공학석사 학위논문

어텐션 메커니즘을 이용한 강인한 표현 학습 기반 추적 흉부엑스선 변화 평가 인공지능 시스템 개발에 대한 연구

A Study on the Development of AI System to Assess Interval Changes on Follow-up Chest Radiographs Using Attention Mechanism for Robust Representation Learning

울산대학교대학원

의과학과

조경진

어탠션 메커니즘을 이용한 강인한 표현 학습 기반 추적 흉부엑스선 변화 평가 인공지능 시스템 개발에 대한 연구

이 논문을 공학석사 학위논문으로 제출함

2023 년 2 월

울산대학교대학원

의과학과

조경진

조경진의 공학 석사학위 논문을 인준함

심사위원 서준범 (인)

심사위원 김남국 (인)

심사위원 이준구 (인)

울산대학교대학원

2023 년 2 월

Abstract

Diagnosing meaningful changes in follow-up CXRs which is one of the main tasks of radiologists in routine clinical practice is challenging because radiologists must distinguish between pathological changes and natural or benign changes and inevitably it requires a large amount of datasets and high quality annotation. However, it is very difficult to acquire those kinds of high-quality annotation, in actual clinical setting. In this paper, a multi-task Siamese convolutional vision transformer (MuSiC-ViT) with an anatomy matching module (AMM) is proposed to mimic the radiologist's cognitive process in classifying CXR pairs (baseline change/no change) and follow-up radiographs to overcome these issues. MuSiC-ViT uses the CNNs and visual transformers (CMTs) model, which combines the CNN and transformer architecture and consists of three main components: a Siamese architecture, an anatomy matching module, and multi-task learning. Since the input was a pair of CXRs, a Siamese network was chosen for the encoder network. The AMM is an attentional module that focuses on related regions in corresponding CXR pairs. To mimic the cognitive process of a radiologist, MuSiC-ViT was trained with multi-task learning and classified normal/abnormal, change/no change, and matching anatomy. A total of 406K CXRs were examined, with 88K pairs with changes and 115K pairs without changes recorded for the training dataset. For the internal validation dataset, 1,620 pairs were used. To show the robustness of MuSiC-ViT, we checked it with two external validation datasets. MuSiC-ViT achieved accuracy and area under the receiver operating characteristic curve (AUC) of 0.728 and 0.797 for the internal validation dataset, 0.614 and 0.784 for the first external validation dataset, and 0.745 and 0.858 for the second external validation dataset, respectively. In summary, we proposed a MuSiC-ViT that may discriminate between change and no-change CXR pairs by comparing baseline and follow-up CXR. By adding AMM, we proved through an ablation study that AMM helps AUC gain. Furthermore, disease loss to distinguish normal and normal of each baseline and follow-up CXR could help the model classify abnormal and normal CXRs in the case of disease. This architecture could be used to develop further CXR follow-up studies and lead to actual applications in clinical settings.

차 례

표 및 그림 차례

Introduction

Deep learning, a subset of machine learning algorithms based on artificial neural networks, has achieved impressive results on a variety of computer vision tasks. Convolutional neural networks (CNNs), a type of deep learning algorithm, gained widespread attention after winning the ImageNet Large-Scale Visual Recognition Challenge in 2012 by significantly reducing the error rate in image classification. In addition to image classification, CNNs have been successful in tasks such as object recognition, semantic segmentation, image reconstruction, depth estimation, and visual question answering.

There is a similar change detection task in medicine, specifically in the interpretation of chest radiographs (CXRs) during follow-up exams. In clinical practice [1], CXR follow-up exams are used to determine whether patients have experienced significant changes over time. However, change detection in medicine is more complex than in remote sensing, where satellite and aerial images can be analyzed using a change map if the field of view of the two images is well-matched. In contrast, changes in follow-up CXRs compared to baseline CXRs may reflect meaningful disease progression and other factors such as differences in patient posture, respiratory rate, and aging, which are not relevant to the initial changes. In addition, there may be other clinical findings present on the follow-up CXR besides the disease diagnosed on the initial image that need to be recognized. As a result, diagnosing changes in CXR follow-up images is extremely challenging.

CXRs are widely used for screening in the medical field. However, interpreting radiographs requires specialized expertise and can be a time-consuming task, leading to the potential for mistakes [4]. In clinical practice, radiologists often compare follow-up CXRs to baseline CXRs to detect changes that may indicate the presence or progression of disease [2,3]. Deep learning methods have been used to assist radiologists in interpreting radiographs [5,6] and have contributed significantly to tasks such as abnormality classification [7-9], detection [10-13], and segmentation [14-17]. These methods can help reduce the workload of radiologists and improve the accuracy of CXR interpretation.

When interpreting follow-up CXRs, radiologists consider several key components. First, they determine whether the CXRs are normal or abnormal at both the baseline and follow-up stages. Next, they match the radiographs by comparing the suspected pathological and anatomic regions of the baseline and follow-up images. Finally, they diagnose a change or no change between the baseline and follow-up CXRs. These steps involve careful analysis of the features present on the CXRs and require specialized expertise and experience.

To address the challenging task of detecting changes in follow-up chest radiographs (CXRs), we developed a multi-task Siamese convolutional vision transformer (MuSiC-ViT) with an anatomy matching module (AMM). A Siamese network is a type of architecture that can decide based on two images. We chose a Siamese [18] network because CXR images from the baseline and follow-up exams must be compared to diagnose changes. The MuSiC-ViT model combines CNNs and visual transformers (CMTs) and is trained with multi-task learning to mimic the cognitive process of a radiologist in classifying CXR pairs and follow-up radiographs. The AMM is an attentional module that focuses on related regions in corresponding CXR pairs. The MuSiC-ViT model has the potential to improve the accuracy and efficiency of CXR change detection in clinical practice.

To mimic the cognitive process of radiologists when interpreting follow-up chest radiographs (CXRs), MuSiC-ViT was trained with three tasks simultaneously. The first task classifies a CXR as normal or abnormal based on the baseline and follow-up images. The second task trains the anatomy matching module (AMM) to focus on the same region in each CXR image. The third task is to determine if there are any changes in the follow-up CXR images. The main contributions of our study are:

- MuSiC-ViT mimics the clinical screening process of a radiologist.
- The AMM incorporated into MuSiC-ViT ensures that the model searches for similar regions in the CXR pairs of baseline and follow-up exams.
- MuSiC-ViT was trained on a large, high-resolution CXR dataset (88,000 changed and 115,000 no-change pairs of 512 × 512 pixels) and validated with one internal and two external datasets.
- MuSiC-ViT can perform classification of lung diseases according to change and no-change.

Materials and Methods

Dataset

This study was approved by the institutional review board (IRB number: 2019-0321), and written informed consent was waived due to the retrospective nature of the study. The CXRs were collected between 2011 and 2018 at the Department of Radiology, Asan Medical Center, with the exception of those without a follow-up CXR. Some examples of CXR pairs from the baseline and follow-up exams are shown in Figure 1.

Figure. 1. Examples of CXR pairs at the beginning and end of the study. (a), (c) No difference between pairs except for posture, diaphragm height, and breath-hold level, (b) Observable consolidation in the left lung area in the baseline radiograph and significant change in consolidation in the follow-up radiograph. (d) A small nodule in the left lung near the apex of the heart on the follow-up radiograph compared with the baseline chest radiograph.

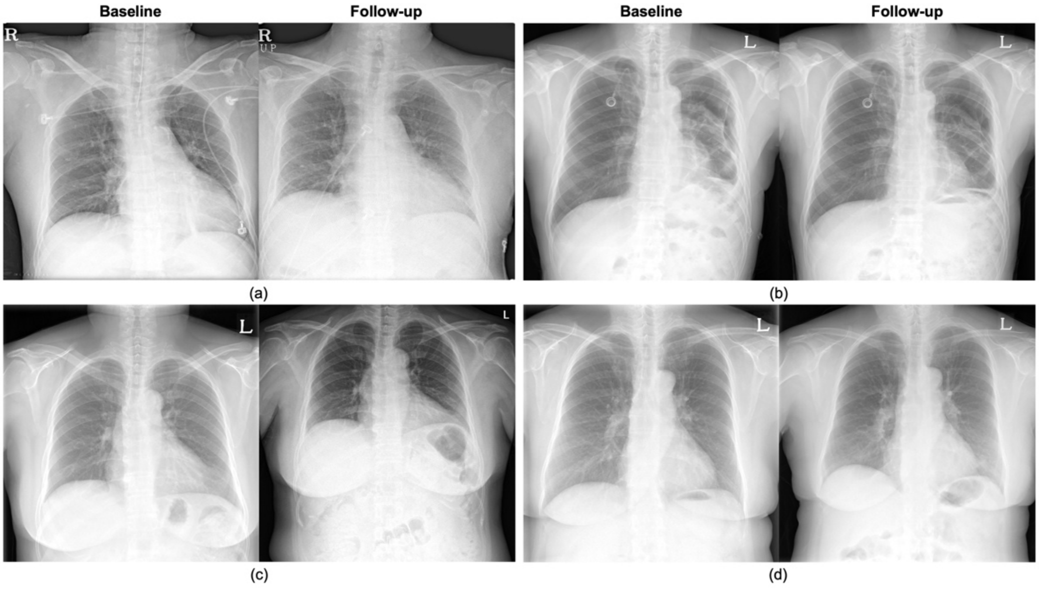
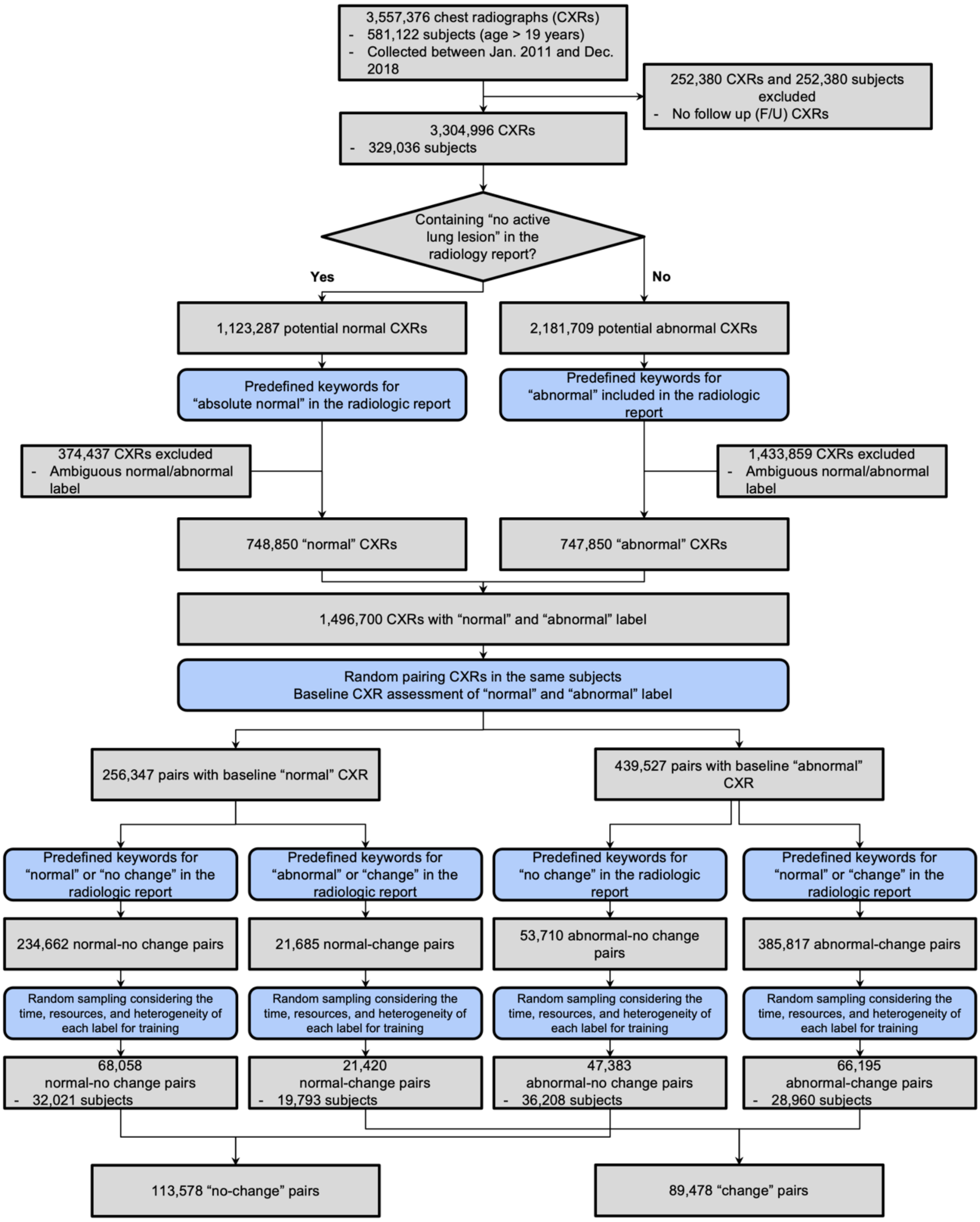


Figure. 2. Flowchart of data collection and labeling criteria in our study.


3,557,376 chest radiographs (CXRs)
- 581,122 subjects (age > 19 years)
- Collected between Jan. 2011 and Dec. 2018
252,380 CXRs and 252,380 subjects excluded
- No follow up (F/U) CXRs
3,304,996 CXRs
- 329,036 subjects
Containing "no active lung lesion" in the radiology report?
Yes
No
1,123,287 potential normal CXRs
2,181,709 potential abnormal CXRs
Predefined keywords for "absolute normal" in the radiologic report
Predefined keywords for "abnormal" included in the radiologic report
374,437 CXRs excluded
- Ambiguous normal/abnormal label
1,433,859 CXRs excluded
- Ambiguous normal/abnormal label
748,850 "normal" CXRs
747,850 "abnormal" CXRs
1,496,700 CXRs with "normal" and "abnormal" label
Random pairing CXRs in the same subjects
Baseline CXR assessment of "normal" and "abnormal" label
256,347 pairs with baseline "normal" CXR
439,527 pairs with baseline "abnormal" CXR
Predefined keywords for "normal" or "no change" in the radiologic report
Predefined keywords for "abnormal" or "change" in the radiologic report
Predefined keywords for "no change" in the radiologic report
Predefined keywords for "normal" or "change" in the radiologic report
234,662 normal-no change pairs
21,685 normal-change pairs
53,710 abnormal-no change pairs
385,817 abnormal-change pairs
Random sampling considering the time, resources, and heterogeneity of each label for training
Random sampling considering the time, resources, and heterogeneity of each label for training
Random sampling considering the time, resources, and heterogeneity of each label for training
Random sampling considering the time, resources, and heterogeneity of each label for training
68,058 normal-no change pairs
- 32,021 subjects
21,420 normal-change pairs
- 19,793 subjects
47,383 abnormal-no change pairs
- 36,208 subjects
66,195 abnormal-change pairs
- 28,960 subjects
113,578 "no-change" pairs
89,478 "change" pairs

Each chest radiograph was automatically labeled based on in-house guidelines, using predefined keywords such as “normal,” “abnormal,” “change,” and “no change,” after being reviewed multiple times by an experienced thoracic radiologist with over 20 years of experience. The “normal” or “abnormal” label for an image pair was determined based on the baseline CXR report, while the “change” or “no change” label for an image pair was determined based on the follow-up CXR examination. CXRs with descriptions similar to the keywords “no active lung lesion” were labeled as "normal," and CXRs that resembled keywords such as “increase*”, “decrease*”, “new*”, “improvement*”, “aggravation*”, “progress*”, “resolution*”, “disappearance*”, “cure*”, and “enlargement*” in the radiology report were labeled as “change.” CXRs with keywords such as “no interval,” “not significant,” and “not remarkable” were labeled as “no change”. Finally, each pair of baseline and follow-up CXRs was classified into one of four categories: “normal, no change,” “normal change,” “abnormal, no change,” and “abnormal change.” “Normal, no change” refers to a pair of CXRs in which the reading in the baseline CXR was normal, and the follow-up image showed no change or was normal. “Normal change” refers to a pair of CXRs in which the baseline CXR was normal and the follow-up CXR was abnormal. “Abnormal, no change” refers to a pair of CXRs in which the baseline CXR was abnormal and the follow-up CXR showed no change. “Abnormal change” refers to a pair of CXRs in which the baseline CXR was abnormal and the follow-up CXR was normal, or in which the predefined keywords used to determine the change were included in the follow-up CXR result. A schematic of the data collection algorithm is shown in Figure 2.

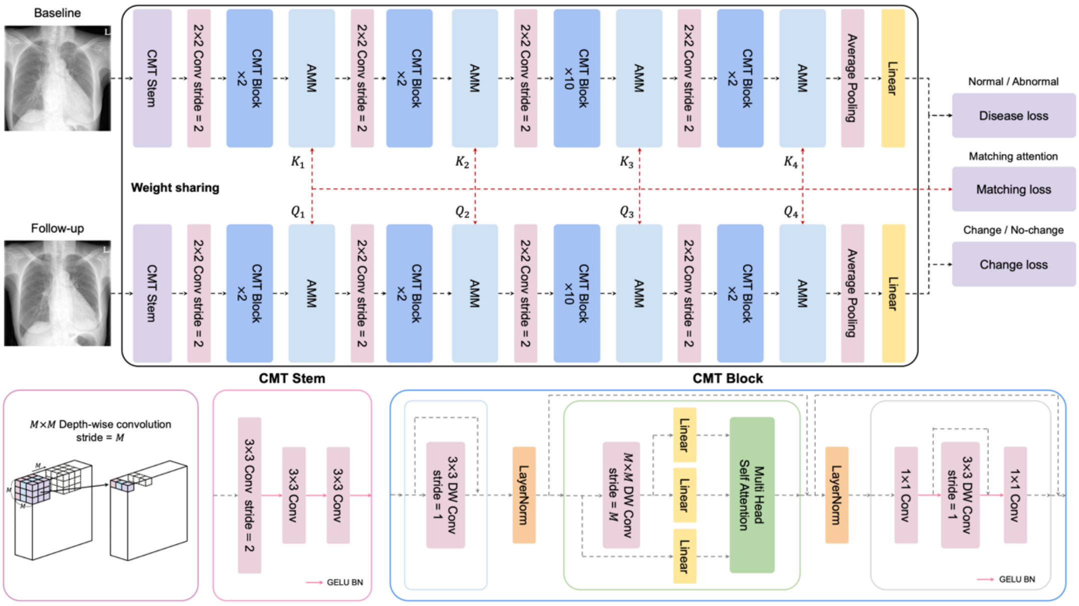
Tyops in the radiology report were checked using the “language-check” package in Python. A total of 406K CXRs were included in the training dataset, which included 161K normal, 245K abnormal, 88K change, and 115K no-change pairs. Visual validation by an experienced thoracic radiologist with over 20 years of experience on a randomly selected subset of the dataset showed that the overall accuracy of the change and no-change labeling from the training dataset obtained from the radiology reports was approximately 80%.

For the internal validation dataset, 810 pairs of change and no-change CXRs were randomly selected. In addition, two external validation datasets were used. One was randomly selected from the CheXpert dataset [19], which contained 215 pairs of change and no-change CXRs that were fully labeled by an experienced thoracic radiologist with over 15 years of experience. The other dataset, containing 267 change pairs and 266 no-change pairs, was collected from the same center as the training dataset but at different time periods. To evaluate the performance of the model, a thoracic radiology expert confirmed the labels of the internal and external validation datasets.

Multi-task Siamese convolutional vision transformer

MuSiC-ViT has three main parts, as shown in Figure 3. First, a pair of CXR images is provided as input to the Siamese network. Second, local and global structure information is embedded using CMT blocks to extract various features from the CXR images. Third, anatomy matching between the two CXR images is performed during the embedding process using the AMM, which helps the model classify the change/no-change class based on similar regions of a pair of baseline and follow-up CXR images. Finally, the model learns the change/no change classification of the CXR pair and the normal/abnormal classification for each CXR image in parallel.

Figure. 3. The overall workflow of the proposed method includes an anatomy matching module (AMM) and a Siamese network based on convolutional neural networks meets vision transformers (CMTs) as a backbone model. The AMM is based on an attentional mechanism. This AMM is combined with the overall process to attend to similar regions in CXR pairs.



Siamese network architecture

The encoder network uses a pair of baseline and follow-up chest radiographs (CXRs) as input. The Siamese network architecture was chosen because it allows weight sharing and has a efficient computational cost. The encoder network produces two encoded vectors, one for the baseline CXR and one for the follow-up CXR. These two vectors are then concatenated together.

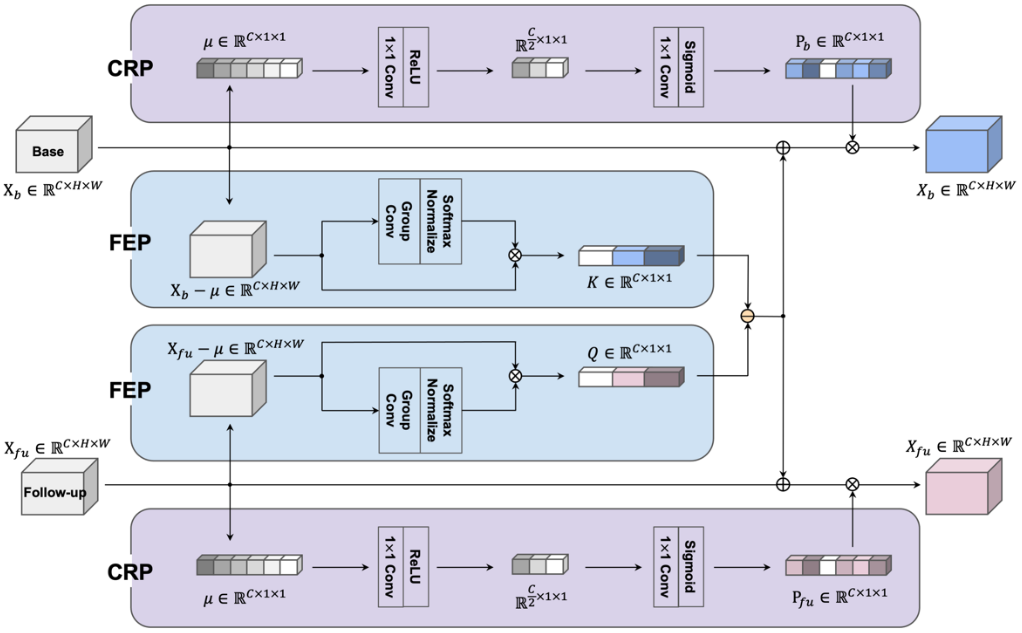
CNNs meet vision transformers

In contrast to a CNN-based model, a vision transformer (ViT) [20-22] based on the "attention" mechanism uses a large receptive field and can be robustly trained on global image features using the attentional mechanism. However, spatial localization may not be considered because the transformer architecture is trained using a sequence of small image patches, and the fixed patch size may make it difficult to extract low-resolution multiscale features, which are important for chest radiograph (CXR) screening [23]. Modern CNN architectures, on the other hand, are effective at generating multiscale features [24]. Therefore, a convolutional multi-headed transformer (CMT) architecture, which combines CNN and ViT, was chosen. CMT can efficiently capture local and global structural information using depth-wise convolution and multi-head self-attention.

Anatomy matching module

An anatomy matching module (AMM) was used as an attentional mechanism to connect two feature maps because the attention method has been successful in natural language processing [25-27] and computer vision [20,28,29]. The AMM is shown in Figure. 4.

Figure. 4. Architecture of anatomy matching module.



AMM consists of a feature extraction part (FEP) and a channel re-calibration part (CRP). The CRP generates $P(\cdot)$, which represents the channel re-calibrating features. The FEP generates $K \in \mathbb{R}^{C\times1\times1}$ and $Q \in \mathbb{R}^{C\times1\times1}$, which represents softmax-normalized features of the baseline and follow-up CXRs, respectively; C denotes the number of channels.

The CRP is the weighted average channel attention. When using FEP and CRP, AMM ensures that K and Q focus on similar feature maps of the two CXRs.

Attend and compare module (ACM) [30] is a method that induces the model to focus on different regions by difference modeling of K and Q produced by FEP in a "single CXR image." After generating two feature maps, K_i and Q_i, the difference of the maps is added to x. Therefore, ACM [30] can be represented as:

$$x = P(x) \otimes \left(x + (K_i - Q_i)\right), \tag{1}$$

where $i \in \{1,2,3,4\}$, i means i-th block index of CMT [24], and $\otimes$ denotes element-wise multiplication**.**

Moreover, ACM uses orthogonal loss to train two feature maps K and Q to obtain different information.

$$L_{orthogonal} = \frac{1}{n}\sum_{i=1}^{n}(K_i \cdot Q_i)/\, C \tag{2}$$

where $n = 4$ and C is the number of channels.

However, our AMM causes the MuSiC-ViT to focus on similar regions by modeling the similarity of K and Q generated by FEP in "two CXR images" The cosine similarity was also used as a loss function to maximize anatomy matching. The process can be represented as follows:

$$cos\ sim(K_i, Q_i) = (K_i \cdot Q_i)/\|K_i\|\|Q_i\| \tag{3}$$

$$L_{matching} = 2 - 2 \cdot \frac{1}{n}\sum_{i=1}^{n} cos\ sim(\ K_i, Q_i) \tag{4}$$

where $n = 4$.

Finally, using the matching loss above, similar anatomy matching was performed between the baseline and follow-up CXRs.

Multi-task learning and loss functions

To mimic the cognitive process of a radiologist, we trained MuSiC-ViT on several tasks, including classifying a pair of baseline and follow-up chest radiograph (CXR) images into change/no change, normal/abnormal, and similar anatomical regions. Two cross-entropy losses were calculated to distinguish the disease and normal (i.e., label) classes from the baseline and follow-up CXR images. One of the cross-entropy loss functions used to determine the change/no change class of a CXR pair is calculated as follows:

$$S(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}}\ , for\ i = 1, \ldots, K \tag{5}$$

$$CE(y, f(x)) = -\sum_i y_i \log f(x_i) \tag{6}$$

$$L_{disease} = CE(\ y_b\ ,\ S(\ W_1 \cdot f(x_b)) + CE\big(\ y_{fu}\ ,\ S(W_2 \cdot f\big(x_{fu}\big))\big) \tag{7}$$

$$L_{change} = CE\big(y_{change}\ ,\ S(W_3 \cdot f(x_b) \oplus f\big(x_{fu}\big))\big) \tag{8}$$

where $S(\cdot)$ is softmax function, $CE(\cdot)$ is cross-entropy, y_b and y_{fu} represent the normal/abnormal disease label of baseline and follow-up CXRs, respectively, y_{change} is the change/no-change class label of the CXR pair, and $\oplus$ denotes vector-wise concatenation. Each

W_1, W_2, and W_3 denotes a multi-layer perceptron.

Finally, MuSiC-ViT could adequately learn to classify change/no-change class of a patient by utilizing the three loss functions (i.e., $L_{matching}$, $L_{disease}$, and L_{change}) regardless of aging and breath-hold level variations, which were not the differences of interest.

Our overall loss function is a weighted sum of the three loss functions. The final loss function is described as:

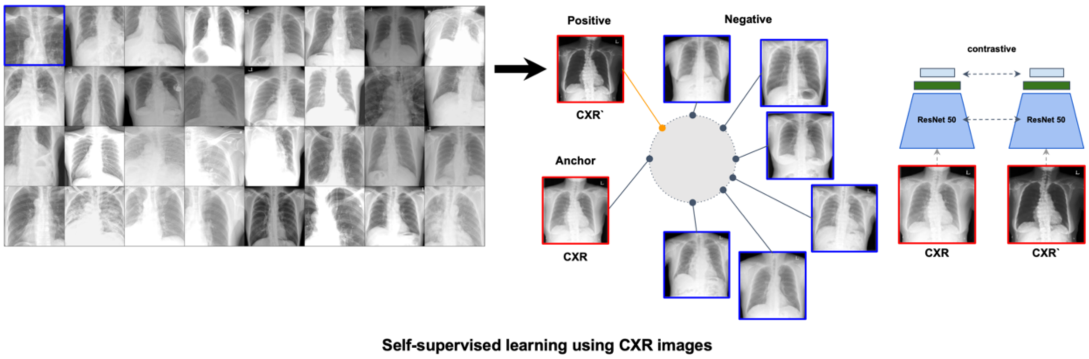
$$L_{total} = \lambda_1 L_{change} + \lambda_2 L_{disease} + \lambda_3 L_{matching} \tag{9}$$

We achieved the best performance when λ_1, λ_2, and λ_3 were set to 1, 0.1, and 0.01, respectively, based on the lambda ablation study in Section III.

Pre-training model using self-supervised learning in CXR

In general, expert radiologists have prior knowledge of various medical domains (e.g., X-rays). From the perspective of deep learning, obtaining such prior knowledge can be seen as learning a well-pretrained model. In natural images, the ImageNet model is often used as a pretrained model. However, there are fundamental differences between medical images and natural images. Medical images, unlike natural images, are acquired according to standardized protocols and are therefore standardized to some extent (e.g., anterior-posterior, posterior-anterior, or lateral). As a result, common deep learning tasks such as classification, recognition, and segmentation of medical images may depend on extremely small differences compared to natural images. These small differences also limit the resolutions and magnifications of medical images, making them more challenging. Resizing medical images to low resolution can degrade image quality and even remove the disease region. Strong augmentation methods, such as rotating, flipping, and cropping, are also limited due to the nature of medical images, as the augmented images may appear inappropriate in a medical context. In addition, medical image assessment and annotation is extremely difficult and costly, as it can only be performed by experts. Self-supervised learning (SSL) can be a solution to these problems. SSL is a pretraining strategy that uses unlabeled data to learn a pretext task, combining supervised and unsupervised learning. Using this SSL method, we trained the model using the properties of the longitudinal dataset we have. We trained MoCo-v2 [31], a pretrained model that is suitable for CXRs, rather than the ImageNet pretrained model that is commonly used for natural images. Finally, using the pretrained model we trained, the supervised learning results for change/no-change classification are presented in Section III. Figure 5 shows the SSL method.

Figure. 5. Self-supervised learning method using CXR dataset.



Training details

Various data augmentation techniques were used during the training phase, including shifting, scaling, rotating, sharpening, adding motion blur, median blur, optical distortion, Gaussian noise, and contrast-limited adaptive histogram equalization. The resolution of the input image was set to 512 × 512. For training, 1 GPU (NVIDIA Titan RTX) and a batch size of 14 were used. All models were implemented in Pytorch (version 1.7.1). The AdamW [32] with a multi-step learning rate decay was used during training. The learning rate was set as 0.00001.

Statistical analyses

DeLong's test [33] for two receiver operating characteristic curves (ROC for change/no-change classification was used to compare the performance of each model. R statistical version 4.2.0 (R Foundation for Statistical Computing, Vienna, Austria) was used for statistical analyzes. Statistical results were analyzed with two-sided P values, with statistical significance set at 0.05 alpha.

Results

Various model architecture comparison study

The commonly used models in medical image were comparatively studied. Table 1 presents the performances of CMT-Ti, a CNN-based model, a vision transformer, and MuSiC-ViT for the internal and external validation datasets. Inception-v3 [34], ResNet-50 [35], DenseNet-121 [36], EfficientNet-b3 [37], EfficientNet-v2 [38], and ConvNeXt [39] were chosen for the CNN-based models, which are commonly used encoder networks. Next, ViT-B [20], Swin v2 [40], CoaT [41], MLP-Mixer [42], and ResMLP [43] were chosen as the widely used vision transformer models.

All architectures were changed to Siamese to achieve change/no-change classification. As a result, all validation results showed better performance with MuSiC-ViT than the other architectures. Vision transformer (PVT [44] and CaiT [45] achieved AUCs of 0.500 and 0.506, respectively) models without convolution could not be trained well using high-resolution images (512 × 512), except for the original vision transformer model. Furthermore, CMT-Ti had the best performance among the other architectures apart from MuSiC-ViT. In addition, for better representation learning, ResNet-50, which is commonly used in medical imaging, and ResNet-50 trained with MoCo-v2 using CXR dataset, and MuSiC-ViT were comparatively studied. Table 2 presents the performances of ResNet-50, SSL ResNet-50 and MuSiC-ViT model. All validation results showed worse results than MuSiC-ViT.

Table 1. Performance comparison based on various model architectures using the three validation datasets

Model architecture		Param (M)	Training (h) Testing (s)	Internal validation dataset				External validation dataset 1				External validation dataset 2			
				SPE	SEN	ACC	AUC	SPE	SEN	ACC	AUC	SPE	SEN	ACC	AUC
CNN	Inception-v3	22.84M	31h/36s	0.790	0.568	0.679	0.732***	0.884	0.265	0.574	0.665***	0.866	0.449	0.659	0.723***
	ResNet-50	24.55M	36h/24s	0.806	0.564	0.685	0.749***	**0.902**	0.279	0.591	0.639***	0.828	0.453	0.642	0.721***
	DenseNet-121	7.48M	32h/46s	0.752	0.595	0.674	0.722***	0.828	0.363	0.595	0.662***	0.869	0.532	0.702	0.758***
	EfficientNet-b3	11.49M	29h/44s	0.790	0.557	0.674	0.741***	0.888	0.270	0.579	0.655***	0.884	0.506	0.696	0.655***
	EfficientNet-v2	21.23M	49h/48s	0.736	0.594	0.665	0.712***	0.726	0.437	0.581	0.649***	0.866	0.543	**0.705**	**0.760*** **
	ConvNeXt	28.59M	67h/29s	0.751	0.544	0.648	0.699***	0.833	0.4	0.616	**0.689** **	0.914	0.415	0.666	0.736***
DNN	MLP-Mixer	32.18M	52h/24s	0.706	**0.656**	0.681	0.727***	0.754	**0.516**	**0.635**	0.660***	0.787	0.521	0.655	0.705***
Transformer	Swin v2	28.35M	54h/37s	0.665	0.611	0.638	0.692***	0.842	0.349	0.595	0.637***	0.892	0.433	0.664	0.742***
	ViT-B	88.28M	33h/17s	**0.816**	0.448	0.632	0.677***	**0.902**	0.200	0.551	0.618***	**0.978**	0.102	0.542	0.633***
	ResMLP	20.21M	56h/16s	0.704	0.649	0.677	0.724***	0.707	0.414	0.561	0.609***	0.758	0.411	0.585	0.617***
	CoaT	11.01M	42h/38s	0.709	0.651	0.680	0.734***	0.795	0.335	0.565	0.609***	0.724	0.555	0.640	0.698***
	CMT-Ti	31.62M	29h/59s	0.721	**0.682**	**0.701**	**0.762** **	0.795	**0.488**	**0.642**	0.674***	0.772	**0.626**	0.700	0.757***
	MuSiC-ViT (ours)	31.81M	39h/70s	**0.817**	0.638	**0.728**	**0.797**	**0.930**	0.298	0.614	**0.784**	**0.899**	0.589	**0.745**	**0.858**

Note: number of model parameters (Param); specificity (SPE); sensitivity (SEN); accuracy (ACC); area under receiver operating characteristics curve (AUC); million (M); hours (h); seconds (s); P-value < 0.05 (*); P-value < 0.01 (**); P-value < 0.001(***). DeLong's test was conducted to compare the model performance between MuSiC-ViT and the other models.

Table 2. Performance comparison based on self-supervised learning method using the three validation datasets

Model architecture		Param (M)	Training (h) Testing (s)	Internal validation dataset				External validation dataset 1				External validation dataset 2			
				SPE	SEN	ACC	AUC	SPE	SEN	ACC	AUC	SPE	SEN	ACC	AUC
Self-supervised learning	SSL ResNet-50	24.55 M	36h/24 s	0.759	0.547	0.652	0.701^{***}	0.758	0.437	0.598	0.611^{***}	0.769	0.506	0.638	0.660^{***}
Supervised learning	ResNet-50	24.55 M	36h/24 s	0.806	0.564	0.685	0.749^{***}	**0.902**	0.279	0.591	0.639^{***}	0.828	0.453	0.642	0.721^{***}
	MuSiC-ViT (ours)	31.81 M	39h/70 s	**0.817**	0.638	**0.728**	**0.797**	**0.930**	0.298	0.614	**0.784**	**0.899**	0.589	**0.745**	**0.858**

Note: number of model parameters (Param); specificity (SPE); sensitivity (SEN); accuracy (ACC); area under receiver operating characteristics curve (AUC); million (M); hours (h); seconds (s); self-supervised learning (SSL); P-value < 0.05 (*); P-value < 0.01 (**); P-value < 0.001(***). DeLong's test was conducted to compare the model performance between MuSiC-ViT and the other models.

Ablation studies

To evaluate the performance of the MuSiC-ViT model for change detection in CXR follow-up images, several metrics were used: the area under the receiver operating characteristic curve (AUC), sensitivity (SEN), specificity (SPE), and accuracy (ACC). Ablation studies were also conducted to assess the contribution of each component to the performance of MuSiC-ViT. The results of the ablation studies are shown in Table 3. When the anatomy matching module (AMM) was added to the model, the performance improved significantly in both the internal and external validation datasets. The best model achieved SPE, SEN, ACC, and AUC values of 0.817, 0.638, 0.728, and 0.797 for the internal validation dataset, 0.930, 0.298, 0.614, and 0.784 for the first external validation dataset, and 0.899, 0.589, 0.745, and 0.858 for the second external validation dataset, respectively. Figure. 6 shows saliency maps of the change pair made with Grad-CAM [46]. We generated the intersection saliency maps of Grad-CAM data between the baseline and follow-up CXRs to illustrate the heatmap for both pictures. The "change" lesion between the baseline and follow-up CXRs was then seen by overlaying the heatmap over the baseline CXR.

Figure 6. Saliency maps with baseline CXR, follow-up CXR, and Grad-CAM data are shown as examples of "change" prediction pair (from left to right). The radiologist judged that certain diseased region on the baseline and follow-up CXR images represented a change. (a) Reduced right pleural effusion and atelectasis lesions are shown on a pair of baseline and follow-up CXR pictures. (b) Images demonstrating pleural effusion lesions and a pneumothorax that have grown larger.

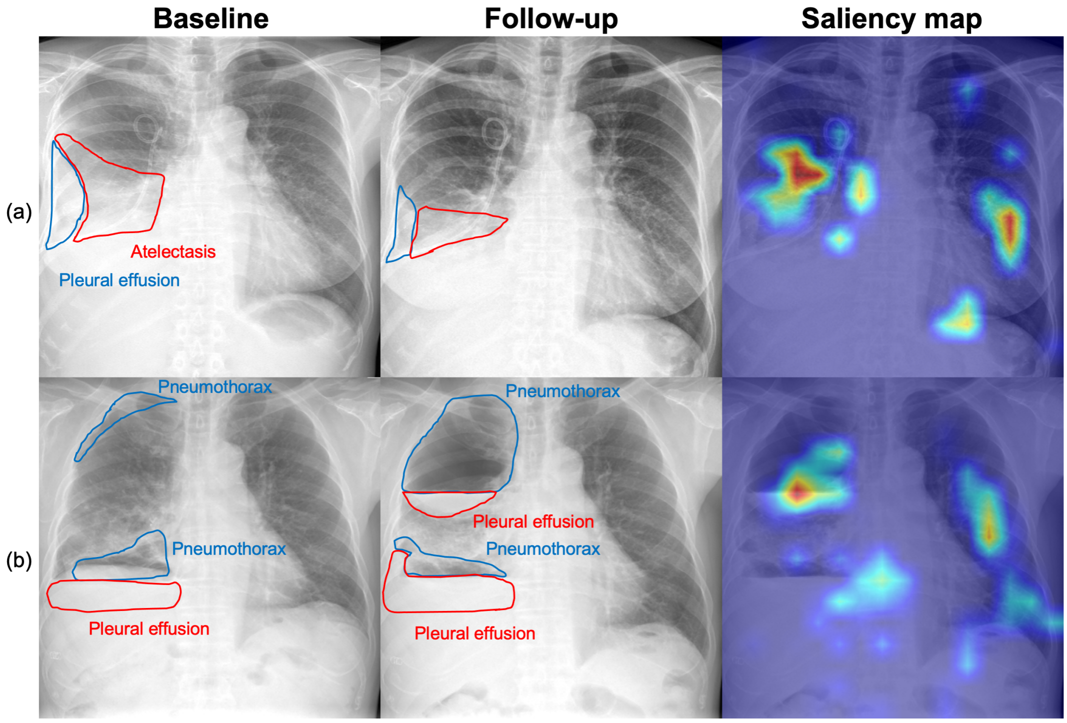


Table 3. Three-validation dataset comparison of ablation studies

Network module	Internal validation				External validation 1				External validation 2			
	SPE	SEN	ACC	AUC	SPE	SEN	ACC	AUC	SPE	SEN	ACC	AUC
Siamese CMT	0.721	**0.682**	0.701	0.762**	0.795	**0.488**	**0.642**	0.674	0.772	**0.626**	0.700	0.757***
Siamese CMT + D	0.773	0.586	0.680	0.736***	0.777	0.386	0.581	0.641**	0.806	0.551	0.679	0.722***
Siamese CMT+ AMM	**0.779**	0.637	**0.708**	**0.779**	**0.865**	**0.461**	**0.663**	**0.754**	**0.896**	**0.604**	**0.751**	**0.836**
Siamese CMT + AMM + D (MuSiC-ViT)	**0.817**	**0.638**	**0.728**	**0.797**	**0.930**	0.298	0.614	**0.784**	**0.899**	0.588	**0.745**	**0.858**

Note: area under receiver operating characteristics curve (AUC); P-value 0.05 (*); P-value 0.01 (**); and P-value 0.001 (***); anatomy matching module (AMM); normal/abnormal disease label (D); specificity (SPE); sensitivity; and accuracy. To compare the model performance of MuSiC-ViT and the other models, DeLong's test was performed. The best performance is denoted by red, and the second-best performance is denoted by blue.

The CMT's baseline model performance, which was employed as the encoder in MuSiC-ViT, is displayed in Table 4. The model's hyperparameter was set for CMT architectural variations in accordance with the recommendation made by Guo J et al. [24]. Since CMT-Ti had the best AUC for the internal validation dataset, it was chosen as the backbone model.

Table 4. Model parameters of the ablation studies on the internal validation dataset

Backbone size	Model parameter	SPE	SEN	ACC	AUC
CMT-XTi	16.38M	0.730	0.660	0.695	0.754
CMT-Ti	31.62M	0.721	**0.682**	**0.701**	**0.762**
CMT-XS	45.81M	**0.772**	0.484	0.628	0.693^{***}
CMT-S	59.99M	0.747	0.475	0.611	0.662^{***}
CMT-B	88.70M	0.756	0.526	0.641	0.693^{***}

Note: extra-tiny (XTi); tiny (Ti); extra-small (XS); small (S); big (B); specificity (SPE); sensitivity (SEN); accuracy (ACC); area under receiver operating characteristics curve (AUC); P-value < 0.05 (*); P-value < 0.01 (**); P-value < 0.001 (***). DeLong's test was performed to compare the model performance between CMT-Ti and the other models.

Results from the lambda ablation study for MuSiC-ViT, which uses multiple losses to simultaneously solve multi-task problems, are shown in Table 5. The coefficients λ_1, λ_2, and λ_3 were used for change, disease, and matching losses. Due to the limited time and GPU resources, λ_1 and λ_3 were set to 1 when seeking the best lambda for λ_2, and λ_1 and λ_2 were set to 1 when seeking the best lambda for λ_3. Finally, we achieved the best AUC of 0.797 with the final lambda values when λ_1, λ_2, and λ_3 were set to 1, 0.1 and, 0.01, respectively.

Table 5. Lambda ablation study with the loss function using the internal validation dataset

Lambda ratio	SPE	SEN	ACC	AUC
λ_1: λ_2: λ_3=1:1:1	0.806	0.614	0.710	0.784
λ_1: λ_2: λ_3=1:1:0.1	0.774	0.651	0.712	0.781
λ_1: λ_2: λ_3=1:1:0.01	0.809	0.631	0.720	0.791
λ_1: λ_2: λ_3=1:1:0.001	0.768	0.668	0.718	0.783
λ_1: λ_2: λ_3=1:0.5:1	0.765	0.675	0.720	0.778
λ_1: λ_2: λ_3=1:0.1:1	0.732	**0.701**	0.717	0.787
λ_1: λ_2: λ_3=1:0.01:1	0.749	0.688	0.719	0.783
λ_1: λ_2: λ_3=1:0.1:0.01	**0.817**	0.638	**0.728**	**0.797**

Note: specificity (SPE); sensitivity (SEN); accuracy (ACC); area under receiver operating characteristics curve (AUC); lambda for change/no-change (λ_1); lambda for disease loss (λ_2);

lambda for matching loss (λ_3); P-value < 0.05 (*); P-value < 0.01 (**); P-value < 0.001(***). DeLong's test was conducted to compare the model performance between the best-performing model and the other models.

Discussion

The radiologic reading of follow-up CXRs is one of the most crucial jobs [47-49] due to the importance of follow-up CXRs in patient monitoring and abnormality detection. However, in actual clinical situations, they may result in an enormous task for radiologists, preventing quick reporting and delaying diagnosis. Therefore, by reducing the number of CXR pairs to read, the model-based triage or automatic screening system could be utilized to improve the radiologic workflow and patients' safety if it may accurately detect no-change or clinically significant alterations in CXR pairs. When a big change takes place, it would give the clinician more time and focus.

Overall, the MuSiC-ViT model demonstrated good performance in detecting changes in CXR follow-up images, with AUC values of 0.797 for the internal validation dataset, 0.784 for the first external validation dataset, and 0.858 for the second external validation dataset. The model also showed high sensitivity and specificity, with values ranging from 0.638 to 0.930 for sensitivity and 0.614 to 0.899 for specificity. In addition, the model achieved good accuracy, with values ranging from 0.614 to 0.745. The results of the ablation studies indicated that the anatomy matching module (AMM) contributed significantly to the performance of the model. Overall, the MuSiC-ViT model appears to be a promising tool for assisting radiologists in the detection of changes in CXR follow-up images.

Multi-task learning of change/no-change classification as well as normal/abnormal classification, and matching similar anatomic regions were carried out for a follow-up CXR classification. The performance improved as the lambda of anatomy matching loss decreased because MuSiC-ViT tended to focus more readily on matching anatomically comparable parts than on change/no-change classification across multiple tasks. The AMM significantly impacted the change/no-change classification performance in the ablation study of the matching module and extra information. The performance increased when the lambda of disease loss decreased because normal/abnormal classification was simpler than change/no-change classification. After an ablation investigation of the lambda of the loss function, the coefficients for the change, disease, and matching losses were set to 1, 0.1, and 0.01, respectively.

In addition, based on internal and external validations, our technique outperformed and generalized better than CNN-based and ViT-based models. Both employing CMT-Ti to get over the low-resolution issue brought on by ViT's patch embedding and CNN's inadequate generalizability, performance was improved. Based on the quality of the training dataset, the enhanced performance gains of AMM and extra illness labels were also contrasted with the base CMT-Ti model.

As last, the SSL ResNet-50 model trained in experiments to have for better representation learning did not perform better than MuSiC-ViT. In this experiment, it was shown that an appropriate self-supervised learning pretext task suitable for the medical image was required, not just hard augmentation learning.

Our study has several advantages. First, to the best of our knowledge, this is the first study to simulate the cognitive process of follow-up CXR classification used by radiologists without focusing on a single lesion. A previous study [50] have focused on a single lesion or a constrained setting. In contrast, our MuSiC-ViT performed well in classifying specific disorders with changes or no changes. Second, our MuSiC-ViT used multi-task learning to address the challenging clinical situations involved in follow-up CXR classification using an AMM, following the same procedure as a radiologist.

However, this study does have some limitations. First, according to the radiologist's visual scoring findings collected through random sampling, the change/no change labeling accuracy of our training dataset was around 80%. This was due to the challenges of handling large amounts of real clinical data. Instead of using a complex radiologist screening method, the labels for the training dataset were created using rules for natural language processing. Therefore, in order to do further research, we will need to collect and examine training data using a stress test. Second, the accuracies of the internal and external validations may not appear to be very high compared to the results of the binary classification task. However, since even radiologists may mistakenly classify a change in a CXR due to its difficulty, our findings may still be considered reasonable in some cases. Third, the CheXpert dataset was subjected to thorough histogram equalization pre-processing. However, because this image processing technique is not commonly used on real-world data, there may be differences in domain characteristics between the datasets used for training and external validation. Fourth, there has been no previous research on follow-up CXR classification, so it may be difficult to strictly evaluate our approach.

Singh et al. [49] conducted a deep learning-based follow-up classification, but their study was only focused on certain findings (e.g., lines and tubes, pneumothorax, fibrosis, pulmonary nodules, and masses). Additionally, the change detection data used in the field of remote sensing (such as roads, structures, and croplands) are mainly stationary or barely altered. Therefore, it is difficult to compare our method to the change detection methods used in the study by Singh et al., which used the change map as the ground truth. Finally, more ablation studies, stress tests, and parameter searches are needed to evaluate the model's sensitivity and robustness.

Conclusion

In conclusion, we propose MuSiC-ViT, a model that is able to perform one of the primary responsibilities of radiologists, which is comparing baseline and follow-up CXRs to distinguish between pairs of CXRs with change and no change. MuSiC-ViT can compare similar regions of interest in each CXR by incorporating an AMM (Automatic Matching Module). The AMM allows the model to focus more on lesions while ignoring the patient's natural fluctuations, such as breath-hold level, aging, and posture. Disease loss can help the model differentiate between abnormal and normal CXRs in the case of disease. This architecture may serve as a catalyst for future CXR experiments and result in practical clinical applications. In further research, our MuSiC-ViT model may be expanded to consider radiology reports simultaneously, to generate linguistic explanations based on follow-up CXR, which may be more useful for radiologists in actual clinical applications [51-53]

References

1 Habra, M. A. & Vassilopoulou-Sellin, R. Contribution of routine chest x-ray in the long-term follow-up of patients with differentiated thyroid carcinoma. *Thyroid* **16**, 303-306 (2006).

2 Alarcón-Rodríguez, J. *et al.* Radiological management and follow-up of post-COVID-19 patients. *Radiología (english edition)* **63**, 258-269 (2021).

3 Scholtz, J.-E. *et al.* Incidental pulmonary nodules in emergent coronary CT angiography for suspected acute coronary syndrome: Impact of revised 2017 Fleischner Society Guidelines. *Journal of cardiovascular computed tomography* **12**, 28-33 (2018).

4 Tigges, S., Roberts, D. L., Vydareny, K. H. & Schulman, D. A. Routine chest radiography in a primary care setting. *Radiology* **233**, 575-578 (2004).

5 Tang, Y.-X. *et al.* Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ digital medicine* **3**, 1-8 (2020).

6 Meedeniya, D. *et al.* Chest X-ray analysis empowered with deep learning: A systematic review. *Applied Soft Computing*, 109319 (2022).

7 Zebin, T. & Rezvy, S. COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization. *Applied Intelligence* **51**, 1010-1021 (2021).

8 Ho, T. K. K. & Gwak, J. Utilizing knowledge distillation in deep learning for classification of chest X-ray abnormalities. *IEEE Access* **8**, 160749-160761 (2020).

9 Sarkar, A., Vandenhirtz, J., Nagy, J., Bacsa, D. & Riley, M. Identification of images of COVID-19 from chest X-rays using deep learning: comparing COGNEX VisionPro deep learning 1.0™ software with open source convolutional neural networks. *SN Computer Science* **2**, 1-16 (2021).

10 Frid-Adar, M., Amer, R., Gozes, O., Nassar, J. & Greenspan, H. COVID-19 in CXR: From detection and severity scoring to patient disease monitoring. *IEEE journal of biomedical and health informatics* **25**, 1892-1903 (2021).

11 Aminu, M., Ahmad, N. A. & Noor, M. H. M. Covid-19 detection via deep neural network and occlusion sensitivity maps. *Alexandria Engineering Journal* **60**, 4829-

4855 (2021).

12 Bhandary, A. *et al.* Deep-learning framework to detect lung abnormality–A study with chest X-Ray and lung CT scan images. *Pattern Recognition Letters* **129**, 271-278 (2020).

13 Fernando, C., Kolonne, S., Kumarasinghe, H. & Meedeniya, D. in *2022 2nd International Conference on Advanced Research in Computing (ICARC).* 165-170 (IEEE).

14 Brown, M. S., Wilson, L. S., Doust, B. D., Gill, R. W. & Sun, C. Knowledge-based method for segmentation and analysis of lung boundaries in chest X-ray images. *Computerized medical imaging and graphics* **22**, 463-477 (1998).

15 Gordienko, Y. *et al.* in *International conference on computer science, engineering and education applications.* 638-647 (Springer).

16 Hesamian, M. H., Jia, W., He, X. & Kennedy, P. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging* **32**, 582-596 (2019).

17 Kumarasinghe, K., Kolonne, S., Fernando, K. & Meedeniya, D. U-Net Based Chest X-ray Segmentation with Ensemble Classification for Covid-19 and Pneumonia. *International Journal of Online & Biomedical Engineering* **18** (2022).

18 Koch, G., Zemel, R. & Salakhutdinov, R. in *ICML deep learning workshop.* 0 (Lille).

19 Irvin, J. *et al.* in *Proceedings of the AAAI conference on artificial intelligence.* 590-597.

20 Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

21 Parmar, N. *et al.* in *International Conference on Machine Learning.* 4055-4064 (PMLR).

22 Yang, G., Tang, H., Ding, M., Sebe, N. & Ricci, E. Transformers solve the limited receptive field for monocular depth prediction. *arXiv e-prints*, arXiv: 2103.12091 (2021).

23 Monday, H. N. *et al.* COVID-19 Diagnosis from Chest X-ray Images Using a Robust Multi-Resolution Analysis Siamese Neural Network with Super-Resolution

Convolutional Neural Network. *Diagnostics* **12**, 741 (2022).

24 Guo, J. *et al.* Cmt: Convolutional neural networks meet vision transformers. *arXiv preprint arXiv:2107.06263* (2021).

25 Lin, Z. *et al.* A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

26 Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

27 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

28 Liu, Z. *et al.* in *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 10012-10022.

29 Touvron, H. *et al.* in *International Conference on Machine Learning.* 10347-10357 (PMLR).

30 Kim, M., Park, J., Na, S., Park, C. M. & Yoo, D. in *European Conference on Computer Vision.* 576-592 (Springer).

31 Chen, X., Fan, H., Girshick, R. & He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

32 Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

33 DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845 (1988).

34 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2818-2826.

35 He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770-778.

36 Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4700-4708.

37 Tan, M. & Le, Q. in *International conference on machine learning.* 6105-6114

(PMLR).

38 Tan, M. & Le, Q. in *International Conference on Machine Learning.* 10096-10106 (PMLR).

39 Liu, Z. *et al.* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 11976-11986.

40 Liu, Z. *et al.* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12009-12019.

41 Xu, W., Xu, Y., Chang, T. & Tu, Z. in *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 9981-9990.

42 Tolstikhin, I. O. *et al.* Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems* **34**, 24261-24272 (2021).

43 Touvron, H. *et al.* Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404* (2021).

44 Wang, W. *et al.* in *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 568-578.

45 Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G. & Jégou, H. in *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 32-42.

46 Selvaraju, R. R. *et al.* in *Proceedings of the IEEE international conference on computer vision.* 618-626.

47 Mañá, J. *et al.* Multidisciplinary approach and long-term follow-up in a series of 640 consecutive patients with sarcoidosis: Cohort study of a 40-year clinical experience at a tertiary referral center in Barcelona, Spain. *Medicine* **96** (2017).

48 Dobson, M. *et al.* What is the value of the lateral chest radiograph in the follow-up thoracic lymphoma? *European radiology* **7**, 1110-1113 (1997).

49 Singh, R. *et al.* Deep learning in chest radiography: detection of findings and presence of change. *PloS one* **13**, e0204155 (2018).

50 Macdonald, C., Jayathissa, S. & Leadbetter, M. Is post-pneumonia chest X-ray for lung malignancy useful? Results of an audit of current practice. *Internal Medicine Journal* **45**, 329-334 (2015).

51 Moon, J. H., Lee, H., Shin, W. & Choi, E. Multi-modal understanding and generation

for medical images and text via vision-language pre-training. *arXiv preprint arXiv:2105.11333* (2021).

52 Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747* (2020).

53 Liu, G. *et al.* in *Machine Learning for Healthcare Conference.* 249-269 (PMLR).

국문요약

일상적인 임상 환경에서 방사선 전문의의 주요 책임 중 하나는 환자 상태 변화를 식별하기 위해 follow-up 흉부 방사선 사진(CXR)을 진단하는 것이다. 방사선 전문의는 질병 변화를 자연적 또는 양성 변화와 구별해야 하기 때문에 follow-up CXR 의 의미 있는 변화를 진단하는 것은 많이 어렵다. 본 논문은 baseline 와 follow-up CXR 쌍에 대하여 변화 여부를 분류하기 위한 방사선 전문의의 인지 프로세스를 모방하기 위해 해부학적 일치 모듈(AMM)이 있는 다중 작업 샴 컨볼루션 비전 변환기 (MuSiC-ViT)를 제안한다. MuSiC-ViT 는 CNN 과 Vision transformer 를 결합한 CMT (CNNs meet Vision transformers) 모델을 사용하며 Siamese 아키텍처, AMM(Anatomy-Matching Module) 및 다중 작업 학습의 세 가지 주요 구성 요소가 있다. 입력이 baseline 과 follow-up 영상 한 쌍의 CXR 이기 때문에 인코더 네트워크에는 Siamese 네트워크가 선택되었다. AMM 은 해당 CXR 쌍의 관련 영역에 초점을 맞춘 attention 모듈이다. 방사선 전문의의 인지 과정을 모방하기 위해 MuSiC-ViT 는 다중 작업 학습, 정상/비정상, 변화/무변화, 해부학적 일치 모듈로 훈련되었다. 총 406,000 개의 CXR 이 연구에 사용되었으며 훈련 데이터 세트에 대해 88,000 개의 change 및 115,000 개의 no-change 쌍이 획득되었다. 내부 검증 데이터 세트의 경우 1,620 쌍이 사용되었고, MuSiC-ViT 의 강인한 성능을 보여주기 위해 두 개의 외부 검증 데이터 세트로 검증하였다. MuSiC-ViT 는 내부 검증 데이터 세트의 경우 0.728 및 0.797, 첫 번째 외부 검증 데이터 세트의 경우 0.614 및 0.784, 두 번째 외부 검증 데이터 세트의 경우 각각 0.745 및 0.858 의 수신기 작동 특성 곡선 아래의 정확도와 면적을 달성하였다. 결론적으로 방사선과 전문의의 주요 업무 중 하나인 기준선과 후속 CXR

을 비교하여 변화된 CXR 쌍과 변화 없는 CXR 쌍을 구분할 수 있는 MuSiC-ViT 를 제안한다. AMM 을 추가함으로써 MuSiC-ViT 는 각 CXR 에서 유사한 관심 영역을 비교할 수 있다. 정상/비정상 손실 함수는 모델이 비정상 및 정상 CXR 을 분류하는 데 도움이 될 수 있지만 AMM 을 사용하면 모델이 호흡 변화 수준, 노화 및 자세와 같은 환자의 자연스러운 변화를 무시하면서 병변에 더 집중할 수 있다. 이 아키텍처는 follow-up CXR 연구에 영감을 주고 임상 환경에서 실제 적용으로 이어질 수 있다. 향후 연구에서 우리의 MuSiC-ViT 모델은 후속 CXR 을 기반으로 하는 언어적 설명을 생성할 가능성을 위해 방사선 보고서를 동시에 고려하도록 확장될 수 있으며, 이는 실제 임상 적용에서 방사선 전문의에게 더 유용할 수 있다.