different ratios of labeled training data. As the proportion of labeled data increases from 10% to 100%, there is a consistent improvement in performance metrics. When the training data is limited to 10%, all methods except for ViT-S(IN), MAE(X-ray), and dual-MAE(X-ray) achieve an AUC below 0.6. At 50% of the training data, dual-MAE(X-ray) significantly outperforms other methods with an AUC of 0.733, demonstrating its effectiveness, followed by MAE at 0.686. With 100% of the training data, dual-MAE achieves the highest AUC of 0.752, while MAE achieves 0.732. Overall, the proposed dual-MAE consistently outperforms other methods across all data proportions, particularly excelling with limited data.

From these results, the roles of the online network and the target network in dual-MAE become apparent. MAE consists solely of the online network, whereas dual-MAE comprises both the online network and the target network. Thus, the performance difference between the two can be attributed to the presence of the target network. As shown in Table 6, when fine-tuning with only 10% of the labeled data, MAE achieves an AUC of 0.636, slightly higher than dual-MAE's AUC of 0.620. However, with 50% of the labeled data, the AIC pf

TABLE 6. AUC comparison of methods at different labeled training data Ratios. (Highest values are in bold and values in parenthesis represent standard deviation).

Methods	10%	50%	100%
ResNet-34 (random)	0.551 (0.093)	0.639 (0.040)	0.667 (0.016)
ViT-S (random)	0.560 (0.017)	0.579 (0.027)	0.644 (0.011)
ResNet-34 (IN)	0.589 (0.059)	0.639 (0.053)	0.694 (0.061)
ViT-S (IN)	0.623 (0.023)	0.661 (0.004)	0.655 (0.012)
ResNet-34 (X-ray)	0.569 (0.086)	0.670 (0.035)	0.694 (0.006)
ViT-S (X-ray)	0.576 (0.011)	0.611 (0.036)	0.654 (0.000)
MAE (X-ray)	**0.636 (0.047)**	0.686 (0.048)	0.732 (0.009)
dual-MAE (X-ray)	0.620 (0.038)	**0.733 (0.008)**	**0.752 (0.003)**

MAE increases from 0.636 to 0.686, a 0.050 improvement, whereas the AUC of dual-MAE rises from 0.620 to 0.733, an impressive 0.113 increase. These results suggest that the target network in dual-MAE significantly enhances feature learning, particularly when only limited labeled data is available. This underscores the effectiveness of utilizing dual-MAE for feature extraction and fine-tuning in scenarios with limited labeled data availability. MAE, composed solely of an online network, demonstrates robust performance since it is designed to extract global representations from partial observations.

V. CONCLUSION

In this study, we introduced the dual-MAE, a novel architecture designed for the effective classification of multi-labeled pediatric thoracic diseases. The dual-MAE consists of both an online network and a target network, each incorporating encoder and decoder modules. The model was optimized by minimizing mean squared errors for reconstructed images. Due to the limited availability of pediatric chest X-ray data, we employed a two-step approach: initially pretraining the dual-MAE network using adult chest X-ray images, followed by fine-tuning on pediatric chest X-ray data.

The experimental results demonstrated that dual-MAE (X-ray) achieved the highest mean AUC of 0.752 when the masking ratio was set to 75%, outperforming other methods. Additionally, this study emphasized the consistent high AUC scores of dual-MAE(X-ray) across various thoracic disease categories. Further analysis illustrated that the dual-MAE(X-ray) outperformed the MAE(X-ray) in terms of AUC across various masking ratios, demonstrating its effectiveness in improving the multi-label classification performance for pediatric thoracic diseases. The dual-MAE(X-ray) and MAE(X-ray) models outperformed the ResNet-34 and ViT-S models across various performance metrics. Notably, ResNet-34(IN) and ViT-S(IN) benefited from pretraining on a larger ImageNet dataset of 1.28 million images, compared to the 303,349 adult chest X-ray images used for pretraining MAE(X-ray) and dual-MAE(X-ray), indicating a 4.2-fold increase in image quantity. Nevertheless, dual-MAE(X-ray) and MAE(X-ray) demonstrated superior performance. When comparing the performance between dual-MAE(X-ray) and MAE(X-ray), the proposed dual-MAE(X-ray) exhibited better performance in terms of AUC and precision but not in sensitivity and F1-score. However, interestingly, overall performance when fine-tuning dual-MAE(X-ray) with a 50% labeled data ratio was comparable to that of MAE(X-ray) with 100% labeled data.

By varying the proportion of labeled training data to 10%, 50%, and 100%, we explained the roles of the online network and target network in dual-MAE. The online network, which follows the vanilla MAE architecture, extracts global representations from partial observations. The target network, on the other hand, aids the model in efficiently learning features even in scenarios with limited labeled data availability.

Although dual-MAE(X-ray) exhibited better performance compared to ResNet-34, ViT-S, and MAE(X-ray) models, there are limitations to consider. Firstly, dual-MAE(X-ray) represents one approach of self-supervised learning, and its superiority needs to be validated by comparison with other self-supervised learning methods such as SimCLR, MOCO, and BYOL [17], [28], [29]. Secondly, it is crucial to establish the generality of the proposed method through experiments with diverse datasets. Future research will aim to address these limitations.

REFERENCES

[1] A. Hart and E. Y. Lee, "Pediatric chest disorders: Practical imaging approach to diagnosis," in *Diseases of the Chest, Breast, Heart and Vessels 2019–2022: Diagnostic and Interventional Imaging*. Cham, Switzerland: Springer, 2019, ch. 10, pp. 107–125. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK553873/

[2] K. Grimwood and A. B. Chang, "Long-term effects of pneumonia in young children," *Pneumonia*, vol. 6, pp. 101–114, Oct. 2015.

[3] K.-C. Chen, H.-R. Yu, W.-S. Chen, W.-C. Lin, Y.-C. Lee, H.-H. Chen, J.-H. Jiang, T.-Y. Su, C.-K. Tsai, T.-A. Tsai, C.-M. Tsai, and H. H.-S. Lu, "Diagnosis of common pulmonary diseases in children by X-ray images and deep learning," *Sci. Rep.*, vol. 10, no. 1, p. 17374, Oct. 2020.

[4] S. Padash, M. R. Mohebbian, S. J. Adams, R. D. E. Henderson, and P. Babyn, "Pediatric chest radiograph interpretation: How far has artificial intelligence come? A systematic literature review," *Pediatric Radiol.*, vol. 52, no. 8, pp. 1568–1580, Apr. 2022.

[5] T. Yoon and D. Kang, "Bimodal CNN for cardiovascular disease classification by co-training ECG grayscale images and scalograms," *Sci. Rep.*, vol. 13, no. 1, p. 2937, Feb. 2023.

[6] M. I. Hossain, M. Zunaed, M. K. Ahmed, S. M. J. Hossain, A. Hasan, and T. Hasan, "ThoraX-PriorNet: A novel attention-based architecture using anatomical prior probability maps for thoracic disease classification," *IEEE Access*, vol. 12, pp. 3256–3273, 2024.

[7] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 590–597.

[8] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, ‘‘ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[9] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, ‘‘CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,’’ 2017, *arXiv:1711.05225*.

[10] D. Bhusal and S. Prasad Panday, ‘‘Multi-label classification of thoracic diseases using dense convolutional network on chest radiographs,’’ 2022, *arXiv:2202.03583*.

[11] U. Hasanah, C. Avian, J. T. Darmawan, N. Bachroin, M. Faisal, S. W. Prakosa, J. S. Leu, and C. T. Tsai, ‘‘CheXNet and feature pyramid network: A fusion deep learning architecture for multilabel chest X-ray clinical diagnoses classification,’’ *Int. J. Cardiovasc. Imag.*, vol. 40, pp. 709–722, Dec. 2023.

[12] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, ‘‘Transfusion: Understanding transfer learning for medical imaging,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, ‘‘Masked autoencoders are scalable vision learners,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[14] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, ‘‘Self pre-training with masked autoencoders for medical image classification and segmentation,’’ 2022, *arXiv:2203.05573*.

[15] X. Xing, G. Liang, C. Wang, N. Jacobs, and A.-L. Lin, ‘‘Self-supervised learning application on COVID-19 chest X-ray image classification using masked AutoEncoder,’’ *Bioengineering*, vol. 10, no. 8, p. 901, Jul. 2023.

[16] T. Yoon and D. Kang, ‘‘Enhancing pediatric pneumonia diagnosis through masked autoencoders,’’ *Sci. Rep.*, vol. 14, no. 1, p. 6150, Mar. 2024.

[17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, and B. Piot, ‘‘Bootstrap your own latent—A new approach to self-supervised learning,’’ in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ‘‘An image is worth 16×16 words: Transformers for image recognition at scale,’’ 2020, *arXiv:2010.11929*.

[19] M.-L. Zhang and Z.-H. Zhou, ‘‘A review on multi-label learning algorithms,’’ *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[20] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, ‘‘Delving into masked autoencoders for multi-label thorax disease classification,’’ in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3577–3589.

[21] H. H. Pham, N. H. Nguyen, T. T. Tran, T. N. M. Nguyen, and H. Q. Nguyen, ‘‘PediCXR: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children,’’ *Sci. Data*, vol. 10, no. 1, p. 240, Apr. 2023.

[22] C. Wu, X. Zhang, Y. Wang, Y. Zhang, and W. Xie, ‘‘K-diag: Knowledge-enhanced disease diagnosis in radiographic imaging,’’ 2023, *arXiv:2302.11557*.

[23] *ChestX-Ray14 Dataset Repository*. Accessed: Jun. 25, 2024. [Online]. Available: https://nihcc.app.box.com/v/ChestXray-NIHCC

[24] *CheXpert Dataset Repository*. Accessed: Jun. 25, 2024. [Online]. Available: https://stanfordmlgroup.github.io/competitions/chexpert/

[25] *PediCXR Dataset Repository*. Accessed: Jun. 25, 2024. [Online]. Available: https://physionet.org/content/vindr-pcxr/1.0.0/

[26] K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Deep residual learning for image recognition,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, ‘‘BERT: Pre-training of deep bidirectional transformers for language understanding,’’ in *Proc. NAACL-HLT*, vol. 1, 2018, p. 2.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. A. Hinton, ‘‘Simple framework for contrastive learning of visual representations,’’ in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.

[29] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, ‘‘Momentum contrast for unsupervised visual representation learning,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

TAEYOUNG YOON received the B.E. degree in healthcare information technology from Inje University, South Korea, in 2023, where he is currently pursuing the master’s degree. His current research interests include deep learning and medical image processing.

DAESUNG KANG received the Ph.D. degree in biomedical engineering from the University of Florida, USA, in 2016. He is currently an Assistant Professor with the Medical Information Technology Department, Inje University, South Korea. His current research interests include self-supervised learning and semi-supervised learning for medical data.