

XFMamba: Cross-Fusion Mamba for Multi-View Medical Image Classification

Xiaoyu Zheng^{1*}, Xu Chen¹, Shaogang Gong², Xavier Griffin³, and Greg Slabaugh^{1**}

¹ Digital Environment Research Institute (DERI), Queen Mary University of London, London, UK

² School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

³ Blizzard Institute - Faculty of Medicine and Dentistry, Queen Mary University of London, London, UK

Abstract. Compared to single view medical image classification, using multiple views can significantly enhance predictive accuracy as it can account for the complementarity of each view while leveraging correlations between views. Existing multi-view approaches typically employ separate convolutional or transformer branches combined with simplistic feature fusion strategies. However, these approaches inadvertently disregard essential cross-view correlations, leading to suboptimal classification performance, and suffer from challenges with limited receptive field (CNNs) or quadratic computational complexity (transformers). Inspired by state space sequence models, we propose *XFMamba*, a pure Mamba-based cross-fusion architecture to address the challenge of multi-view medical image classification. XFMamba introduces a novel two-stage fusion strategy, facilitating the learning of single-view features and their cross-view disparity. This mechanism captures spatially long-range dependencies in each view while enhancing seamless information transfer between views. Results on three public datasets, MURA, CheXpert and DDSM, illustrate the effectiveness of our approach across diverse multi-view medical image classification tasks, showing that it outperforms existing convolution-based and transformer-based multi-view methods. Code is available at <https://github.com/XZheng0427/XFMamba>.

Keywords: Multi-view · Cross fusion · Medical image classification · State space models.

1 Introduction

In many clinical applications, such as the detection of orthopedic fractures, radiologists analyze multiple views within a diagnostic examination. Typically each view provides complementary insights, but their correlation can also be leveraged to improve predictive accuracy. Numerous recent studies aim to develop

* Corresponding author: xiaoyu.zheng@qmul.ac.uk

** Corresponding author: g.slabaugh@qmul.ac.uk

multi-view networks inspired by the multi-view analysis performed by radiologists, leading to the creation of more robust and performant models [1, 2].

Previous approaches in multi-view image analysis rely on Convolutional Neural Networks (CNN) or Transformers. While CNN-based approaches are recognized for their scalability and linear computational complexity in the number of image pixels, their reliance on local receptive fields introduces a bias toward local feature extraction, limiting their ability to directly capture long-range dependencies [3, 4]. Moreover, CNNs often employ a weight-sharing mechanism across different view features of the input image, limiting their flexibility in adapting to unseen or low-quality medical images. To address the limited context understanding of CNNs, methods such as CVT [5] and MV-HFMD [25] utilize CNNs as the backbone for local feature extraction while incorporating cross-view attention mechanisms to transfer features between unregistered views. With the emergence of Transformers, transformer-based methods such as T-MVF [7] enhance visual modeling capabilities by leveraging global view features and dynamically adaptable weights. However, the attention mechanism in transformer-based methods results in quadratic computational complexity relative to input size and limitations on token size, posing substantial challenges to efficiency [8]. Building on the architecture of vision transformers, multi-view Swin Transformer approaches such as MV-Swin-T [6] and WT-MVSNet [9] aim to improve efficiency by reducing the dimensions or strides of processing windows. However, this optimization comes at the cost of limiting cross-view feature interactions.

To address the aforementioned limitations, Selective Structured State Space Models, also referred to as Mamba [8], have garnered increasing attention for their ability to selectively forget or propagate information, achieve global feature coverage, and utilize dynamic weights with linear computational complexity. Mamba has demonstrated exceptional performance in natural language processing [8] and computer vision tasks [11, 17]. Furthermore, extensive research has investigated its potential in medical imaging, including image segmentation [10] and image classification [12]. Multi-view BI-Mamba [26] employs an early fusion mechanism to capture and integrate multi-view information. However, these recent studies incorporate Mamba as a plug-and-play module without a task-specific, in-depth design for multi-view medical imaging. Additionally, the exploration of Mamba in multi-view medical imaging tasks for cross-image fusion remains limited, particularly in architectural adaptations and fusion strategies.

Inspired by these advantages, we propose XFMamba, a pure Mamba network for multi-view cross-fusion, tailored to address the challenges of multi-view unregistered medical image classification. XFMamba integrates a four-stage encoder and two-stage fusion modules. The encoder effectively captures multi-scale features across multiple views. Furthermore, the two-stage fusion module is designed to exchange and align cross-view information and enhance multi-view integration. Comprehensive experiments conducted on the MURA musculoskeletal radiographs dataset [13], the CheXpert chest X-ray dataset [14], and the CBIS-DDSM mammography dataset [15] demonstrate that XFMamba outperforms

state-of-the-art models across three varied clinical problems. The key contributions of this work can be summarized as follows:

1. Our cross-fusion work marks the successful application of state space models, specifically Mamba, in multi-view medical image classification.
2. We propose a two-stage, Mamba-based fusion mechanism designed to efficiently extract and seamlessly integrate information across multiple views.
3. Comprehensive evaluations across three varied datasets demonstrate the superior accuracy and efficiency of our method, establishing a new benchmark for Mamba’s potential in the multi-view medical imaging domain.

2 Methods

2.1 Preliminaries

State Space Models State Space Models (SSMs) [16] constitute a category of sequence-to-sequence modeling frameworks distinguished by their time-invariant dynamics, formally known as linear time-invariant (LTI) systems. SSMs, with their linear computational complexity, efficiently capture underlying dynamics by implicitly mapping to latent states, formally expressed as:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), y(t) = \mathbf{C}h(t) + \mathbf{D}x(t). \quad (1)$$

Where, $x(t) \in \mathbb{R}$ denotes the input, $h(t) \in \mathbb{R}^N$ denotes the hidden state, and $y(t) \in \mathbb{R}$ denotes the output. $h'(t)$ refers to the time derivative of $h(t)$. Additionally, $\mathbf{A} \in \mathbb{R}^{N \times N}$ refers to the state matrix. $\mathbf{B} \in \mathbb{R}^{N \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times N}$ and $\mathbf{D} \in \mathbb{R}$ denote the projection matrices.

Mamba [8] employs the zero-order hold (ZOH) discretization method to convert ordinary differential equations (ODEs) into discrete functions, making it particularly well-suited for deep learning applications. The discretization process applies a time step Δ to transform the continuous state parameters \mathbf{A} and \mathbf{B} into discrete parameters $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, which can be defined as:

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}), \overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1} \exp(\Delta\mathbf{A} - I) \cdot \Delta\mathbf{B}. \quad (2)$$

Then, Eq. (1) takes the following form:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, y_t = \mathbf{C}h_t + \mathbf{D}x_t \quad (3)$$

Eq. (3) represents the fundamental operation within the SSM module.

2.2 Proposed Architecture

As shown in Fig. 1, the proposed approach consists of a four-stage multi-scale encoder and a two-stage cascade fusion module. For each input image, the encoder is comprised of four visual state space modules (VSSMs), incorporating down-sampling operations and sequentially cascaded to extract hierarchical image features across multiple levels. Next, features extracted from the two branches are

processed and integrated through a two-stage cascade fusion module, which consists of a Cross-View Swapping Mamba (CVSM) block for local feature cross-view correlation (shallow fusion) and a Multi-View Combination Mamba (MVCM) block for holistic feature multi-view correlation (deep fusion). Finally, the refined and merged features are fed into a classifier to produce the prediction.

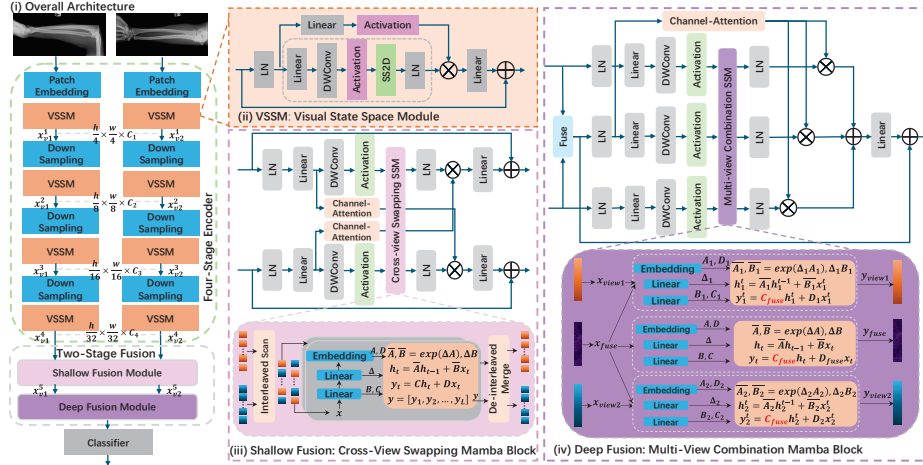


Fig. 1. XFMamba architecture. (i) The overall architecture is composed of a four-stage encoder and two-stage fusion module. (ii) Visual State Space Module (VSSM) for feature extraction. (iii) Cross-view swapping Mamba (CVSM) block for shallow fusion. (iv) Multi-view combination Mamba (MVCM) block for deep fusion.

XFMamba Encoder. The multi-scale encoder consists of two four-stage feature extraction branches that process two greyscale images with different views as input, defined as $X_{v1}, X_{v2} \in \mathbb{R}^{H \times W \times 1}$. The multi-scale encoder adopts VMamba as the backbone [17]. Inspired by ViT [18], it starts by partitioning the input into patches to produce feature maps, i.e., $X_{v1,v2}^1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_1}$. Next, the features are progressively processed through three stages of downsampling and VSSM, as shown in Fig. 1(i) and (ii), resulting in multi-scale feature representations, i.e., $X_{v1,v2}^2 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_2}$, $X_{v1,v2}^3 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_3}$, and $X_{v1,v2}^4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$. We implement the VSSM using Selective Scan 2D (SS2D) modules proposed by VMamba [17]. As illustrated in Fig. 1(ii), the features are processed through a sequence comprising linear projection (Linear), depth-wise convolution (DWConv), and the Sigmoid Linear Unit (SiLU) activation function. Subsequently, an SS2D module with a multiplicative connection is utilized to capture spatially long-range (wider context) information.

Shallow Fusion: Cross-View Swapping Mamba. As illustrated in Fig. 1(iii), we introduce a CVSM block that leverages an interleaving mechanism to enhance cross-view features by incorporating information from the other view. The two input features are initially processed through linear layers and depth-wise convolutions, followed by SiLU activation, and are then passed into the cross-view interleaving SS2D. Specifically, the cross-view interleaving SS2D mechanism is firstly applied on X_{v1}^4 and X_{v2}^4 , facilitating efficient feature interleaving across multiple views without incurring additional computational overhead, as shown in Eq. 4.

$$M(C) = \begin{cases} 1, & \text{if } C \bmod 2 = 0, \\ 0, & \text{if } C \bmod 2 = 1. \end{cases} \quad \tilde{X}_{v1/v2}^5[:, C] = \begin{cases} \tilde{X}_{v2/v1}^4[:, C], & \text{if } M(C) = 1, \\ \tilde{X}_{v1/v2}^4[:, C], & \text{if } M(C) = 0. \end{cases} \quad (4)$$

Here, $\tilde{X}_{v1/v2}^5, \tilde{X}_{v1/v2}^4 \in \mathbb{R}^{B \times N \times C}$ ($N = H \times W$) represent the flattened features. $M(C)$ defines the channel-wise mask, comprising binary values where 0 signifies no interleaving and 1 denotes interleaving. Subsequently, the interleaved cross-view features are processed through their corresponding Mamba blocks to extract long-range cross-view spatial information. The resulting sequences are then reshaped and merged to generate the output scanned features using the de-interleaving merge mechanism. This results in the features X_{v1}^5 and X_{v2}^5 from both views. We adopt a shared channel-attention mechanism, i.e., Squeeze-and-Excitation (SE) [20], to reweight their channels in a cross-view manner and adaptively learn the inter-model relationships. Crucially, cross-multiplication is employed to process two channels instead of merely scaling each feature map by its own weights. Specifically, the attention derived from the global context of v_1 is used to scale the channels of v_2 's feature map, and vice versa.

Deep Fusion: Multi-View Combination Mamba. In the previous CVSM block, features from two views interleave to enhance each other, capturing localized cross-view interactions. However, to generate a more holistic representation that integrates critical information from both views at a deeper level, we introduce the MVCM block. Unlike CVSM, which focuses on localized feature interleaving, MVCM performs a more comprehensive fusion by aggregating and refining the outputs of the CVSM block. This transition from shallow fusion (local feature mixing) to deep fusion (global integration) ensures that the final representation effectively captures higher-level fused features from both views, as illustrated in Fig. 1(iv). Specifically, the outputs X_{v1}^5 and X_{v2}^5 from the CVSM block are first fused as one of the inputs. Additionally, two separate view branches are introduced, each processing features from different views as input. Subsequently, the three branches are processed through linear layers, depth-wise convolution layers, and SiLU activation before being fed into the multi-view combination SSM. Within the multi-view combination SSM, the system matrix \mathbf{C}_{fuse} information is exchanged across multiple selective scan modules to enhance multi-view feature fusion, inspired by [22, 23]. The Mamba selection mechanism generates the system matrices \mathbf{B} , \mathbf{C} and Δ from linear projection

layers, facilitating the context-awareness capability of the model. As specified in Eq 3, the matrix \mathbf{C} decodes information from the hidden state h_t to produce the output y_t . Therefore, we employ the matrix \mathbf{C}_{fuse} of fused-branch to decode two separate view branches enhancing the model’s multi-view context-awareness ability. The modified Eq. 3 is applicable to the multi-view combination SSM and is defined as follows:

$$\begin{aligned} y_{v1}^t &= \mathbf{C}_{fuse} h_{v1}^t + \mathbf{D}_{v1} x_{v1}^t, & y_{v1} &= [y_{v1}^1, y_{v1}^2, \dots, y_{v1}^n], \\ y_{v2}^t &= \mathbf{C}_{fuse} h_{v2}^t + \mathbf{D}_{v2} x_{v2}^t, & y_{v2} &= [y_{v2}^1, y_{v2}^2, \dots, y_{v2}^n], \\ y_{fuse}^t &= \mathbf{C}_{fuse} h_{fuse}^t + \mathbf{D}_{fuse} x_{fuse}^t, & y_{fuse} &= [y_{fuse}^1, y_{fuse}^2, \dots, y_{fuse}^n]. \end{aligned} \quad (5)$$

Where, $x_{v1/v2/fuse}^t$ defines as the input at time step t , while $y_{v1/v2/fuse}$ represents the output of the multi-view combination SSM. After the multi-view combination SSM process, we propose a fuse-multiplication mechanism to compute channel attention on the normalized original fused view-specific features, adaptively learning the fused-view channel relationships to enhance the feature representation of a single view. Finally, additive fusion is applied to combine the three feature branches, resulting in a final holistic fused feature.

3 Experiments

3.1 Datasets and Evaluation Metrics

Datasets. To evaluate the effectiveness of our approach, we conducted experiments on the MURA [13], CheXpert [14], and CBIS-DDSM [15] datasets. The **MURA** dataset is the musculoskeletal abnormality detection dataset, which is a binary classification task and contains two view X-ray images for each patient. We have selected patients with two views and divided the patients into random subsets for training (35,185), validation (3,158) and testing (3,283). The **CheXpert** dataset comprises chest X-ray images annotated for 13 distinct observations with positive and negative labels. We selected patients that included both frontal and lateral views and randomly divided the samples into subsets for training (23,559), validation (3,926) and testing (3,928). The **CBIS-DDSM** is a mammography dataset with craniocaudal (CC) and mediolateral-oblique (MLO) views, which is a binary classification to predict benign versus malignant cases for each CC/MLO pair. The training, validation, and test splits comprise 857, 215, and 274 CC/MLO pair samples, respectively. We applied the cropped method for CBIS-DDSM, outlined by [24], adopting thresholding to position a fixed-size cropping window that captures the breast while minimizing the inclusion of an empty background. For all three datasets, we resized the images to 224×224 prior to inputting them into the models.

3.2 Experimental Details

Our XFMamba approach was implemented using PyTorch and trained on an NVIDIA A100 GPU. We adopted the Adam optimizer with the initial learning

rate $1e^{-4}$ and weight decay $1e^{-5}$ to update the model parameters. The model was trained for 100 epochs with a batch size of 16. We initialized our XFMamba model with the pre-trained weights provided by VMamba [17] for the XFMamba feature encoder, constructing three variants of the model with different sizes: tiny (XFMamba-T), small (XFMamba-S), and base (XFMamba-B).

3.3 Comparison with State-of-the-Art Methods

To evaluate the performance, we compared the AUROC values of our XFMamba method with CVT [5], MVC-NET [2], MV-Swin-T [6], MV-HFMD [25], and BI-Mamba [26]. We follow the experimental protocol [5], which is optimized for the cross-entropy loss and repeated over four training runs. The quantitative results are presented in Table 1, showcasing the classification performance along with a comparison of model sizes across the three datasets. Our tiny model outperforms other methods on the MURA and CheXpert datasets while maintaining the second-fewest number of parameters. Additionally, our small and base methods achieve the highest AUROC values across all three datasets. This superiority arises from the capability of our proposed XFMamba model to efficiently extract features from each individual view and while leveraging cross-view features.

Table 1. AUROC and model size (millions of parameters) comparisons on MURA (2 classes), CheXpert (13 classes), and CBIS-DDSM (2 classes) datasets. The mean and standard deviation of AUROC over four training runs.

Method	Backbone	Params(M)	MURA	CheXpert	CBIS-DDSM
CVT	Resnet18	39.95	0.876 ± 0.003	0.909 ± 0.002	0.697 ± 0.002
BI-Mamba	ViM-s	37.39	0.878 ± 0.004	0.905 ± 0.002	0.611 ± 0.003
MV-HFMD	ViT-s	54.07	0.886 ± 0.002	0.909 ± 0.003	0.569 ± 0.002
MVC-NET	ResNet26	71.10	0.838 ± 0.003	0.892 ± 0.002	0.702 ± 0.003
MV-Swin-T	Swin-T	58.32	0.711 ± 0.002	0.882 ± 0.005	0.566 ± 0.002
XFMamba	VMamba-t	39.29	0.898 ± 0.002	0.917 ± 0.003	0.664 ± 0.002
XFMamba	VMamba-s	58.73	0.910 ± 0.003	0.918 ± 0.002	0.752 ± 0.004
XFMamba	VMamba-b	90.54	0.904 ± 0.004	0.919 ± 0.002	0.761 ± 0.003

Furthermore, Fig. 2(Left) compares the model computational complexity on the CBIS-DDSM dataset. The results indicate that our tiny model has the lowest FLOPs among the compared approaches while still achieving competitive performance. Fig. 2(right) presents the qualitative results for GradCAM [27] heatmap examples in the CBIS-DDSM dataset against other methods. The white mask of the full mammogram image is the ROI of abnormalities. XFMamba’s heatmaps are more precise and tend to focus on the ROI mask, whereas other models’ heatmaps may highlight irrelevant areas. However, in failed cases, all models struggled to distinguish between abnormal and normal regions due to the limited visibility of the pathological structure.

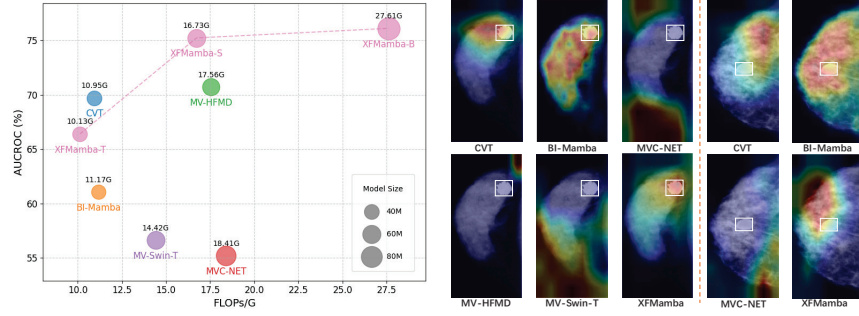


Fig. 2. (Left) Computational complexity comparison on CBIS-DDSM dataset. The size of each circle denotes the model size, i.e., parameters. (Right) Qualitative results for successful cases and failed cases (right of the orange dotted line) using different methods on the CBIS-DDSM dataset.

3.4 Ablation Study

As illustrated in Table 2, we performed an ablation study with XFmamba-S on the CheXpert dataset. Compared to the complete model, eliminating the

Table 2. XFmamba ablation study on CheXpert.

Num.	Views	Fusion Type	CVSM-Block	MVCM-Block	AUROC↑
1	Lateral	no	✗	✗	0.8961
2	Frontal	no	✗	✗	0.9081
3	Frontal & Lateral	early-fusion	✗	✗	0.9090
4	Frontal & Lateral	late-fusion	✗	✗	0.9092
5	Frontal & Lateral	cross-fusion	✓	✗ (concat)	0.9129
6	Frontal & Lateral	cross-fusion	✓	✗ (add)	0.9136
7	Frontal & Lateral	cross-fusion	✗	✓	0.9144
8	Frontal & Lateral	cross-fusion	✓	✓	0.9184

CSVM-Block results in a 0.40% performance reduction. To evaluate the impact of removing the MVCM-Block, we test two simple fusion methods (addition and concatenation) reducing performance by 0.48% and 0.55%, respectively. Eliminating both blocks and employing late fusion for feature integration results in a 0.92% performance decline. Additionally, we compare our proposed cross-fusion model with both early-fusion and late-fusion models utilizing concatenation, finding that XFmamba outperforms them by 0.94% and 0.92%, respectively. To further assess the effectiveness of multiple views, we compare the lateral view, frontal view, and their combination, observing that incorporating multiple views improves performance by 1.03% and 2.17% compared to using only the lateral or frontal view, respectively.

4 Conclusion and Future Work

In this paper, we have proposed a novel XFMamba network for multi-view medical image classification. Our model combines a four-stage encoder and two-stage fusion module, enabling the effective learning of single-view features and their cross-view combination. Our cross-view swapping Mamba block enriches feature across views through cross-view information integration. The multi-view combination Mamba block further enhances multi-view information fusion through a multi-view selective scan mechanism. Comparison experiments and ablation studies demonstrate the effectiveness of our approach. In future work, our XFMamba network can serve as a baseline for multi-modality models that combine different imaging modalities. We also plan to explore clinical applications where multiple views are registered, or exhibit geometric relationships, leveraging our approach to enhance feature alignment and information fusion.

Acknowledgments. This work forms part of the research portfolio of the National Institute for Health Research Barts Biomedical Research Centre (NIHR203330). This work was also supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through Keystone project funding from Responsible AI UK (KP0016). This research utilised Queen Mary’s Andrena HPC facility, supported by QMUL Research-IT.

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

References

1. M. Seeland, P. Mader: Multi-view classification with convolutional neural networks. *PLOS ONE* **16**(1), e0245230 (2021)
2. X. Zhu, Q. Feng: MVC-Net: Multi-view chest radiograph classification network with deep fusion. In: *IEEE 18th International Symposium on Biomedical Imaging (ISBI 2021)*, pp. 554–558. IEEE (2021)
3. K. Karthik, S.S. Kamath: A deep neural network model for content-based medical image retrieval with multi-view classification. *The Visual Computer* **37**(7), 1837–1850 (2021)
4. H. Yu, L.T. Yang, Q. Zhang, D. Armstrong, M. Deen: Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. *Neurocomputing* **444**, 92–110 (2021)
5. G. Van Tulder, Y. Tong, E. Marchiori: Multi-view analysis of unregistered medical images using cross-view transformers. In: M. de Bruijne, P. Cattin, S. Cotin, et al. (eds.) *MICCAI 2021, LNCS*, vol. 12903, pp. 104–113. Springer, Cham (2021)
6. S. Sarker, P. Sarker, G. Bebis, A. Tavakkoli: MV-Swin-T: Mammogram classification with multi-view swin transformer. In: *IEEE International Symposium on Biomedical Imaging (ISBI 2024)*. IEEE (2024)
7. Z. Sun, H. Jiang, L. Ma, Z. Yu, H. Xu: Transformer based multi-view network for mammographic image classification. In: L. Wang, Q. Dou, P.T. Fletcher, S. Speidel, S. Li (eds.) *MICCAI 2022, LNCS*, pp. 46–54. Springer, Cham (2022)

8. A. Gu, T. Dao: Mamba: Linear-time sequence modeling with selective state spaces. In: First Conference on Language Modeling (2024). Available at: <https://openreview.net/forum?id=tEYskw1VY2>
9. J. Liao, Y. Ding, Y. Shavit, et al.: WT-MVSNet: Window-based transformers for multi-view stereo. *Adv. Neural Inform. Process. Syst.* **35**, 8564–8576 (2022)
10. J. Ma, F. Li, B. Wang: U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722* (2024)
11. L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang: Vision Mamba: Efficient visual representation learning with bidirectional state space model. In: *Proc. 41st Int. Conf. Mach. Learn. (ICML 2024)*, vol. 235, Article 2584, pp. 62429–62442. *JMLR.org* (2024)
12. Y. Yue, Z. Li: MedMamba: Vision Mamba for medical image classification. *arXiv preprint arXiv:2403.03849* (2024)
13. P. Rajpurkar, J. Irvin, A. Bagul, et al.: MURA: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957* (2017)
14. J. Irvin, P. Rajpurkar, M. Ko, et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proc. AAAI Conf. Artif. Intell.* **33**(1), 590–597. AAAI Press (2019)
15. R.S. Lee, F. Gimenez, A. Hoogi, et al.: A curated mammography dataset for use in computer-aided detection and diagnosis research. *Scientific Data* **4**(1), 170177 (2017)
16. A. Gu, K. Goel, C. Ré: Efficiently modeling long sequences with structured state spaces. In: *Proc. 9th Int. Conf. Learning Representations (ICLR 2021)*, pp. 1–15. OpenReview (2021)
17. Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, Y. Liu: VMamba: Visual state space model. In: *38th Conf. Neural Inform. Process. Syst. (NeurIPS 2024)*
18. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby: An image is worth 16×16 words: transformers for image recognition at scale. In: *Int. Conf. Learning Representations (ICLR 2021)*
19. Z. Li, H. Pan, K. Zhang, et al.: MambaDFuse: A Mamba-based dual-phase model for multi-modality image fusion. *arXiv preprint arXiv:2404.08406* (2024)
20. J. Hu, L. Shen, G. Sun: Squeeze-and-excitation networks. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR 2018)*, pp. 7132–7141. IEEE (2018)
21. Z. Li, H. Pan, K. Zhang, et al.: MambaDFuse: A Mamba-based dual-phase model for multi-modality image fusion. *arXiv preprint arXiv:2404.08406* (2024)
22. C.F.R. Chen, Q. Fan, R. Panda: CrossViT: Cross-attention multi-scale vision transformer for image classification. In: *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV 2021)*, pp. 357–366. IEEE (2021)
23. Z. Wan, P. Zhang, Y. Wang, et al.: Sigma: Siamese mamba network for multi-modal semantic segmentation. In: *IEEE/CVF Winter Conf. Applications of Computer Vision (WACV 2025)*, pp. 1734–1744. IEEE (2025)
24. N. Wu, J. Phang, J. Park, et al.: Deep neural networks improve radiologists’ performance in breast cancer screening. *IEEE Trans. Med. Imaging* **39**(4), 1184–1194 (2019)
25. S. Black, R. Souvenir: Multi-view classification using hybrid fusion and mutual distillation. In: *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV 2024)*, pp. 270–280. IEEE (2024)

26. Z. Yang, J. Zhang, G. Wang, et al.: Cardiovascular disease detection from multi-view chest X-rays with BI-Mamba. In: M.G. Linguraru, Q. Dou, A. Feragen, et al. (eds.), MICCAI 2024, LNCS, pp. 134–144. Springer, Cham (2024)
27. R.R. Selvaraju, M. Cogswell, A. Das, et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE Int. Conf. Comput. Vis. (ICCV 2017), pp. 618–626. IEEE (2017)