Cardiovascular Disease Detection from Multi-View Chest X-rays with BI-Mamba

Zefan Yang[1], Jiajin Zhang[1], Ge Wang[1],
Mannudeep K. Kalra[2], and Pingkun Yan[1]⋆

[1] Department of Biomedical Engineering and Center for Biotechnology and
Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, USA
[2] Department of Radiology, Massachusetts General Hospital, Harvard Medical
School, Boston, MA, USA

Abstract. Accurate prediction of Cardiovascular disease (CVD) risk in medical imaging is central to effective patient health management. Previous studies have demonstrated that imaging features in computed tomography (CT) can help predict CVD risk. However, CT entails notable radiation exposure, which may result in adverse health effects for patients. In contrast, chest X-ray emits significantly lower levels of radiation, offering a safer option. This rationale motivates our investigation into the feasibility of using chest X-ray for predicting CVD risk. Convolutional Neural Networks (CNNs) and Transformers are two established network architectures for computer-aided diagnosis. However, they struggle to model very high resolution chest X-ray due to the lack of large context modeling power or quadratic time complexity. Inspired by state space sequence models (SSMs), a new class of network architectures with competitive sequence modeling power as Transfomers and linear time complexity, we propose Bidirectional Image Mamba (BI-Mamba) to complement the unidirectional SSMs with opposite directional information. BI-Mamba utilizes parallel forward and backward blocks to encode long-range dependencies of multi-view chest X-rays. We conducted extensive experiments on images from 10,395 subjects in National Lung Screening Trail (NLST). Results show that BI-Mamba outperforms ResNet-50 and ViT-S with comparable parameter size, and saves significant amount of GPU memory during training. Besides, BI-Mamba achieves promising performance compared with previous state of the art in CT, unraveling the potential of chest X-ray for CVD risk prediction.

Keywords: Cardiovascular disease detection · chest X-ray · state space sequence models.

1 Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, representing a significant public health challenge [12]. Accurate prediction of CVD

⋆ Corresponding author: yanp2@rpi.edu

risk using imaging techniques is vital in medical practice. Such predictive capability is key to devising effective health management strategies, enabling early intervention, and potentially saving lives by mitigating disease progression. Previous studies have demonstrated that imaging features in computed tomography (CT) can help predict CVD risk [1,21]. Nonetheless, CT scanning entails notable radiation exposure, which may result in adverse health effects for patients. In contrast, chest X-ray stands as a more accessible imaging technique, emitting considerably lower levels of radiation compared to CT scans, thereby offering a safer option for patients. This rationale underpins our investigation into the feasibility of using chest X-ray for predicting CVD risk.

Recent strides in artificial intelligence, especially deep neural networks, have impressed the research community by their capability to capture effective imaging features for computer-aided diagnosis [26]. Among all deep learning models [22–25], Convolutional Neural Networks (CNNs) [9] and Transformers [20] have become two pivotal network architectures. CNNs (such as U-Net [15] and ResNet [6]) employ shared-weight convolutional kernels to capture local features and is invariant to spatial translation. Transformers (such as ViTs [2] and Swin Transformer [10]) divide an image into a sequence of small patches and model multi-head self-attention between patches, encoding dense long-range dependencies. However, both CNNs and Transformers struggle to model very high resolution images due to the inherent locality of CNNs and the quadratic time complexity with respect to sequence length of Transformers. This problem is profound in chest X-ray disease detection since a typical chest radiographs contains more than a million pixels [7] and previous work [5] shows that downscaling chest radiographs harms disease detection performance. The problem is even more challenging when using multiple views of chest X-ray imaging, which proves to bring performance benefits by previous studies [8, 16, 19]. Hence, an efficient architecture with long-range dependency modeling ability is demanded.

Recently, structured state space sequence models (S4) [4] have emerged as a promising class of network architectures for efficient sequence modeling. Mamba [3] improves S4 with a selective mechanism, allowing the model to selectively forget or propagate information, achieving state-of-the-art on language and genomics modeling. With the diagonally structured state matrix and hardware-aware memory management, Mamba has linear time and memory complexity with respect to input sequence length. Inspired by these merits, recent works have transferred the Mamba network to the vision domain, such as U-Mamba [11] for medical image segmentation and Vision Mamba (Vim) [27] for natural image processing.

In this study, we propose a novel Bidirectional Image Mamba (BI-Mamba) model to efficiently process multi-view high-resolution chest radiographs for CVD risk prediction. To enrich the unidirectional Mamba with opposite directional information, BI-Mamba utilizes parallel forward and backward blocks operating in recurrent mode. This structure captures extensive long-range information into a classification token for CVD risk prediction. BI-Mamba sets it apart from Transformers by its linear scaling with respect to image size, in con-
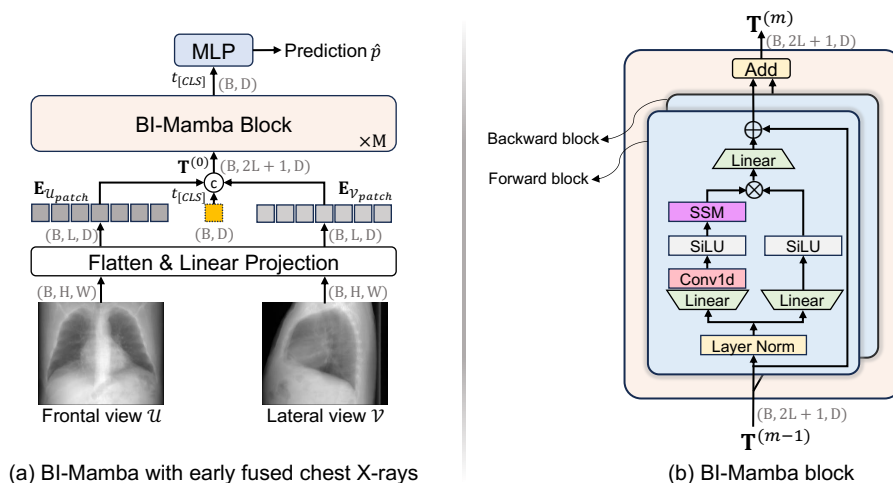
Fig. 1: Network architecture of the BI-Mamba model.

trast to Transformer's quadratic complexity. We use data from 10,395 subjects in the National Lung Screening Trail (NLST) [18] dataset to test BI-Mamba performance for CVD risk prediction in chest X-ray. Experimental results show that BI-Mamba achieves an Area under the Receiver Operating Characteristic Curve (AUROC) of 0.8243, outperforming ViT-S and ResNet-50 with comparable parameter size. BI-Mamba shows exciting memory efficiency with at least 30.5% less allocated memory than ResNet-50 and ViT-S. Overall, our method shows promising performance compared with previous state of the art (0.8710) in CT [1], unraveling the potential to use chest X-ray for CVD risk prediction.

2   Method

The proposed model for CVD risk prediction in chest X-ray consists of an input layer, a novel BI-Mamba model, and a multilayer perceptron (MLP) for output (Fig.1(a)). This section first describes the preliminaries of the state space models that are used for sequence modeling in BI-Mamba. Then, we elaborate the specific operations inside the BI-Mamba block and the details of the early fusion strategy. Last, we analyze the computational complexity of the proposed model.

2.1   Bidirectional Image Mamba (BI-Mamba)

Preliminaries State space models represent dynamic physical systems by means of ordinary differential equations parameterized by $(\mathbf{A}, \mathbf{B}, \mathbf{C})$. Mamba [3] uses a discrete form of the continuous state space model. The discretization is achieved by transforming differential integral to summation across a time step $\boldsymbol{\Delta}$. With this discretization, the state space model can be written in a recurrent form,

where the output $y \in \mathbb{R}$ at t is determined by the hidden state $h \in \mathbb{R}^N$ at $t-1$ and the input $x \in \mathbb{R}$ at t:

$$\begin{aligned} h(t) &= \overline{\mathbf{A}}h(t-1) + \overline{\mathbf{B}}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{1}$$

where $\overline{\mathbf{A}} \in \mathbb{R}^{N \times N} = f_{\mathbf{A}}(\mathbf{\Delta}, \mathbf{A})$, $\overline{\mathbf{B}} \in \mathbb{R}^{N \times 1} = f_{\mathbf{B}}(\mathbf{\Delta}, \mathbf{A}, \mathbf{B})$, and $\mathbf{C} \in \mathbb{R}^{1 \times N}$ are state matrices. The pair $(f_{\mathbf{A}}, f_{\mathbf{B}})$ is a discretization rule. The above equation represents the core operation inside the SSM module in Fig.1(b).

BI-Mamba Block BI-Mamba belongs to the family of state space sequence models (SSMs) that receive token sequences as input. Given a gray-scale medical image $\mathcal{U} \in \mathbb{R}^{H \times W}$ as input, it has to first rearrange into a sequence of small patches $\mathcal{U}_{patch} \in \mathbb{R}^{J \times (P \times P)}$, where (H, W) is the image size, J is the number of patches, and (P, P) is the patch size. A linear projection layer then maps $\mathcal{U}_{patch}$ to patch embeddings $\mathbf{E}_{\mathcal{U}_{patch}} \in \mathbb{R}^{J \times D}$, where D is the embedding dimension. A learnable classification token $\mathbf{t}_{[CLS]}$ is then initialized and placed at the middle of $\mathbf{E}_{\mathcal{U}_{patch}}$. Different from the position-agnostic $\mathbf{t}_{[CLS]}$ in ViTs [2], a middle $\mathbf{t}_{[CLS]}$ proves to be better than left-ended one in Mamba [27]. The input to BI-Mamba is denoted by:

$$\mathbf{T}^{(0)} = [\mathbf{E}_{\mathcal{U}_{patch}}^{(1)}, ..., \mathbf{t}_{[CLS]}, ..., \mathbf{E}_{\mathcal{U}_{patch}}^{(J)}] + \mathbf{E}_{pos}, \tag{2}$$

where $\mathbf{T}^{(0)}, \mathbf{E}_{pos} \in \mathbb{R}^{L \times D}$ are the input sequence and positional embeddings and $L = J + 1$ is the sequence length.

A BI-Mamba block (Fig.1(b)) consists of parallel forward and backward blocks that encode the input sequence in opposite directions, producing representations conditioning on both left and right context. Inside the forward or backward block, $\mathbf{T}^{(m-1)}$ is first normalized and projected to two variables $x, z \in \mathbb{R}^{L \times E}$ in expanded state dimension E. E is multiple times larger than D. x is the input to the SSM module, representing the input sequence. z is the variable for gating dependent on the input sequence.

$$\begin{aligned} \mathbf{T}'^{(m-1)} &= \mathbf{Norm}(\mathbf{T}^{(m-1)}), \\ x &= \mathbf{Linear}^x(\mathbf{T}'^{(m-1)}), \\ z &= \mathbf{Linear}^z(\mathbf{T}'^{(m-1)}). \end{aligned} \tag{3}$$

The projected input sequence x then passes through a 1D convolution and activation unit $x'_d = \mathbf{SiLU}(\mathbf{Conv1d}_d(x))$ to create non-linearity, where $x'_d \in \mathbb{R}^{L \times E}$. The subscript $d \in \{forward, backward\}$ is used to indicate features or operation units specific to a directional block.

A key property of Mamba is that the SSM parameters are input-dependent, as apposed to time- (*i.e.*, patch position) and input-invariant, making it powerful to model images with large appearance shift. BI-Mamba makes it even more powerful by letting the SSM parameters $(\mathbf{\Delta}_d, \mathbf{A}_d, \mathbf{B}_d, \mathbf{C}_d)$ condition on context

direction. The SSM parameters are generated by the linear projection of x'_d and initialized parameters. Discretizing $(\mathbf{A}_d, \mathbf{B}_d)$ given $\boldsymbol{\Delta}_d$ to obtain the form in Eq. (1) is achieved by a simple multiplication discretization rule $(f_\mathbf{A}, f_\mathbf{B})$.

$$\overline{\mathbf{A}}_d = \boldsymbol{\Delta}_d \otimes \mathbf{A}_d, \overline{\mathbf{B}}_d = \boldsymbol{\Delta}_d \otimes \mathbf{B}_d, \tag{4}$$

where $\boldsymbol{\Delta}_d \in \mathbb{R}^{L \times E}$, $\overline{\mathbf{A}}_d, \overline{\mathbf{B}}_d \in \mathbb{R}^{L \times E \times N}$, and $\mathbf{C}_d \in \mathbb{R}^{L \times N}$. N is the latent dimension in SSM, typically set to a small value (*e.g.* 16). The SSM parameterized by $(\overline{\mathbf{A}}_d, \overline{\mathbf{B}}_d, \mathbf{C}_d)$ recurrently processes the input variable x'_d along the forward or backward direction, producing the output sequence $y_d \in \mathbb{R}^{L \times E}$, which is then filtered by z via element-wise product. Then, the $\mathbf{Linear^T}$ maps y_d back to the embedding dimension D.

$$
\begin{aligned}
y_d &= \mathbf{SSM}_d(x'_d) \otimes \mathbf{SiLU}(z), \\
\mathbf{T}_d^{(m)} &= \mathbf{Linear^T}(y_d) + \mathbf{T}^{(m-1)}.
\end{aligned}
\tag{5}
$$

The final output $\mathbf{T}^{(m)}$ is the summation of $\mathbf{T}_{forward}^{(m)}$ and $\mathbf{T}_{backward}^{(m)}$, and hence jointly encapsulates both left-to-right and right-to-left dependencies. The classification token from the output of the last BI-Mamba block $\mathbf{T}^{(M)}$ is projected by a MLP and sigmoid function to produce the prediction $\hat{p}$, which is then supervised by a cross-enropy loss.

2.2   BI-Mamba for Multi-View Image Analysis

Previous studies have applied different fusion strategies to integrate multi-view chest X-rays for disease detection, such as middle fusion [19] and late fusion [8, 16]. Inspired by the efficient linear complexity in sequence length of BI-Mamba, we propose a novel early input patch concatenation strategy for multi-view combination. This early fusion manner allows BI-Mamba to sufficiently model the synergy between chest X-ray views. The input sequence $\mathbf{T}^{(0)}$ is rewritten as:

$$\mathbf{T}^{(0)} = [\mathbf{E}_{\mathcal{U}_{patch}}^{(1)}, ..., \mathbf{E}_{\mathcal{U}_{patch}}^{(L)}, \mathbf{t}_{[CLS]}, \mathbf{E}_{\mathcal{V}_{patch}}^{(1)}, ..., \mathbf{E}_{\mathcal{V}_{patch}}^{(L)}] + \mathbf{E}_{pos}, \tag{6}$$

where $\mathbf{E}_{\mathcal{U}_{patch}}, \mathbf{E}_{\mathcal{V}_{patch}}$ denote the patch embeddings of the frontal view $\mathcal{U}$ and lateral view $\mathcal{V}$ respectively and $\mathbf{T}^{(0)} \in \mathbb{R}^{(2L+1) \times D}$.

2.3   Computational complexity

The state space model independently maps each channel of the input $x \in \mathbb{R}^{L \times E}$ to the output $y \in \mathbb{R}^{L \times E}$ through a higher dimensional state $h \in \mathbb{R}^N$. This results in EN number of effective states per input. With $\mathbf{A} \in \mathbb{R}^{N \times N}$ representing by diagonal structure [4], computing them over an input of length L requires $O(LEN)$ time. In contrast, self-attention in ViTs has a time complexity of $O(L^2D)$. Since self-attention has quadratic scaling to sequence length, it is much more inefficient when handling long sequences. Following [27], we set the number of blocks M to 24, embedding dimension D to 384, expanded state dimension E to 768, and SSM latent dimension to 16.

3   Experiments and Results

3.1   Datasets

This study uses the National Lung Screening Trail (NLST) dataset [18] in experiments. We obtained the same low dose CT scans from NLST and the same CVD risk labels as the state-of-the-art method for CVD prediction in CT [1]. We then simulated chest radiographs via parallel projection through CT volumes [13] to create chest X-ray images for direct performance comparison with the model on CT images. We simulated chest radiographs in both frontal and lateral views.[3]

   In total, the dataset has 33,413 low dose CT scans from 10,395 subjects. From each CT scan, we generated frontal and lateral chest radiographs, resulting in 66,826 images. Subjects are categorized into high and low CVD risk groups according to their exam reports and causes of death for deceased subjects. Specifically, a subject is deemed having high CVD risk when there is cardiovascular abnormality reported in the exam report or the subject died of CVD according to the ICD-9 code. A subject is considered normal when the subject has no CVD-related medical history and no reported cardiovascular abnormality in CT, and did not die of circulatory system diseases. The numbers of subjects with high and low CVD risk are 2,962 and 7,433 respectively. The chest X-ray data is divided into training (70%, 7,268 subjects), validation (10%, 1,042 subjects), and test splits (20%, 2,085 subjects) for model development and evaluation.

3.2   Experimental details

We set the model input chest X-ray size to 512 by default. For results in Fig.2(a), we vary image size to 448, 384, and 224. For sequence models (ViT-S and BI-Mamba), we use a patch size 16. During model training, we apply random crop (aspect ratio ranging from (0.75, 1.3)) and horizontal flip as image augmentation. We initialized our BI-Mamba model with the weights released by the Vision Mamba team [27]. ViTs were initialized with weights from DINOv2 [14]. ResNet-50 was initialized with weights pretained on ImageNet. We set the initial learning rate to 5e-6, weight decay to 1e-8, and batch size to 24. We used the AdamW optimizer and cosine learning rate scheduler. The number of training epochs is 30. Models are trained on a NVIDIA DGX-1 server with 8x A100 GPUs. At the inference time, we compute the area under the receiver operating characteristic (AUROC) curve in the test split to evaluate classification performance. Source code is publicly available at: https://github.com/RPIDIAL/BI-Mamba.

3.3   Results of CVD Risk Prediction

We compared our BI-Mamba model with other methods using different network architectures and view combination strategies (Table 1). The benchmarking networks include ResNet-50 and ViT-S. ResNet-50 combines features from two chest

---

[3] Details of chest X-ray simulation are shown in Supplementary Material.

Table 1: Results of different network architectures and view combination strategies. p-value is computed with DeLong's algorithm [17] that compares the statistical difference between AUROC values.

Architecture	View Combination	# params	AUROC	p-value
ResNet-50	Output concat. [16]	25M	0.7900	2e-5
	Cross Attention [19]	29M	0.7914	4e-5
ViT-S	`[CLS]` token concat.	21M	0.7877	4e-5
	Input patch concat.	21M	0.7597	5e-8
BI-Mamba	`[CLS]` token concat.	25M	0.8106	0.1297
	Input patch concat.	25M	**0.8243**	-

Table 2: Risk prediction results using each view individually.

Architecture	# params	embed dim	init. weights	AUROC	
				Frontal	Lateral
ResNet-50	25M	-	sup. pretrain.	0.7831	0.7800
ViT-S	21M	384	DINOv2 pretrain.	0.7712	0.7680
ViT-B	86M	768	DINOv2 pretrain.	0.7649	0.7660
ViT-L	300M	1024	DINOv2 pretrain.	0.7770	0.7816
BI-Mamba	25M	384	sup. pretrain.	**0.7924**	**0.7946**

X-ray views after the global average pooling in each branch [16]. The ResNet-50 with cross attention [19] includes a cross attention block after the third block of ResNet-50. ViT-S and BI-Mamba both use a `[CLS]` token as their final output. They combine the `[CLS]` tokens from the last layer outputs of the two views for prediction. ViT-S and BI-Mamba with input patch concatenation directly combine two patch sequences at the input layer and have only one `[CLS]` token. ViT-S conventionally places the `[CLS]` token at the left-end. BI-Mamba puts it at the middle between the two input patch sequences. Table 1 shows that BI-Mamba with input patch concatenation attains a 0.8243 AUROC and outperforms all the other networks and combination strategies. Under the same combination strategy, ViT-S attains a 0.7597 AUROC, significantly inferior to BI-Mamba's performance ($p < 0.05$). This result can be attributed to their different mechanisms for sequence modeling. A middle `[CLS]` in BI-Mamba independently aggregate information from one view in recurrent propagation. Conversely, ViT-S uses self-attention, which may be deficient to model sequences from two views, since there are not always strong correspondences to attend to between the query from one view and keys and values from the other view.

Besides, we compared BI-Mamba with the state of the art method Tri-2D network [1] for CVD risk prediction. Tri-2D attains a 0.8710 AUROC in low-dose CT. BI-Mamba presents a promising result of 0.8243 in chest X-rays.
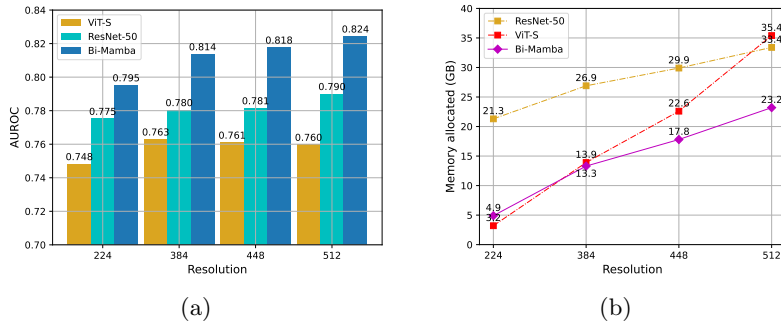
Fig. 2: Results on different image resolution and comparison of memory footprint at training time.

3.4   Ablation studies

Results on different image resolution To investigate whether image resolution is a bottleneck of CVD risk prediction in chest X-ray, we gradually increase image size from 224 to 512. We apply output concatenation to ResNet-50 and input patch concatenation to ViT-S and BI-Mamba in experiments. As shown in Fig. 2(a), increasing image size steadily improves BI-Mamba's performance, leading to an AUROC margin of 2.9% between image size 512 and 224. We expect the tendency to extrapolate with higher image resolution, which we will validate in our future work.

Analysis of computational efficiency Computational efficiency is an inevitable factor to consider when input size scales. BI-Mamba and ViTs have computational complexities of $O(L^2D)$ and $O(LEN)$ respectively. ViTs have quadratic time complexity with respect to input length, while BI-Mamba has linear time complexity, making it more efficient with high-definition inputs. This is evidenced by 34.5% less allocated memory for BI-Mamba at image resolution 512 (Fig.2(b)).

Risk prediction results using individual views This section provides single-view results to benchmark the benefits of multi-view chest radiographs for CVD risk prediction. We load ImageNet supervised pretraining weights for ResNet-50 and BI-Mamba, and perform fine-tuning. We load DINOv2 general purpose features for ViT-S, ViT-B, and ViT-L, and perform linear probing. Results are shown in Table 2

4   Conclusions

We propose a novel Bidirectional Image Mamba (BI-Mamba) model for cardiovascular disease (CVD) risk prediction in multi-view chest X-rays. BI-Mamba

contains parallel forward and backward blocks in recurrent mode to encode rich long-range dependencies of two input concatenated chest X-rays. BI-Mamba shows superior modeling power than ViT-S and ResNet-50, while it only has linear time complexity and saves at least 30.5% training-time GPU memory. Compared with previous state of the art in CT (0.8710), BI-Mamba attains a promising 0.8243 AUROC, unlocking the gate towards a low-dose, inexpensive, and accurate computed-aided CVD risk prediction technique in chest X-ray.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chao, H., Shan, H., Homayounieh, F., Singh, R., Khera, R.D., Guo, H., Su, T., Wang, G., Kalra, M.K., Yan, P.: Deep learning predicts cardiovascular disease risks from lung cancer screening low dose computed tomography. Nature Communications **12**(1), 2963 (2021)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
4. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
5. Haque, M.I.U., Dubey, A.K., Danciu, I., Justice, A.C., Ovchinnikova, O.S., Hinkle, J.D.: Effect of image resolution on automated classification of chest x-rays. Journal of Medical Imaging **10**(4), 044503–044503 (2023)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
8. Kim, D.: Chexfusion: Effective fusion of multi-view features using transformers for long-tailed chest x-ray classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2702–2710 (2023)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

11. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
12. Martin, S.S., Aday, A.W., Almarzooq, Z.I., Anderson, C.A., Arora, P., Avery, C.L., Baker-Smith, C.M., Barone Gibbs, B., Beaton, A.Z., Boehme, A.K., et al.: 2024 heart disease and stroke statistics: A report of us and global data from the american heart association. Circulation **149**(8), e347–e913 (2024)
13. Moturu, A., Chang, A.: Creation of synthetic x-rays to train a neural network to detect lung cancer. Journal Beyond Sciences Initiative, University of Toronto, in Toronto (2018)
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
16. Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., Xu-Wilson, M.: Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. arXiv preprint arXiv:1804.07839 (2018)
17. Sun, X., Xu, W.: Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Processing Letters **21**(11), 1389–1393 (2014)
18. Team, N.L.S.T.R.: Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine **365**(5), 395–409 (2011)
19. van Tulder, G., Tong, Y., Marchiori, E.: Multi-view analysis of unregistered medical images using cross-view transformers. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 104–113. Springer (2021)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
21. Zeleznik, R., Foldyna, B., Eslami, P., Weiss, J., Alexander, I., Taron, J., Parmar, C., Alvi, R.M., Banerji, D., Uno, M., et al.: Deep convolutional neural networks to predict cardiovascular risk from computed tomography. Nature communications **12**(1), 715 (2021)
22. Zhang, J., Chao, H., Dasegowda, G., Wang, G., Kalra, M.K., Yan, P.: Revisiting the trustworthiness of saliency methods in radiology ai. Radiology: Artificial Intelligence **6**(1), e220221 (2023)
23. Zhang, J., Chao, H., Dhurandhar, A., Chen, P.Y., Tajer, A., Xu, Y., Yan, P.: Spectral adversarial mixup for few-shot unsupervised domain adaptation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 728–738. Springer (2023)
24. Zhang, J., Chao, H., Dhurandhar, A., Chen, P.Y., Tajer, A., Xu, Y., Yan, P.: When neural networks fail to generalize? a model sensitivity perspective. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 11219–11227 (2023)
25. Zhang, J., Chao, H., Yan, P.: Toward adversarial robustness in unlabeled target domains. IEEE Transactions on Image Processing **32**, 1272–1284 (2023)
26. Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Disease-informed adaptation of vision-language models. arXiv preprint arXiv:2405.15728 (2024)

27. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)