Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Efficient Dual-View Mamba-CNN Framework for Edge-Enabled Intelligent Mammogram Analysis

Kun Xu

202212281015@std.uestc.edu.cn

University of Electronic Science and Technology of China

Pin-Han Ho

University of Electronic Science and Technology of China

Jun Chen

Shenzhen Hospital of Southern Medical University

Limei Peng

University of Electronic Science and Technology of China

Boyang Deng

Future University Hakodate

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11584902/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Efficient Dual-View Mamba-CNN Framework for Edge-Enabled Intelligent Mammogram Analysis

Kun Xu[1,2,3], Pin-Han Ho[2*], Jun Chen[3], Limei Peng[2], Boyang Deng[4]

[1]University of Electronic Science and Technology of China, Chengdu, 611731, Sichuan, China.
[2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, 518000, Guangdong, China.
[3]Shenzhen Hospital of Southern Medical University, Shenzhen, 518100, Guangdong, China.
[4]Future University Hakodate, Hakodate, 0418655, Hokkaido, Japan.

*Corresponding author(s). E-mail(s): p4ho@uwaterloo.ca;

Abstract

Reliable mammogram classification is essential for intelligent breast cancer screening, yet deploying high-performance models in resource-constrained healthcare settings remains challenging. Single-view methods overlook complementary information in craniocaudal (CC) and mediolateral oblique (MLO) views, while complex networks impose substantial computational and memory demands. We propose MCCV-Net, an efficient dual-view Mamba-CNN framework for edge-enabled mammogram analysis. Its Mamba branch captures global context, whereas the CNN branch extracts local lesion features. Cross-attention enables interaction between CC and MLO features; incremental feature integration progressively combines multi-level representations, and a view-consistency loss regularizes cross-view learning. To facilitate resource-constrained inference, we further develop a Knowledge Distillation and View-Asymmetric Student (KDVA-S), comprising three full CPMC modules in the MLO branch, three half-width CPMC-Lite modules in the CC branch, and three fusion stages. A

frozen MCCV-Net teacher provides feature-, fusion-, and output-level supervision. On CBIS-DDSM and VinDr-Mammo, MCCV-Net achieves AUCs of 78.63% and 97.66%, respectively, while KDVA-S obtains 76.13% and 95.75%. Distillation improves the student's AUC by 3.58 and 1.49 percentage points over direct training. On an RTX 4090 workstation, KDVA-S reduces parameters and FLOPs by 40.0%, peak memory by 31.8%, and end-to-end latency by 30.8%, while increasing throughput from 19.2 to 27.8 cases/s. These results demonstrate an effective balance between diagnostic performance and computational efficiency, supporting local inference on resource-constrained edge devices and potentially reducing reliance on centralized image processing. The source code is available at https://github.com/kylemessi/MCCV-Net.

Keywords: Artificial intelligence, Deep learning, Breast cancer diagnosis, Dual-view mammography, State space model, Mammogram classification

1 Introduction

Breast cancer screening is an important application of intelligent healthcare connectivity. As medical imaging increasingly shifts toward distributed acquisition, edge-assisted processing, and cloud-based clinical decision support, mammogram analysis must be accurate, computationally scalable, and suitable for heterogeneous deployment environments. In such settings, the communication infrastructure connects imaging devices, local processing units, and centralized learning or decision services, while AI models convert high-dimensional visual data into clinically useful predictions. The reliability and efficiency of the analytical model therefore directly affect the timeliness and scalability of intelligent screening services.

Mammography illustrates the image-analysis requirements of connected healthcare. A routine examination includes two complementary projections for each breast: the craniocaudal (CC) view and the mediolateral oblique (MLO) view. These views provide partially overlapping but non-identical anatomical evidence that can support joint interpretation. Deep learning methods have been successfully applied to mass detection [1–3], microcalcification analysis [4, 5], and breast lesion segmentation [6, 7]. However, many existing classifiers process only one view or require region-of-interest

annotations, which can increase annotation and preprocessing burdens and limit the scalability of automated screening pipelines.

For image-analysis components intended for connected screening services, three modeling challenges are particularly relevant. First, paired views must be integrated while preserving view-specific evidence, since independent predictions may not fully exploit their complementary diagnostic information. Second, full-field mammograms contain subtle lesions within a large anatomical background, requiring both long-range contextual modeling and fine-grained local feature extraction. Third, limited labeled data and differences in acquisition conditions complicate the development of generalizable classifiers. These challenges motivate a joint investigation of representation quality, cross-view consistency, and computational demands. They also indicate that feature fusion alone should not be assumed to reduce network traffic or deployment cost.

Conventional CNNs are effective at extracting local morphology but have limited receptive fields, whereas self-attention-based Transformers can model global relationships at a substantial computational cost for high-resolution images. A practical model for intelligent healthcare networks should consequently balance global dependency modeling, local diagnostic sensitivity, multi-view collaboration, and computational efficiency. These requirements motivate the hybrid Mamba–CNN and progressive dual-view design developed in this work.

Recent advances in enhancing models' long-range dependency learning and overcoming the limitations of local receptive fields have demonstrated promising results in medical imaging. In particular, architectures based on Vision Transformer (ViT) [8] and its variants [9–11] have achieved state-of-the-art performance in image recognition and semantic segmentation tasks. ViT applies Transformer encoders to sequences of image patches, enabling interactions between spatially distant regions through

self-attention. Dense prediction Transformers further extend this approach to spatially structured outputs [12]. Furthermore, TransUNet [13] established a benchmark architecture in medical image segmentation by integrating CNNs with Transformers, overcoming the limitations of CNN structures in modeling long-range dependencies and handling complex structures with ambiguous boundaries. Subsequent studies have strengthened this hybrid paradigm through boundary-sensitive FCN–Transformer integration [14] and dual-path axial-Transformer modeling that jointly captures global context and local structure in ultrasound images [15]. Although Transformer-based models have demonstrated powerful global modeling capabilities in medical image analysis, the computational complexity of the self-attention mechanism is substantial, posing a significant computational burden for high-resolution full-field mammogram images. Outside medical imaging, Li et al. proposed PMVT for plant disease identification on mobile devices, combining inverted residual structures with attention-enhanced Transformer encoding [16]. This provides a cross-domain example of lightweight hybrid feature extraction, rather than evidence of performance on mammography.

Additionally, although ViT incorporates spatial information through positional encoding, its sensitivity to pixel-level structures and fine-grained local patterns remains limited. To overcome these limitations, several novel architectures have been proposed in recent years, among which Mamba [17] has attracted considerable attention. Built upon the selective state space model (SSM), Mamba enables efficient long-range dependency modeling with linear computational complexity, making it particularly suitable for processing high-resolution full-field medical images. U-Mamba integrates state-space modeling with a convolutional U-Net architecture for biomedical image segmentation [18], demonstrating the relevance of hybrid local and long-range feature extraction to medical imaging. In more specialized applications, Polyp-Mamba [19] proposed a dedicated architecture for polyp segmentation, incorporating a hybrid

multi-frequency fusion encoder and a gated selection decoder to reconcile Mamba's linear complexity with the requirement for cross-scale semantic modeling. Furthermore, a growing body of research has focused on improving Mamba's selective scanning mechanism. For instance, H-vmunet [20] introduced a higher-order 2D selective scan (H-SS2D), which aims to suppress redundant information and generate cleaner global representations. Morphology-guided Mamba–CNN segmentation with boundary-aware auxiliary supervision provides an additional example of combining state-space context modeling with structural priors [21].

These modeling developments address one component of the broader computation–communication problem in connected healthcare: extracting useful diagnostic information from complex observations. Federated mammography learning [22] and semantic-guided feature mixing for federated long-tailed data [23] provide complementary examples of learning across distributed data sources. However, federated training, communication scheduling, and representation transmission are separate design problems from the within-model CC–MLO fusion investigated here. In particular, feature interaction between the two branches is not a communication protocol, and the linear sequence complexity of a Mamba branch does not establish low end-to-end latency or energy consumption for the complete network.

Accordingly, this paper proposes MCCV-Net as a diagnostic inference model for full-field mammograms, with connected screening as its prospective application context. The architecture combines Mamba-based global dependency modeling with residual convolutional extraction of local detail and progressively integrates paired CC and MLO representations. The present study evaluates classification performance, model parameter counts, and the effect of dual-view consistency, rather than claiming a deployed edge–cloud system. Its main contributions are summarized as follows:

1. We develop a hierarchical dual-view Mamba–CNN feature extractor that combines global state-space modeling with local convolutional representations for mammographic classification.
2. We introduce progressive dual-view feature fusion to integrate complementary CC and MLO evidence while retaining view-specific information.
3. We formulate a dual-view feature consistency loss to regularize paired representations from the same breast.
4. We evaluate MCCV-Net on CBIS-DDSM and VinDr-Mammo, obtaining AUC scores of 78.63% and 97.66%, respectively, and analyze the contributions of multi-view input and the consistency constraint.

These results provide an application-level basis for subsequent computation–communication co-design in connected screening. Establishing suitability for resource-constrained or networked deployment will additionally require measurements of inference latency, memory and energy consumption, and communication overhead under specified system conditions.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed method in detail. Section 4 presents the experimental settings and discusses the results. Finally, Section 5 concludes the paper and outlines future research directions.

2 Related Work

2.1 Mamba-based Models in Medical Imaging

Medical image segmentation is a fundamental and critical task in medical image analysis. Its success depends on a model's dual capabilities: precisely delineating anatomical or pathological boundaries at the fine-grained level, such as organ contours and tumor

shapes, while comprehensively understanding global spatial context at the macro level to accurately interpret multi-scale complex structures and relationships within images.

Current mainstream model paradigms face inherent challenges in achieving this dual objective. Convolutional Neural Networks (CNNs), while possessing strong local feature extraction capabilities, suffer from inherently limited receptive fields that constrain their ability to model long-range spatial dependencies. This limitation makes it difficult for CNN-based models to accurately delineate complex anatomical structures that require comprehensive global context. Vision Transformers (ViTs) obtain global modeling capabilities through self-attention mechanisms; however, their computational complexity grows quadratically with image resolution, imposing heavy computational burdens on high-resolution medical images, such as whole-slide pathology images and 3D MRI scans. Moreover, their context-agnostic attention mechanisms may be inefficient when processing homogeneous backgrounds.

Recent research advances indicate that Mamba-based architectures are emerging as a promising new paradigm to address these challenges [24, 25]. Outside medical imaging, Feng et al. investigated an epistemic uncertainty-aware Mamba framework for emotional stress assessment from spontaneous micro-gestures [**?**]. This provides a cross-domain example of uncertainty-aware state-space modeling, although its temporal recognition task differs from mammographic classification.

By leveraging their inherent advantages in efficiently modeling long-range dependencies and adaptive state space representations, Mamba models demonstrate significant performance gains over traditional methods, particularly in challenging scenarios such as high-dimensional 3D volumetric segmentation and heterogeneous multimodal data fusion, including MRI, CT, and PET images. Their ability to efficiently capture both fine local details and overall contextual information makes them powerful tools for improving the accuracy and robustness of medical image interpretation.

Figure 1 illustrates the distinction between ROI-based and full-field mammogram analysis.

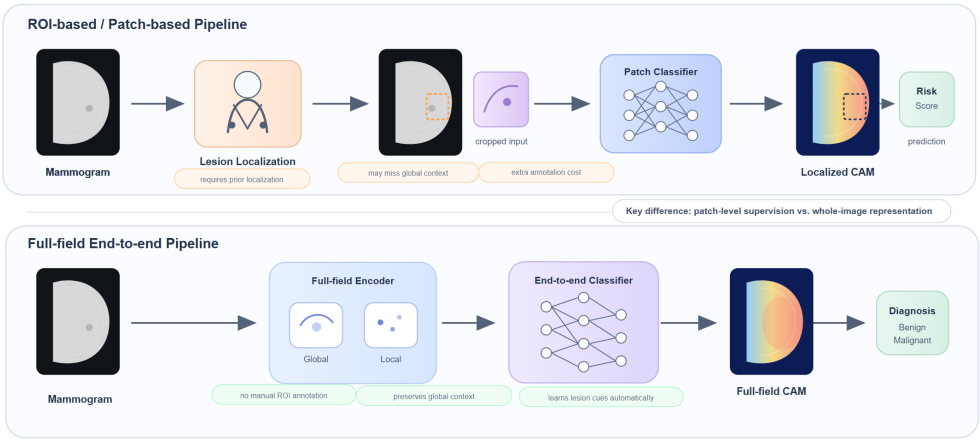

Fig. 1 The difference between full-field and ROI-based methods. The ROI-based method, shown in the upper branch, utilizes pre-annotated positional information during input to assist the model in feature extraction. In contrast, the full-field method, shown in the lower branch, takes the full-size image as input without relying on such prior localization cues, which places greater demand on the model's ability to focus on discriminative features from key regions.

2.2 Full-field Mammogram Analysis

In the field of 2D medical image segmentation, DA-MambaNet [26] proposes an end-to-end segmentation model that integrates multi-attention Mamba blocks for feature extraction and axial residual blocks for feature recovery, achieving state-of-the-art performance on multiple public medical imaging datasets. MambaMIM [27] introduces an innovative pre-training framework, whose core contribution is the Tokenized State Space Sequence Interpolation (TOKI) strategy. This strategy effectively addresses the limitations of traditional masked image modeling (MIM) in causal dependency modeling, thereby significantly enhancing the model's capability to capture long-range dependencies in 3D medical images. LoG-VMamba [28] introduces a local-global visual state space model that employs a straightforward scanning strategy to concurrently

capture both local and global dependencies, effectively addressing the high computational cost associated with complex scanning strategies in high-dimensional medical image analysis.

To address the challenges of multimodal segmentation in whole-brain scenarios, Tang et al. [29] proposed an innovative method for PET/MR dual-modal images. This approach introduces a cross-fusion mechanism that effectively integrates functional and structural information, enabling high-precision whole-brain parcellation and providing new insights for medical image analysis and the diagnosis of neurological disorders. MedMamba [30], as one of the first general visual Mamba frameworks for medical image classification, combines the local feature extraction capability of convolutional layers with the long-range dependency modeling ability of SSMs, enabling efficient processing of medical images from diverse imaging modalities. Qu et al. compared MRI radiomics-based machine-learning models for differentiating common sellar-region lesions, providing a complementary example of quantitative feature-based lesion classification outside breast imaging [**?**].

Considering the complementary nature of Mamba and CNN models, recent years have witnessed growing research interest in hybrid architectures that leverage CNNs for local feature extraction and Mamba for global dependency capture. Such designs enable a strong complementarity between local perception and global modeling capabilities.

Fully automated full-field mammography image classification without human intervention is a core task in CAD for breast cancer screening. These images present significant analytical challenges: they are extremely high-resolution, often exceeding 3000×4000 pixels, while pathological features, such as microcalcifications or spiculated masses, typically occupy only minute spatial areas, usually less than 5% of the total image area. Early CAD frameworks primarily adopted an ROI-centric paradigm, where expert radiologists manually delineated suspicious areas for subsequent analysis

[31–33]. While this approach enables the deployment of computationally lightweight models, such as support vector machines or shallow convolutional neural networks, it suffers from two critical limitations: (i) heavy reliance on domain experts for labor-intensive annotation, which is prone to inter-observer variability; and (ii) irreversible loss of global anatomical context during lesion extraction, hindering comprehensive morphological assessment based on breast tissue architecture.

The current paradigm is shifting toward end-to-end analysis of raw full-field mammograms, driven by synergistic advances in model architecture and multi-scale feature fusion. Emerging hybrid architectures, particularly Mamba-CNN collaborative frameworks, are reshaping the technological frontier. The Mamba architecture achieves global context modeling with linear complexity through structured state space models, efficiently capturing complex lesion-tissue relationships in megapixel-scale inputs. This holistic approach not only mimics radiologists' diagnostic workflow, which inherently integrates local findings with overall breast structure, but also preserves critical spatial relationships that may be obscured by local-slice-based methods. Consequently, modern full-field analysis pipelines can improve diagnostic reliability while alleviating annotation bottlenecks, paving the way for more autonomous and context-aware computer-aided diagnosis systems that can be integrated into clinical decision protocols.

Early full-field methodologies commonly decomposed the task into a two-stage pipeline: initial detection of regions of interest across the full image, followed by patch-level classification. For instance, the GMIC model [34] employed a memory-efficient, low-parameter network to identify informative regions across the whole mammogram. It subsequently applied a high-capacity network to extract detailed features from these localized areas and finally integrated both global and local information through a fusion module to generate predictions. In a related effort, Sun et al. [35] proposed a multi-scale region selection network (MRSN) for end-to-end full-field mammogram

classification. This approach eliminates the need for lesion segmentation or ROI annotations by automatically screening and focusing on tumor regions at the feature level, enabling the whole-image classifier to achieve performance comparable to patch-based classifiers. For the task of breast segmentation, Larroza et al. [36] fine-tuned the general-purpose Segment Anything Model (SAM) to achieve precise delineation of the breast region and pectoral muscle. This work represents one of the first adaptations of SAM for breast segmentation in mammograms, demonstrating superior robustness and generalization capabilities through evaluations across multiple centers and datasets.

2.3 Mammogram Classification

Contemporary research in mammographic image classification has undergone a paradigm shift from isolated single-image analysis toward clinically grounded diagnostic frameworks that more closely mirror radiologists' decision-making workflows. This evolution is driven by two critical clinical imperatives. First, radiologists inherently leverage complementary diagnostic evidence from multiple perspectives, specifically bilateral craniocaudal and mediolateral oblique views of both breasts, to resolve ambiguities in lesion morphology and reduce false positives. Consequently, state-of-the-art systems now prioritize multi-view fusion architectures that algorithmically replicate this comparative assessment protocol by dynamically weighting view-specific features while preserving spatial correspondence across projections. Second, diagnostic accuracy in breast cancer screening can be further enhanced by cross-modal contextual integration, such as synthesizing full-field mammograms with prior screening histories, ultrasound elastography, or genomic risk markers to resolve indeterminate findings. Pioneering frameworks have therefore employed heterogeneous modality alignment modules, such as cross-attention Transformers or contrastive representation learners,

to fuse these disparate data streams while preserving modality-specific diagnostic signatures. Related methodological developments outside mammography further illustrate the importance of selective information exchange. LR-ASD [37] combines lightweight audio-visual feature extraction, channel-attention fusion, and temporal modeling for active speaker detection. This study offers methodological context for efficient fusion, although its audio-visual setting differs from the paired X-ray projections considered here.

Contemporary research paradigms have increasingly moved beyond the diagnostic constraints of isolated single-projection mammography and have pivoted toward integrated multi-temporal and multi-geometric frameworks that better replicate radiologists' holistic interpretive workflows. For instance, SwinCAMF-Net [38] employs a cross-modal attention fusion module to align and integrate 2D multi-view images, 3D lesion volume data, and clinical metadata, leading to strong performance in both classification and lesion segmentation tasks. Meanwhile, Amelia et al. [22] explored, for the first time in a federated learning context, the use of curriculum learning to enhance multi-site breast cancer classification performance. Semantic-guided feature mixing has also been proposed to reduce bias in federated learning with long-tailed data, offering a possible methodological reference for future multi-center mammography studies [23].

To address the core challenges of limited data volume and high annotation costs in medical imaging, researchers have also explored hybrid architectures and supervision-robust learning strategies. In the broader field of recognition with limited supervision, Feng et al. identify reliable modalities and transfer knowledge from more reliable to less reliable modalities for few-shot action recognition [39]. This offers a cross-domain reference for reliability-aware information exchange, rather than direct evidence for dual-view mammographic classification. Dual-student mean-teacher learning with prototype-based pseudo-label refinement has been used to exploit unlabeled medical

images [40], while teacher-guided adaptive label correction and feature perturbation improve segmentation robustness under noisy annotations [41]. One representative hybrid architecture is FeatFusion-ViT [42], which combines the global attention mechanism of Vision Transformers with local texture information extracted from traditional histogram of oriented gradients (HOG) features. This hybrid approach has demonstrated superior performance over single-model architectures on datasets such as CBIS-DDSM while also improving the interpretability of model decisions.

In recent years, numerous studies have also focused on developing comprehensive frameworks that simulate the entire clinical diagnostic process. A representative model is LAS-GAM [43], which not only integrates a patient's four standard-view images but also incorporates a global module, simulating a radiologist's comprehensive assessment, and a lesion screening module, simulating lesion localization by comparing symmetrical regions. This design improves diagnostic performance, achieving an AUC of 0.817 on the DDSM dataset, while reducing reliance on finely annotated lesion data and enhancing its practical utility.

Furthermore, models such as MammoViT [44] illustrate the trend of hybrid architectural innovation. MammoViT combines ResNet50 for feature extraction with a Vision Transformer to capture global dependencies and employs the SMOTE technique to mitigate class imbalance, thereby accomplishing multi-class BI-RADS classification.

3 Methodology

3.1 Approach Overview

As illustrated in Fig. 2, the proposed dual-view classification framework consists of two core components: (i) a dual-view feature extraction module based on a hybrid Mamba-CNN architecture, and (ii) a progressive cross-view feature fusion module. Each feature extractor processes one mammographic view, i.e., the craniocaudal (CC)

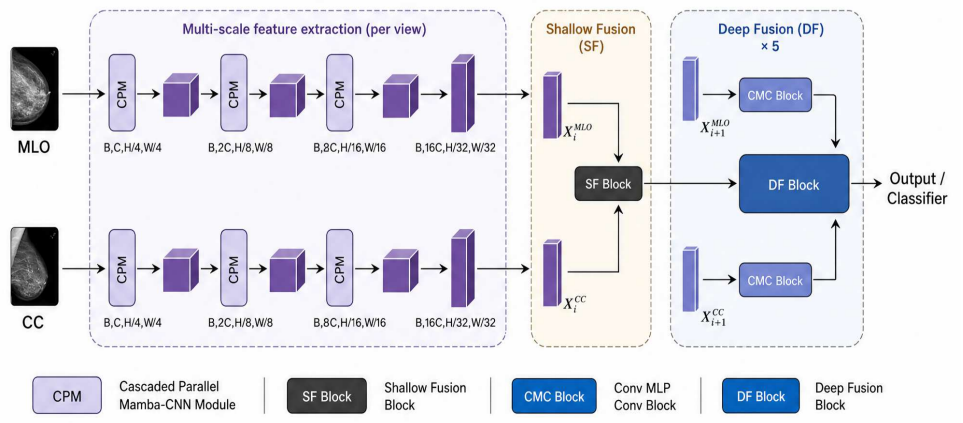

Fig. 2 Overall architecture of the proposed MCCV-Net framework. Images from two views are fed into a feature extraction network for multi-scale feature learning. The extracted features are subsequently processed through a hierarchical dual-view fusion pipeline, comprising an initial shallow fusion stage for geometric alignment followed by multi-stage deep fusion layers that reconcile semantic discrepancies across views. The unified representation is then fed into the diagnostic classifier.

view or the mediolateral oblique (MLO) view, and synergistically integrates the State Space Model (SSM)-based Mamba backbone with hierarchical convolutional operations. This hybrid design exploits the global context modeling capability of Mamba with linear computational complexity $\mathcal{O}(n)$ to capture long-range anatomical dependencies, while the CNN branch extracts discriminative local morphological features, such as microcalcification clusters and spiculated mass margins, through multi-scale receptive fields. The progressive fusion module dynamically aligns and integrates multi-scale features from the two views using anatomically guided cross-view attention. Finally, the fused representations are fed into a classification head, which is further regularized by the proposed view-consistency constraint.

Dual-view mammograms are independently processed by two symmetric Mamba-CNN hybrid streams, as shown in Fig. 2. Each stream constructs a hierarchical multi-scale feature representation by progressively integrating Mamba-derived global contextual information with CNN-derived local details at different resolutions. Features at corresponding scales from the CC and MLO views are then fed into the proposed progressive cross-view fusion module. Specifically, this module performs

coarse-to-fine cross-view interaction by first aligning low-resolution semantic representations and then progressively incorporating high-resolution fine-grained details through scale-adaptive attention weights. The fused multi-scale features are subsequently aggregated for final prediction. The optimization process is jointly supervised by the standard classification loss and the proposed view-consistency loss, encouraging diagnostic consistency between the CC and MLO views while preserving view-specific complementary information.

3.2 Cascaded Parallel Mamba-CNN Module

To harmonize global sequence modeling with local perceptual representation, we propose a Cascaded Parallel Mamba-CNN (CPMC) module, a hybrid architecture that integrates State Space Models (SSMs) with residual convolutional networks, as illustrated in Fig. 3. The proposed module is employed in two dedicated feature extraction streams to process dual-view mammograms, i.e., the craniocaudal (CC) and mediolateral oblique (MLO) views. Each stream consists of a cascade of CPMC blocks, which progressively extract hierarchical feature representations from the corresponding mammographic view. Within each CPMC block, the input features are dynamically partitioned by a learnable channel-gating mechanism into two complementary pathways:

(i) A residual CNN branch is designed to preserve high-frequency diagnostic details, such as microcalcification clusters and mass margins. This branch consists of two sequential residual layers, each of which applies convolution, batch normalization, max-pooling, and ReLU activation. Skip connections are introduced to facilitate gradient propagation and alleviate feature degradation during deep representation learning.

(ii) An SSM branch based on the Mamba architecture is employed to capture long-range spatial dependencies with linear computational complexity, thereby avoiding

the quadratic computational overhead commonly associated with Transformer-based self-attention mechanisms.

The CPMC module is designed to construct a dual-view symmetric hierarchical feature extraction network, with the goal of overcoming the inherent limitations of conventional single-architecture models in medical image analysis. Specifically, the module separately models global long-range dependencies and local fine-grained details through its parallel Mamba and CNN branches, respectively. The Mamba branch captures global anatomical context and long-range spatial interactions, whereas the CNN branch preserves local lesion-specific morphology, such as mass textures, spiculated margins, and microcalcification patterns. By further integrating these complementary representations through the progressive dual-view fusion module, the proposed framework achieves synergistic enhancement of global and local features, thereby improving its ability to discriminate diverse mammographic lesions.

As illustrated in Fig. 3, each input mammogram is first passed through a dynamic channel-gating mechanism within the CPMC module. The learnable gating weights partition the feature tensor along the channel dimension into two complementary subsets, which are then processed in parallel by two dedicated pathways. One subset is fed into the residual CNN branch, where it passes through two cascaded residual layers. Each residual layer consists of a 3×3 convolution followed by batch normalization, ReLU activation, and max-pooling. The skip connections facilitate stable optimization and help preserve diagnostically critical microstructures, such as microcalcification boundaries and spiculated mass margins. Meanwhile, the other subset is forwarded to the SSM branch, which captures long-range spatial dependencies through selective state space modeling. The outputs of the two branches are subsequently integrated via channel shuffling, enabling effective cross-pathway feature exchange. This design jointly enhances local morphological fidelity and global anatomical context, thereby supporting robust dual-view mammographic interpretation. For each mammographic

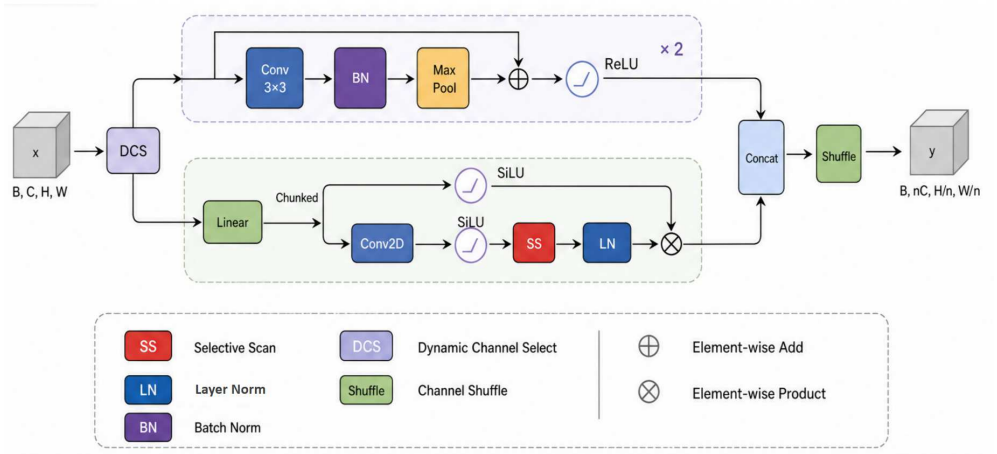

Fig. 3 Overview of Cascaded Parallel Mamba-CNN module.

view, the preprocessed image is fed into a stack of four CPMC blocks to progressively construct hierarchical feature representations. After each CPMC block, the spatial resolution is reduced by a factor of two along both the height and width dimensions, resulting in one-quarter of the original spatial area, while the number of channels is doubled. Without loss of generality, we take the MLO view as an example. Given the input feature tensor of a CPMC block, $A^{MLO} \in \mathbb{R}^{B\times C\times H\times W}$, where B, C, H, and W denote the batch size, number of channels, height, and width, respectively, a dynamic channel selection mechanism is first applied to partition the input into two complementary feature subsets: $A^{MLO}_{Re} \in \mathbb{R}^{B\times C/2\times H\times W}$ and $A^{MLO}_{Mam} \in \mathbb{R}^{B\times C/2\times H\times W}$. The former is forwarded to the residual CNN branch, while the latter is processed by the Mamba-based SSM branch.

Let the input feature tensor be $\mathbf{x} \in \mathbb{R}^{B\times C\times H\times W}$. The dynamic channel selection operation can be formulated as

$$A_{Re}, A_{Mam} = \text{Select}(\mathbf{x}), \tag{1}$$

where Select$(\cdot)$ denotes a learnable channel selection operation. It adaptively partitions the input tensor along the channel dimension into two subsets, A_{Re} and A_{Mam}, which serve as the inputs to the residual CNN branch and the Mamba branch, respectively. Both subsets have the same shape, i.e., $A_{Re}, A_{Mam} \in \mathbb{R}^{B\times C/2\times H\times W}$.

The residual CNN branch aims to preserve local fine-grained morphological details, such as microcalcification boundaries and spiculated mass margins. Given the input $A_{Re} \in \mathbb{R}^{B\times C/2\times H\times W}$, two cascaded residual blocks are applied to extract local discriminative features. Each residual block consists of a 3×3 convolution, batch normalization, ReLU activation, and a residual shortcut. The residual transformation is defined as

$$\mathcal{R}(\mathbf{z}) = \mu\left(\mathbf{z} + \mathcal{F}_{\mathrm{BN}}\left(\mathbf{W}_q(\mathbf{z})\right)\right), \tag{2}$$

where $\mathbf{W}_q$ denotes a 3×3 convolution, $\mathcal{F}_{\mathrm{BN}}(\cdot)$ represents batch normalization, and $\mu(\cdot)$ denotes the ReLU activation function. The output of the residual CNN branch is obtained by

$$\mathcal{B}^{up}(A_{Re}) = \mathcal{D}_{Re}\left(\mathcal{R}\left(\mathcal{R}(A_{Re})\right)\right), \tag{3}$$

where $\mathcal{D}_{Re}(\cdot)$ denotes the downsampling operation, such as max-pooling or strided convolution. The resulting feature tensor is $\mathcal{B}^{up}(A_{Re}) \in \mathbb{R}^{B\times C\times H/2\times W/2}$.

The Mamba-based SSM branch is designed to capture long-range spatial dependencies with linear computational complexity. Given the input $A_{Mam} \in \mathbb{R}^{B\times C/2\times H\times W}$, a channel expansion layer is first employed to increase the channel dimension. The expanded feature is then split into two streams along the channel dimension:

$$m_1, m_2 = \mathrm{Chunk}\left(\mathcal{F}_{\mathrm{EX}}(A_{Mam})\right), \tag{4}$$

where $\mathcal{F}_{\mathrm{EX}}(\cdot)$ denotes the channel expansion module and Chunk$(\cdot)$ represents the channel-wise splitting operation.

The first stream is processed by a selective state scanning pathway composed of convolution, SiLU activation, selective scan, and layer normalization, while the second stream serves as a gating signal. The output of the Mamba branch is formulated as

$$\mathcal{T}(A_{Mam}) = \mathcal{F}_{\mathrm{LN}}\left(\mathcal{P}_{\mathrm{SS}}\left(\varphi\left(\mathbf{W}_K(m_1)\right)\right)\right) \odot \varphi(m_2), \tag{5}$$

where $\mathbf{W}_K$ denotes a 3×3 convolution, $\varphi(\cdot)$ is the SiLU activation function, $\mathcal{P}_{\mathrm{SS}}(\cdot)$ represents the selective state scanning mechanism, $\mathcal{F}_{\mathrm{LN}}(\cdot)$ denotes layer normalization, and $\odot$ indicates element-wise multiplication.

To match the spatial scale and channel dimension of the residual CNN branch, a downsampling and projection operation is further applied:

$$\mathcal{B}^{low}(A_{Mam}) = \mathcal{D}_{Mam}\left(\mathcal{T}(A_{Mam})\right), \tag{6}$$

where $\mathcal{D}_{Mam}(\cdot)$ denotes the downsampling and channel projection operation. The output of the Mamba branch is $\mathcal{B}^{low}(A_{Mam}) \in \mathbb{R}^{B\times C\times H/2\times W/2}$.

The outputs of the residual CNN branch and the Mamba branch are concatenated along the channel dimension:

$$\mathbf{Z}_{cat} = \mathcal{B}^{up}(A_{Re}) \oplus \mathcal{B}^{low}(A_{Mam}), \tag{7}$$

where $\oplus$ denotes channel-wise concatenation. Thus, $\mathbf{Z}_{cat} \in \mathbb{R}^{B\times 2C\times H/2\times W/2}$.

To promote information exchange between the CNN-derived local features and the Mamba-derived global representations, a channel shuffle operation is subsequently applied:

$$\mathcal{M}(\mathbf{x}) = \mathrm{Shuffle}(\mathbf{Z}_{cat}), \tag{8}$$

where Shuffle(·) denotes the channel shuffling mechanism. The output of the CPMC block is therefore $\mathcal{M}(\mathbf{x}) \in \mathbb{R}^{B\times 2C\times H/2\times W/2}$. This output is then fed into the next CPMC block to progressively build a hierarchical representation with enlarged channel capacity and reduced spatial resolution.

3.3 Dual-view Feature Fusion Module

As illustrated in Fig. 4, we design a hierarchical cascaded dual-view feature fusion module to progressively aggregate cross-view information through staged integration. Given the dual-view feature inputs, denoted as X_i^{MLO} and X_i^{CC}, the two views are first concatenated and projected by the Dual-view Shallow Fuse module to establish preliminary anatomical correspondences between the mediolateral oblique (MLO) and craniocaudal (CC) views. This process yields an initial fused representation, denoted as F_i, which serves as the input to the subsequent Dual-view Deep Fuse module.

The Dual-view Deep Fuse module further refines the fused representation through five cascaded fusion stages. Each stage adopts a multi-branch architecture that integrates convolutional operations for preserving local morphological details, MLP layers for channel-wise feature recalibration, and State Space Models (SSMs) for modeling long-range spatial dependencies. Layer normalization and residual connections are incorporated to stabilize optimization and facilitate effective information interaction across views. Specifically, after feature extraction, the multi-scale view-specific features X_i^{MLO} and X_i^{CC} are first rapidly integrated by the shallow fusion module to generate the preliminary fused feature F_i. The deep fusion module then takes F_i together with the view-specific memory states from the previous stage, i.e., M_{i-1}^{MLO} and M_{i-1}^{CC}, as inputs. These memory states are exclusively involved in the SSM-based computations, enabling the model to retain and propagate view-specific contextual information across fusion stages. The module produces three outputs, including the

updated fused representation and the refreshed memory states for the MLO and CC views, which are subsequently propagated to the next fusion stage.

Through this five-stage iterative fusion process, the proposed module progressively reconciles morphological details and semantic consistencies between the MLO and CC views. Consequently, it achieves fine-grained cross-view alignment and complementary enhancement of dual-view mammographic features.

The input to the dual-view fusion module consists of two enhanced feature maps extracted from the corresponding view-specific branches, denoted as X_i^{MLO} and X_i^{CC}. First, a Shallow Fuse layer is applied to generate an initial cross-view representation. Specifically, the two feature maps are concatenated along the channel dimension to obtain a preliminary fused feature. This concatenated feature is then fed into an MLP-based transformation, which performs nonlinear feature mapping while compressing the channel dimension. Through this process, the Shallow Fuse layer extracts a more compact and discriminative fused representation for subsequent deep fusion.

The shallow fused feature is then forwarded to the Deep Fuse module, which consists of five cascaded fusion layers for progressive feature refinement. At the i-th fusion layer, the module receives three inputs: X_i^{SF}, X_{i-1}^{MLO}, and X_{i-1}^{CC}. Here, X_i^{SF} denotes the shallow fused feature at the current layer, while X_{i-1}^{MLO} and X_{i-1}^{CC} represent the MLO-view and CC-view memory features propagated from the previous SSM module, respectively. By leveraging these view-specific memory features, the SSM module captures cross-view contextual dependencies and progressively enhances the fused representation, producing a more discriminative output feature Z_i^o.

At the i-th Deep Fuse stage, the SSM module takes three inputs, namely $\boldsymbol{Z}_{i-1}^{\mathrm{CC}}$, $\boldsymbol{Z}_{i-1}^{\mathrm{MLO}}$, and $\boldsymbol{Z}_i^{\mathrm{SF}}$, where $i = 1, 2, \ldots, 5$ denotes the stage index of the cascaded Deep Fuse module. To obtain the view-specific inputs for the SSM module, the memory features propagated from the previous stage are first processed by a CMC

transformation:

$$\boldsymbol{Z}_{i-1}^{\mathrm{CC}} = \mathcal{F}_{\mathrm{CMC}}\left(\boldsymbol{X}_{i-1}^{\mathrm{CC}}\right) \in \mathbb{R}^{B\times 16C\times H/32\times W/32}, \tag{9}$$

$$\boldsymbol{Z}_{i-1}^{\mathrm{MLO}} = \mathcal{F}_{\mathrm{CMC}}\left(\boldsymbol{X}_{i-1}^{\mathrm{MLO}}\right) \in \mathbb{R}^{B\times 16C\times H/32\times W/32}. \tag{10}$$

Here, $\mathcal{F}_{\mathrm{CMC}}(\cdot)$ represents the CMC transformation, which consists of a 3×3 convolutional layer, an MLP layer, and a subsequent 3×3 convolutional layer. By combining convolutional local modeling with MLP-based channel interaction, the CMC transformation enhances the view-specific features and projects them into a unified latent space for subsequent SSM-based dependency modeling. The shallow fused feature $\boldsymbol{X}_i^{\mathrm{SF}}$ is first normalized and projected into a unified latent space before being fed into the Spatial-Semantic Module. Specifically, this transformation is defined as

$$\boldsymbol{Z}_i^{\mathrm{SF}} = \mathcal{M}_1\left(\nabla_1\left(\mathcal{F}_{\mathrm{LN}}\left(\boldsymbol{X}_i^{\mathrm{SF}}\right)\right)\right) \in \mathbb{R}^{B\times 16C\times H/32\times W/32}, \tag{11}$$

where $\mathcal{F}_{\mathrm{LN}}(\cdot)$ denotes layer normalization, $\nabla_1(\cdot)$ is an MLP-based projection, and $\mathcal{M}_1(\cdot)$ represents a 3×3 convolutional operation.

Given the projected shallow fused feature $\boldsymbol{Z}_i^{\mathrm{SF}}$ and the view-specific memory features from the previous stage, i.e., $\boldsymbol{Z}_{i-1}^{\mathrm{MLO}}$ and $\boldsymbol{Z}_{i-1}^{\mathrm{CC}}$, the Spatial-Semantic Module performs cross-view dependency modeling as follows:

$$\boldsymbol{C}_i^{\mathrm{SSM}} = \mathcal{F}_{\mathrm{SSM}}\left(\boldsymbol{Z}_i^{\mathrm{SF}}, \boldsymbol{Z}_{i-1}^{\mathrm{MLO}}, \boldsymbol{Z}_{i-1}^{\mathrm{CC}}\right) \in \mathbb{R}^{B\times 16C\times H/32\times W/32}, \tag{12}$$

where $\mathcal{F}_{\mathrm{SSM}}(\cdot)$ denotes the Spatial-Semantic Module. This module jointly models spatial relationships and semantic dependencies among the shallow fused representation and the two view-specific memory features, thereby facilitating cross-view feature alignment and complementary information interaction.

Meanwhile, a residual modulation branch is introduced to preserve and recalibrate the original shallow fused information:

$$\boldsymbol{C}_i^{\mathrm{Re}} = \varphi\left(\nabla_2\left(\mathcal{F}_{\mathrm{LN}}\left(\boldsymbol{X}_i^{\mathrm{SF}}\right)\right)\right) \in \mathbb{R}^{B\times 16C\times H/32\times W/32}, \tag{13}$$

where $\nabla_2(\cdot)$ represents an MLP-based projection and $\varphi(\cdot)$ denotes the SiLU activation function.

Finally, the SSM-enhanced feature is modulated by the residual feature through element-wise multiplication, followed by a nonlinear MLP projection to produce the dual-view fused output:

$$\boldsymbol{C}_i^{o} = \nabla_3\left(\boldsymbol{C}_i^{\mathrm{SSM}} \odot \boldsymbol{C}_i^{\mathrm{Re}}\right) \in \mathbb{R}^{B\times 16C\times H/32\times W/32}, \tag{14}$$

where $\nabla_3(\cdot)$ denotes the final MLP projection and $\odot$ indicates element-wise multiplication.

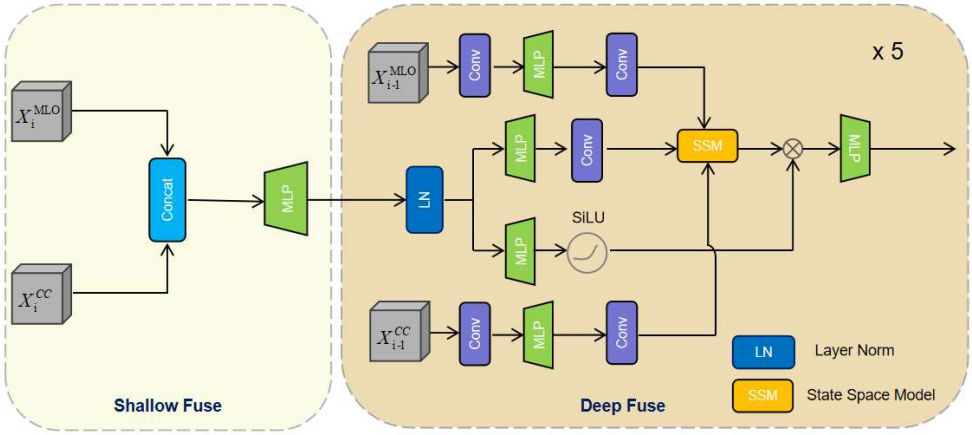

Fig. 4 The overview of Dual-view Feature Fusion Module.

3.4 Loss Function

To train the proposed dual-view model effectively, we adopt a composite loss function composed of the Dual-View Consistency Loss and the binary cross-entropy loss. Since the MLO and CC views of the same breast share consistent anatomical and tissue-structure information, the Dual-View Consistency Loss is introduced to encourage view-invariant feature learning by minimizing the discrepancy between paired view-specific representations.

Given a mini-batch with N paired samples, let $\mathbf{f}_i^{\mathrm{MLO}}$ and $\mathbf{f}_i^{\mathrm{CC}}$ denote the feature vectors extracted from the MLO and CC views of the i-th case, respectively. The Dual-View Consistency Loss is defined based on cosine distance as follows:

$$\mathcal{L}_{\mathrm{DV}} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{\mathbf{f}_i^{\mathrm{MLO}} \cdot \mathbf{f}_i^{\mathrm{CC}}}{\left\| \mathbf{f}_i^{\mathrm{MLO}} \right\|_2 \left\| \mathbf{f}_i^{\mathrm{CC}} \right\|_2} \right), \tag{15}$$

where $\| \cdot \|_2$ denotes the ℓ_2 norm. This loss maximizes the cosine similarity between paired MLO and CC features, thereby promoting cross-view consistency.

For classification supervision, we employ the binary cross-entropy loss:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right], \tag{16}$$

where p_i is the predicted probability and $y_i \in \{0, 1\}$ is the ground-truth label.

The overall loss is formulated as:

$$\mathcal{L}_{\mathrm{Total}} = \alpha \mathcal{L}_{\mathrm{DV}} + \beta \mathcal{L}_{\mathrm{BCE}}, \tag{17}$$

where α and β are hyperparameters that control the relative importance of the two loss terms. The joint optimization of these objectives enables the model to learn discriminative classification features while maintaining robust dual-view consistency.

4 Experiments

The proposed methodology was rigorously evaluated on two public mammography benchmarks: the Curated Breast Imaging Subset of the Digital Database for Screening Mammography and the Vietnam National Hospital of Radiology and Oncology Mammography dataset. Both datasets used in this study are publicly available benchmark datasets rather than self-collected clinical data; the former is distributed by The Cancer Imaging Archive, and the latter is distributed by PhysioNet. These datasets cover diverse clinical scenarios, including variations in imaging equipment, such as Hologic, GE, and Siemens systems, acquisition protocols, patient demographics, breast density, and pathological manifestations, including masses, microcalcifications, and architectural distortions. Therefore, they provide comprehensive validation under real-world heterogeneity.

We first conducted a systematic comparison with classical baseline architectures, including ResNet-50, Vision Transformer, and Vision Mamba, under strictly controlled experimental conditions to establish performance baselines. Subsequently, we benchmarked the proposed MCCV-Net model across two critical diagnostic dimensions: (i) malignancy/benignity classification performance, measured by AUC and ACC, and (ii) diagnostic region-of-interest (ROI) localization capability. In addition, we analyzed the diagnostic value of multi-view fusion by comparing single-view and dual-view CC-MLO processing strategies. Finally, an ablation study on the proposed Dual-View Consistency Loss (DVCL) was conducted to validate its contribution to the overall framework.

4.1 Datasets

4.1.1 CBIS-DDSM

Figure 5 shows an example mammogram with its corresponding lesion mask and lesion region from CBIS-DDSM.

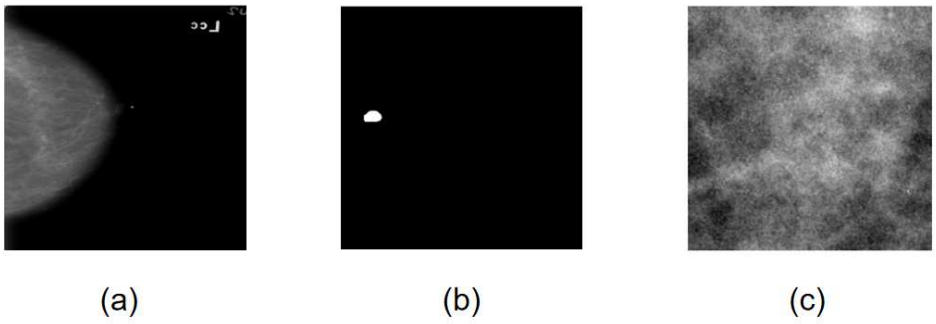

Fig. 5 A sample from the CBIS-DDSM dataset: (a) original image, (b) mask image, and (c) lesion region image.

The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) is a rigorously standardized public dataset derived from the Digital Database for Screening Mammography (DDSM) [45, 46]. It contains 2,462 high-resolution mammograms from 1,231 unique patients with paired bilateral craniocaudal and mediolateral oblique views, which closely mirror clinical screening protocols. Expert board-certified mammographers decompressed the original lossy JPEG images, re-annotated all findings to eliminate ambiguities, and provided pathology-confirmed labels for 753 microcalcification clusters and 891 mass lesions, totaling 1,644 radiologically verified abnormalities. The dataset is also accompanied by pixel-level segmentation masks and bounding boxes formatted in COCO/VOC standards, facilitating integration with modern deep learning frameworks. Each case further includes diagnostic outcomes, including benign or malignant labels, and BI-RADS descriptors. Therefore, CBIS-DDSM serves as an NIH/NCI-endorsed benchmark resource that bridges retrospective data limitations and prospective AI validation by providing radiologist-verified ground truth for developing clinically interpretable computer-aided detection and diagnosis systems.

4.1.2 VinDr-Mammo

The VinDr-Mammo dataset represents a significant advancement in ethnically diverse mammography resources and is distributed as version 1.0.0 through PhysioNet [47, 48].

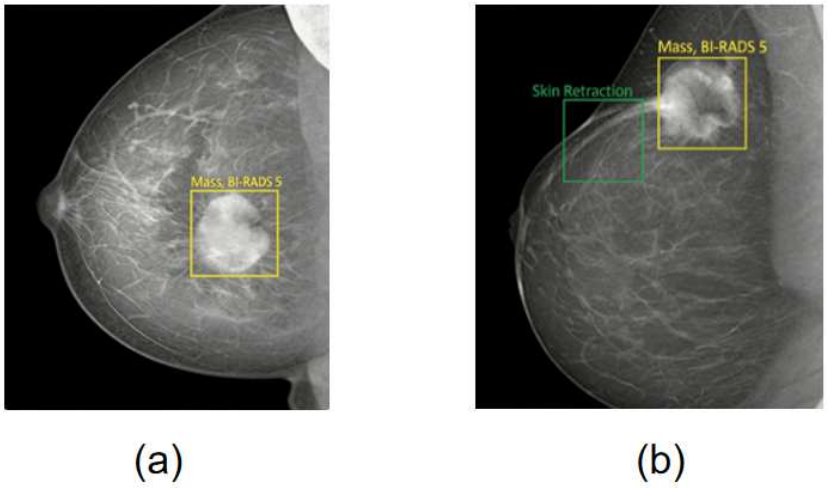

Fig. 6 A sample mammography examination from the VinDr-Mammo dataset assessed as BI-RADS 5: (a) a case containing mass and skin retraction annotations, and (b) a case containing mass annotation information.

It comprises 5,000 four-view screening examinations, totaling over 10,000 DICOM images, with comprehensive breast-level BI-RADS assessments and pixel-level annotations for lesions, including masses, microcalcifications, and architectural distortions. The dataset was rigorously curated under ethical approval protocols, and all images were anonymized and randomly sampled from the Picture Archiving and Communication System (PACS) of Hanoi Medical University Hospital and Military Hospital 108, two of Vietnam's largest tertiary referral centers, between 2018 and 2020. It includes both mediolateral oblique and craniocaudal views for the left and right breasts under standardized clinical acquisition protocols. This Vietnam-sourced dataset addresses critical gaps in Asian breast imaging research by providing radiologist-verified ground truth across diverse breast densities and pathological conditions. It therefore establishes an important benchmark for developing equitable AI systems validated on populations historically underrepresented in global medical imaging repositories. Figure 6 shows example mammograms and lesion annotations from VinDr-Mammo.

Table 1 Hyperparameter configuration of MCCV-Net.

Hyperparameter	CBIS-DDSM	VinDr-Mammo
Batch size	8	16
Learning rate	0.005	0.001
Epochs	100	100
Weight decay	0.001	0.001
Image size	382×382	382×382
Classification threshold	0.5	0.5

4.2 Experimental Setup

In this section, we describe the experimental setup of this study. The experiments were designed with two primary objectives: (i) to validate the effectiveness of the proposed dual-view model, and (ii) to evaluate its capability for full-field mammogram classification by comparing it with existing benchmark methods. The detailed experimental configurations are presented as follows.

To improve model robustness and generalization, we applied several data augmentation strategies to the original mammograms, including random cropping, random horizontal flipping, brightness adjustment, contrast adjustment, and normalization. Since the proposed framework processes full-field mammograms rather than ROI-based inputs, the random cropping operation does not rely on ROI annotations. Therefore, for each dataset, we empirically selected a crop size that preserves the key anatomical regions in most cases. All cropped images were then resized to a fixed resolution of 382×382 pixels. For both datasets, image enhancement and normalization were implemented using the OpenCV Python package. The main training hyperparameters are listed in Table 1.

4.3 Implementation Details

The proposed model was implemented using the PyTorch framework and trained on a workstation equipped with two NVIDIA RTX 4090 GPUs. The Adam optimizer

was employed for model optimization. For the CBIS-DDSM dataset, the initial learning rate was set to 0.005, while for the VinDr-Mammo dataset, it was set to 0.001. The weight decay was set to 0.001 for both datasets. A ReduceLROnPlateau scheduler was adopted to dynamically adjust the learning rate according to the validation performance.

For stable fine-tuning, the learning rates of the pre-trained backbone parameters were set to 5×10^{-5} and 1×10^{-5} for the two datasets, respectively. The model was trained for 100 epochs on both datasets. The batch sizes were set to 8 and 16 for CBIS-DDSM and VinDr-Mammo, respectively. Binary cross-entropy loss was used for classification supervision, and a threshold of 0.5 was applied to convert the predicted probability into the final binary classification result.

4.4 Quantitative Evaluations

We conducted a series of experiments to comprehensively evaluate the proposed MCCV-Net. First, we compared several representative backbone architectures on two benchmark datasets, including classical convolutional neural networks, Transformer-based models, and Mamba-based models. This comparison provides a baseline reference for evaluating the effectiveness of the proposed hybrid Mamba-CNN design.

Second, we compared MCCV-Net with other state-of-the-art methods to demonstrate its advantages in full-field mammogram classification. The evaluation focuses on malignancy/benignity classification performance using commonly adopted metrics such as AUC and ACC. Finally, to further analyze the source of performance improvement and examine whether the model can focus on clinically relevant regions in the full-field setting, we conducted feature attention visualization experiments.

4.4.1 Comparison of Baseline Models on CBIS-DDSM

As shown in Table 2, we first compare several representative baseline architectures on the CBIS-DDSM dataset. Since these baselines are single-view models, they uniformly

Table 2 Comparison of baseline model parameter counts, AUC, and ACC on the CBIS-DDSM dataset.

Model	Params	View	AUC (%)	ACC (%)
Swin-Transformer	19.4 M	Single	63.23	62.23
ResNet-50	33.8 M	Single	67.41	65.02
Mamba-Vision	23.8 M	Single	67.83	65.80

Table 3 Comparison of model parameter counts, AUC, and ACC on the CBIS-DDSM dataset.

Model	Params	View	AUC (%)	ACC (%)
MV-Swin-T	19.4 M	Single	65.08	63.64
CTMA	23.8 M	Single	67.92	64.93
MCCV-Net (Ours)	35.3 M	Single	68.83	66.29
MorphHR-ResNet18	33.8 M	Single	72.57	69.84
OpenCLIP-G/14	33.8 M	Single	74.32	72.70
MV-Swin-T	28.8 M	Multi	66.12	65.69
CTMA	34.8 M	Multi	68.48	66.06
MCCV-Net (Ours)	58.2 M	Multi	78.63	76.31

take single-view mammograms as input. Among the compared methods, Mamba-Vision achieves the best performance, with an AUC of 67.83% and an ACC of 65.80%. Compared with the second-best model, ResNet-50, Mamba-Vision improves AUC and ACC by 0.42 and 0.78 percentage points, respectively. These results indicate that Mamba-based architectures have strong feature extraction capability for mammographic image classification, outperforming both CNN-based and Transformer-based baselines under the same single-view setting.

4.4.2 Comparisons with State-of-the-Art Methods on CBIS-DDSM

We further compare the proposed MCCV-Net with state-of-the-art methods under both single-view and multi-view settings on the CBIS-DDSM dataset. The evaluation metrics include classification accuracy (ACC) and the area under the receiver operating characteristic curve (AUC). In addition, model parameter counts are reported

to provide a reference for computational complexity. The results are summarized in Table 3.

In the single-view setting, MV-Swin-T [49] achieves an AUC of 65.08% and an ACC of 63.64%. CTMA [50] obtains an AUC of 67.92% with 23.8 M parameters, showing competitive parameter efficiency. MorphHR-ResNet18 [51] further improves the AUC to 72.57% by exploiting high-resolution feature modeling. The single-view version of MCCV-Net achieves an AUC of 68.83% and an ACC of 66.29%, outperforming MV-Swin-T and CTMA. This validates the effectiveness of the proposed hybrid Mamba-CNN feature extraction architecture.

The comparison under the multi-view setting reveals more substantial advantages of the proposed model. For MV-Swin-T and CTMA, the gains from multi-view input are relatively limited, with AUC improvements of 1.04 and 0.56 percentage points, respectively. In contrast, MCCV-Net shows a remarkable improvement when dual-view information is incorporated. Its AUC increases from 68.83% in the single-view setting to 78.63% in the multi-view setting, yielding a gain of 9.80 percentage points. Similarly, its ACC increases from 66.29% to 76.31%, corresponding to an improvement of 10.02 percentage points. These results demonstrate that the proposed dual-view feature fusion module can effectively integrate complementary information from the MLO and CC views, thereby improving diagnostic robustness.

Although MCCV-Net has a higher parameter count in the multi-view setting, its performance improvement is substantial. Specifically, the multi-view MCCV-Net achieves the highest AUC and ACC among all multi-view methods. This indicates that the additional model complexity is justified by the significant performance gain obtained through the proposed dual-view fusion strategy.

Table 4 Comparison of model parameter counts, AUC, and ACC on the VinDr-Mammo dataset.

Model	Params	View	AUC (%)	ACC (%)
MV-Swin-T	19.4 M	Single	93.21	92.39
CTMA	23.8 M	Single	94.54	93.28
ImCSOA-ELM2	58.1 M	Single	88.23	86.04
MCCV-Net (Ours)	35.3 M	Single	95.94	95.11
MV-Swin-T	28.8 M	Multi	95.40	94.44
CTMA	34.8 M	Multi	96.18	95.19
MCCV-Net (Ours)	58.2 M	Multi	97.66	96.93

4.4.3 Comparison with State-of-the-Art Methods on VinDr-Mammo

As shown in Table 4, we compare different methods on the VinDr-Mammo dataset under both single-view and multi-view settings. Among these methods, MV-Swin-T, CTMA, and MCCV-Net support both single-view and multi-view configurations, whereas ImCSOA-ELM2 [52] is evaluated as a single-view architecture.

In the single-view setting, MV-Swin-T achieves an AUC of 93.21% and an ACC of 92.39%. CTMA obtains improved performance, with an AUC of 94.54% and an ACC of 93.28%, while maintaining a relatively small parameter count. ImCSOA-ELM2 achieves comparatively lower performance, with an AUC of 88.23% and an ACC of 86.04%. This may be attributed to the sensitivity of chaotic optimization strategies to initial conditions and control parameters, which can affect convergence stability and lead to suboptimal classification performance.

The proposed MCCV-Net achieves the best single-view performance, with an AUC of 95.94% and an ACC of 95.11%. This result indicates that the hybrid Mamba-CNN architecture provides strong representational capability for full-field mammogram classification.

In the multi-view setting, the performance of all models is further improved. The AUC of MV-Swin-T increases from 93.21% to 95.40%, while that of CTMA increases

Table 5 Ablation experiment of Dual-View Consistency Loss on classification performance across CBIS-DDSM and VinDr-Mammo datasets.

Model	Image size	DVCL	Dataset	AUC (%)	ACC (%)
MCCV-Net	224×224	×	CBIS-DDSM	65.27	63.75
MCCV-Net	224×224	✓	CBIS-DDSM	67.95	66.39
MCCV-Net	382×382	×	CBIS-DDSM	71.29	69.42
MCCV-Net	382×382	✓	CBIS-DDSM	78.63	76.31
MCCV-Net	224×224	×	VinDr-Mammo	94.33	93.45
MCCV-Net	224×224	✓	VinDr-Mammo	95.80	94.66
MCCV-Net	382×382	×	VinDr-Mammo	95.04	94.23
MCCV-Net	382×382	✓	VinDr-Mammo	97.66	96.93

from 94.54% to 96.18%. MCCV-Net achieves the highest performance, with an AUC of 97.66% and an ACC of 96.93%. Compared with its single-view version, the multi-view MCCV-Net improves AUC and ACC by 1.72 and 1.82 percentage points, respectively. These results demonstrate the effectiveness of the proposed dual-view fusion mechanism in exploiting complementary information from paired mammographic views.

4.5 Ablation Experiment

Ablation studies were conducted to evaluate the contribution of the proposed Dual-View Consistency Loss (DVCL) and its interaction with input resolution. As shown in Table 5, DVCL consistently improves classification performance on both CBIS-DDSM and VinDr-Mammo.

On the CBIS-DDSM dataset, when the input size is 224×224, introducing DVCL improves AUC from 65.27% to 67.95% and ACC from 63.75% to 66.39%. When the input size is increased to 382×382, the performance gain becomes more pronounced. Specifically, DVCL improves AUC from 71.29% to 78.63%, corresponding to an increase of 7.34 percentage points, and improves ACC from 69.42% to 76.31%, corresponding to an increase of 6.89 percentage points. These results indicate that

high-resolution inputs provide richer anatomical and pathological details, allowing DVCL to better enforce cross-view consistency between the MLO and CC views.

Similar improvements can be observed on the VinDr-Mammo dataset. At the input size of 224×224, DVCL improves AUC from 94.33% to 95.80% and ACC from 93.45% to 94.66%. At the input size of 382×382, DVCL further improves AUC from 95.04% to 97.66% and ACC from 94.23% to 96.93%. These results demonstrate that DVCL is effective across different datasets and that its benefit becomes more substantial when higher-resolution mammograms are used.

The effect of image resolution is also evident. For CBIS-DDSM, increasing the input resolution from 224×224 to 382×382 improves performance both with and without DVCL. Without DVCL, the AUC increases from 65.27% to 71.29%; with DVCL, the AUC increases from 67.95% to 78.63%. This suggests that higher-resolution inputs are particularly important for CBIS-DDSM, where lesion patterns may be subtle and local details are critical for classification.

For VinDr-Mammo, increasing the input size also improves performance. Without DVCL, the AUC increases from 94.33% to 95.04%, while with DVCL, it increases from 95.80% to 97.66%. This indicates that larger input images preserve more fine-grained pathological cues and provide more reliable feature representations for dual-view consistency learning.

Overall, the ablation results confirm that DVCL plays an important role in improving MCCV-Net by encouraging consistent representation learning between paired mammographic views. Moreover, the combination of high-resolution input and DVCL yields the best performance on both datasets, demonstrating the importance of preserving detailed anatomical information for full-field mammogram classification.

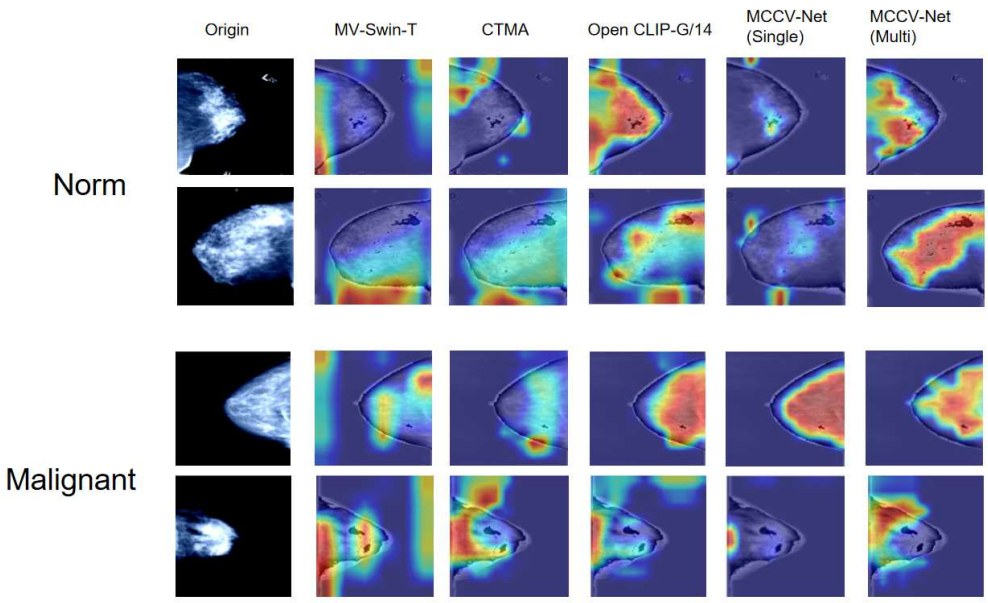

Fig. 7 Feature attention maps of different models on MLO and CC mammographic images.

4.6 Heatmap Visualization Experiment

As shown in Fig. 7, we conducted a heatmap visualization experiment to compare the feature attention patterns of different models on key breast regions. The heatmaps were generated by computing the gradients of features from critical network layers and using the resulting weights to identify image regions that contribute most to the final prediction. These attention weights were then mapped back to the original mammograms to provide an intuitive visual interpretation of model behavior.

In this experiment, representative MLO and CC images, including both normal and malignant samples, were selected to evaluate the region-discriminative capability of different models. The visualization results reveal distinct attention patterns among the compared methods. For normal samples, MV-Swin-T exhibits scattered and localized attention, with excessive responses in irrelevant background regions. For malignant samples, its attention does not consistently concentrate on the true lesion regions and instead disperses across non-abnormal breast tissues.

CTMA also shows limited spatial alignment between high-attention regions and actual lesions in malignant samples. In some cases, its attention shifts toward non-lesion tissues near the breast boundary. For normal samples, CTMA still produces unnecessary local hotspots, indicating potential sensitivity to irrelevant image patterns.

Compared with MV-Swin-T and CTMA, OpenCLIP-G/14 presents more concentrated attention in normal samples and smaller deviations from lesion regions in malignant samples. However, it still suffers from attention dispersion and occasional unwarranted local activation. The single-view version of MCCV-Net produces smoother and more evenly distributed attention maps, with less focus on irrelevant regions. Nevertheless, its overall attention intensity is relatively weak, and the coverage of clinically relevant regions remains insufficient.

In contrast, MCCV-Net (multi-view) demonstrates the most desirable visualization results. For malignant samples, the red and yellow regions in the heatmaps highly coincide with the actual lesion areas in the original images, indicating accurate and concentrated lesion localization. For normal samples, its attention is smoothly distributed over breast tissues without producing unnecessary local hotspots. These observations suggest that the dual-view fusion mechanism enables the model to suppress irrelevant responses while enhancing attention to diagnostically meaningful regions.

Overall, the heatmap visualization results demonstrate that the proposed MCCV-Net possesses superior lesion localization and regional discriminative capabilities. Compared with other models, MCCV-Net exhibits stronger robustness to background noise and better alignment with clinically relevant abnormal regions.

4.7 Edge Deployment and Computational Efficiency Analysis

Beyond diagnostic accuracy, connected screening requires efficient inference to support concurrent cloud workloads and resource-constrained edge devices. MCCV-Net's dual branches and progressive fusion motivate lightweight design to reduce computation, memory requirements, and model-distribution payloads while preserving complementary diagnostic information. This objective aligns with computation–communication co-design for edge–cloud healthcare, although any reduction in image-transmission traffic or communication latency requires separate system-level evaluation.

To address these deployment requirements, a Knowledge Distillation and View-Asymmetric Student (KDVA-S) uses three full CPMC modules in the MLO branch, three CPMC-Lite modules in the CC branch (width coefficient 0.5), and three fusion stages instead of five. The frozen MCCV-Net teacher provides feature-, fusion-, and output-level supervision alongside ground-truth labels; only the student is retained for inference. Using the same reported data partitions and paired-view protocol, students are trained with 384×384 inputs, AdamW (learning rate 2×10^{-4}, weight decay 10^{-4}), a batch size of 4, and three stages of 30 epochs. Table 6 compares classification performance and summarizes RTX 4090 workstation efficiency; end-to-end latency includes input handling, preprocessing, inference, and result preparation.

Complete KDVA-S improves AUC over the directly trained student by 3.58 and 1.49 percentage points on CBIS-DDSM and VinDr-Mammo, respectively, while narrowing the performance gap with the teacher to 2.50 and 1.91 points. Within the reported efficiency configuration, parameters and FLOPs decrease by 40.0%, peak memory by 31.8%, forward latency by 40.0%, and end-to-end latency by 30.8%, while throughput increases from 19.2 to 27.8 cases/s. Together, the classification results support multi-level distillation for improving the compact student's diagnostic performance, while the efficiency summary indicates lower resource requirements and faster inference for the reported deployment configuration.

Table 6 Classification performance and reported deployment efficiency of the lightweight student.

Metric	MCCV-Net teacher	Student no distillation	Student output distillation	Complete KDVA-S
CBIS-DDSM AUC (%)	78.63	72.55	74.85	76.13
CBIS-DDSM ACC (%)	76.31	71.08	72.36	74.51
VinDr-Mammo AUC (%)	97.66	94.26	94.52	95.75
VinDr-Mammo ACC (%)	96.93	93.57	94.74	95.08
Parameters (M)	30.0	18.0	-	-
FLOPs (G)	16.0	9.6	-	-
Peak GPU memory (MB)	1100	750	-	-
Forward latency (ms)	20	12	-	-
End-to-end latency (ms/case)	52	36	-	-
Reported P95 latency (ms)	78	54	-	-
Throughput (cases/s)	19.2	27.8	-	-

5 Conclusion

In this paper, we proposed MCCV-Net, a dual-view hybrid architecture for full-field mammographic image classification. The proposed framework is designed to emulate the diagnostic process of radiologists by jointly considering complementary craniocaudal and mediolateral oblique views. Different from conventional single-view approaches that process mammograms as isolated two-dimensional images, MCCV-Net explicitly models both local lesion morphology and global anatomical context across paired mammographic views. The core of MCCV-Net lies in the integration of convolutional neural networks and Mamba-based state space modeling. The CNN branch captures local fine-grained pathological patterns, such as mass margins and microcalcification distributions, while the Mamba branch models long-range spatial dependencies and global anatomical structures with linear computational complexity. Furthermore, the proposed dual-view feature fusion module progressively integrates complementary information from MLO and CC views, thereby improving cross-view consistency and diagnostic robustness. The Dual-View Consistency Loss further encourages paired view-specific representations to remain anatomically coherent while preserving discriminative features for classification.

Extensive experiments on CBIS-DDSM and VinDr-Mammo demonstrate the effectiveness of the proposed method. Compared with representative CNN-based, Transformer-based, and Mamba-based models, MCCV-Net achieves superior classification performance under both single-view and multi-view settings. Ablation experiments validate the contribution of the Dual-View Consistency Loss, while heatmap visualization results show that MCCV-Net can better focus on clinically relevant lesion regions and suppress irrelevant background responses.

Despite its promising performance, MCCV-Net introduces additional computational complexity due to its dual-view architecture and progressive fusion mechanism. Future work will focus on improving computational efficiency through lightweight network design, model pruning, and knowledge distillation, while preserving diagnostic accuracy and interpretability. Teacher-assistant-based distillation for heterogeneous models provides a methodological reference for investigating knowledge transfer from the hybrid architecture to a compact student network [53]. This possibility remains to be evaluated for mammographic classification. In addition, further validation on larger multi-center clinical datasets will be explored to improve the generalization and clinical applicability of the proposed framework.

Declarations

Funding. This study was supported by grants from the Shenzhen Big Data Industry Chain Sharing and Service Platform of Laboratory Medicine (Grant No. XMHT20251124003)

Ethics approval and consent to participate. This study uses existing mammography datasets distributed by TCIA and PhysioNet, rather than newly collected clinical data [45, 47]. The VinDr-Mammo repository reports that the Institutional Review Boards of Hanoi Medical University Hospital and Hospital 108 approved the original data collection and release of de-identified data, and that patient informed

consent was waived for that original study [47]. These approvals concern the original dataset and do not establish the ethics approval or exemption status of the present secondary analysis.

Author confirmation required before submission: Specify the institutional determination for this secondary analysis, including the reviewing institution or committee, approval or exemption reference where applicable, and the applicable consent requirement or waiver. Public availability of the datasets alone is not used here as a basis for asserting exemption.

Consent for publication. Author confirmation required before submission: Confirm that the manuscript and example mammograms contain no identifiable participant information and that publication of the selected images is permitted by the applicable dataset terms. If this is confirmed and no publication consent is required, replace this note with the appropriate not-applicable statement; otherwise provide the relevant consent details.

Author contributions. **Kun Xu**: Writing—original draft, Writing—review and editing, Methodology, Visualization, Investigation. **Pin-Han Ho**: Writing—review and editing, Conceptualization, Methodology, Resources, Supervision. **Jun Chen**: Conceptualization, Data curation, Resources, Supervision, Funding acquisition. **Limei Peng**: Writing—review and editing, Methodology, Visualization. **Boyang Deng**: Writing—review and editing, Supervision.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process. During manuscript preparation, a generative AI assistant was used to assist with language revision, restructuring of the abstract and introduction, and reference

discovery and formatting. This assistance did not generate the experimental results reported in this article. The authors remain responsible for verifying the accuracy of the text, references, and declarations before submission.

Availability of data and materials. The datasets supporting the conclusions of this article are available in The Cancer Imaging Archive (TCIA) repository for CBIS-DDSM, https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY [45], and in the PhysioNet repository for VinDr-Mammo version 1.0.0, https://doi.org/10.13026/br2v-7517 [47]. CBIS-DDSM is distributed under the Creative Commons Attribution 3.0 license. Access to VinDr-Mammo requires registration and acceptance of the applicable PhysioNet Restricted Health Data Use Agreement; use is subject to the PhysioNet Restricted Health Data License. The corresponding dataset descriptor publications are also included in the reference list [46, 48].

Code availability. The source-code repository is omitted for double-blind review and will be made available after the review process.

References

[1] Su, Y., Liu, Q., Xie, W., Hu, P.: Yolo-logo: A transformer-based yolo segmentation model for breast mass detection and segmentation in digital mammograms. Computer Methods and Programs in Biomedicine **221**, 106903 (2022)

[2] Chen, J., Pan, T., Zhu, Z., Liu, L., Zhao, N., Feng, X., Zhang, W., Wu, Y., Cai, C., Luo, X., *et al.*: A deep learning-based multimodal medical imaging model for breast cancer screening. Scientific Reports **15**(1), 14696 (2025)

[3] Lee, W., Lee, H., Lee, H., Park, E.K., Nam, H., Kooi, T.: Transformer-based deep neural network for breast cancer classification on digital breast tomosynthesis images. Radiology: Artificial Intelligence **5**(3), 220159 (2023)

[4] Lin, Q., Tan, W.-M., Ge, J.-Y., Huang, Y., Xiao, Q., Xu, Y.-Y., Jin, Y.-T., Shao, Z.-M., Gu, Y.-J., Yan, B., et al.: Artificial intelligence-based diagnosis of breast cancer by mammography microcalcification. Fundamental research (2023)

[5] Ilyas, Z., Saleem, A., Suter, D., Schousboe, J.T., Leslie, W.D., Lewis, J.R., Gilani, S.Z.: A hybrid cnn-transformer feature pyramid network for granular abdominal aortic calcification detection from dxa images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 14–25 (2024). Springer

[6] Huang, R., Xu, Z., Xie, Y., Wu, H., Li, Z., Cui, Y., Huo, Y., Han, C., Yang, X., Liu, Z., *et al.*: Joint-phase attention network for breast cancer segmentation in dce-mri. Expert Systems with Applications **224**, 119962 (2023)

[7] Huang, R., Lin, M., Dou, H., Lin, Z., Ying, Q., Jia, X., Xu, W., Mei, Z., Yang, X., Dong, Y., *et al.*: Boundary-rendering network for breast lesion segmentation in ultrasound images. Medical image analysis **80**, 102478 (2022)

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)

[9] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., *et al.*: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12009–12019 (2022)

[10] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16133–16142 (2023)

[11] Pan, J., Bulat, A., Tan, F., Zhu, X., Dudziak, L., Li, H., Tzimiropoulos, G., Martinez, B.: Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In: European Conference on Computer Vision, pp. 294–311 (2022). Springer

[12] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12179–12188 (2021)

[13] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 103280 (2024)

[14] Li, B., Zhou, J., Gou, F., Wu, J.: TransRNetFuse: A highly accurate and precise boundary FCN-transformer feature integration for medical image segmentation. Complex & Intelligent Systems **11**(5), 213 (2025) https://doi.org/10.1007/s40747-025-01847-3

[15] Qiao, S., Pang, S., Luo, G., Sun, Y., Yin, W., Pan, S., Lv, Z.: DPC-MSGATNet: Dual-path chain multi-scale gated axial-transformer network for four-chamber view segmentation in fetal echocardiography. Complex & Intelligent Systems **9**(4), 4503–4519 (2023) https://doi.org/10.1007/s40747-023-00968-x

[16] Li, G., Wang, Y., Zhao, Q., Yuan, P., Chang, B.: PMVT: a lightweight vision transformer for plant disease identification on mobile devices. Frontiers in Plant Science **14**, 1256773 (2023) https://doi.org/10.3389/fpls.2023.1256773

[17] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First Conference on Language Modeling (2024)

[18] Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)

[19] Zhu, X., Wang, W., Zhang, C., Wang, H.: Polyp-mamba: A hybrid multi-frequency perception gated selection network for polyp segmentation. Information Fusion **115**, 102759 (2025)

[20] Wu, R., Liu, Y., Liang, P., Chang, Q.: H-vmunet: High-order vision mamba unet for medical image segmentation. Neurocomputing **624**, 129447 (2025)

[21] Xu, Y., Yang, H., Jiang, Z., Shi, W., Miao, Y., Feng, G., Li, Y.: MGASM-Net: Morphology-guided multi-task learning network with anatomic spatial mamba for 3D airway segmentation. Complex & Intelligent Systems **11**(8), 360 (2025) https://doi.org/10.1007/s40747-025-01995-6

[22] Jiménez-Sánchez, A., Tardy, M., Ballester, M.A.G., Mateus, D., Piella, G.: Memory-aware curriculum federated learning for breast cancer classification. Computer Methods and Programs in Biomedicine **229**, 107318 (2023)

[23] Zhang, J., Xu, Y., Li, S., Liang, F., Duan, H., Dong, Y., Leung, V.C.M., Hu, X.: FedSM: Semantic-guided feature mixup for bias reduction in federated learning with long-tail data. IEEE Internet of Things Journal **13**(6), 12491–12500 (2026) https://doi.org/10.1109/JIOT.2026.3652363

[24] Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 296–306 (2024)

[25] Pei, X., Huang, T., Xu, C.: Efficientvmamba: Atrous selective scan for light weight visual mamba. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 6443–6451 (2025)

[26] Tran, T.-T., Truong, G.-B., Than, N.-L., Hoang, N.-K., Nguyen, D.-H., Pham, V.-T., et al.: A dense attention mamba-based network with adaptive sigmoid fowlkes-mallows loss for enhanced medical image segmentation. Image and Vision Computing, 105778 (2025)

[27] Tang, F., Nian, B., Li, Y., Jiang, Z., Yang, J., Liu, W., Zhou, K.S.: Mambamim: Pre-training mamba with state space token interpolation and its application to medical image segmentation. Medical Image Analysis **103**, 103606 (2025)

[28] Dang, T.D.Q., Nguyen, H.H., Tiulpin, A.: Log-vmamba: local-global vision mamba for medical image segmentation. In: Proceedings of the Asian Conference on Computer Vision, pp. 548–565 (2024)

[29] Tang, H., Huang, Z., Li, W., Wu, Y., Yuan, J., Yang, Y., Zhang, Y., Qin, J., Zheng, H., Liang, D., et al.: Automatic brain segmentation for pet/mr dual-modal images through a cross-fusion mechanism. IEEE Journal of Biomedical and Health Informatics (2024)

[30] Yue, Y., Li, Z.: Medmamba: Vision mamba for medical image classification. arXiv preprint arXiv:2403.03849 (2024)

[31] Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Scientific reports **8**(1), 4165 (2018)

[32] Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. arXiv preprint arXiv:1710.01766 (2017)

[33] Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., *et al.*: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE transactions on medical imaging **39**(4), 1184–1194 (2019)

[34] Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., *et al.*: An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Medical image analysis **68**, 101908 (2021)

[35] Sun, L., Han, B., Jiang, W., Liu, W., Liu, B., Tao, D., Yu, Z., Li, C.: Multi-scale region selection network in deep features for full-field mammogram classification. Medical Image Analysis **100**, 103399 (2025)

[36] Larroza, A., Pérez-Benito, F.J., Tendero, R., Perez-Cortes, J.C., Román, M., Llobet, R.: Breast delineation in full-field digital mammography using the segment anything model. Diagnostics **14**(10), 1015 (2024)

[37] Liao, J., Duan, H., Feng, K., Zhao, W., Yang, Y., Chen, L., Chen, Y.: LR-ASD: Lightweight and robust network for active speaker detection. International Journal of Computer Vision **133**(7), 4749–4769 (2025) https://doi.org/10.1007/s11263-025-02399-2

[38] Narayanam, L.P.R., Rao, T.N., Kumar, D.S.: Swincamf-net: Explainable cross-attention multimodal swin network for mammogram analysis. Diagnostics **15**(23), 3037 (2025)

[39] Feng, W., Zhu, Y., Zhang, R., Wang, C., Ma, F., Wang, X., Li, X.: Active multimodal distillation for few-shot action recognition. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp.

999–1007 (2025). https://doi.org/10.24963/ijcai.2025/112

[40] Zuo, Y., Zhang, R., Guo, X., Kang, C., Lyu, L.: Decoupling mean teacher via dual students: A prototype-based framework for semi-supervised medical segmentation. Complex & Intelligent Systems **11**(8), 353 (2025) https://doi.org/10.1007/s40747-025-01964-z

[41] Liu, S., Zou, M., Liu, N., Li, Y., Zheng, W.: A teacher-guided early-learning method for medical image segmentation from noisy labels. Complex & Intelligent Systems **10**(6), 8011–8026 (2024) https://doi.org/10.1007/s40747-024-01574-1

[42] Telçeken, M.: A new hybrid vision transformer model for medical image classifications: Attention mechanism deepened by hog. Biomedical Signal Processing and Control **114**, 109381 (2026)

[43] Liu, X., Sun, L., Li, C., Han, B., Jiang, W., Yuan, T., Liu, W., Liu, Z., Yu, Z., Liu, B.: Lesion asymmetry screening assisted global awareness multi-view network for mammogram classification. IEEE Transactions on Medical Imaging (2025)

[44] Al Mansour, A.G., Alshomrani, F., Alfahaid, A., Almutairi, A.T.: Mammovit: A custom vision transformer architecture for accurate birads classification in mammogram analysis. Diagnostics **15**(3), 285 (2025)

[45] Sawyer-Lee, R., Gimenez, F., Hoogi, A., Rubin, D.: Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM). The Cancer Imaging Archive. Dataset. Accessed 1 November 2026 (2016). https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY . https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY

[46] Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis

research. Scientific Data **4**, 170177 (2017) https://doi.org/10.1038/sdata.2017.177

[47] Pham, H.H., Nguyen Trung, H., Nguyen, H.Q.: VinDr-Mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography. PhysioNet. Dataset, version 1.0.0. Accessed 1 November 2026 (2022). https://doi.org/10.13026/br2v-7517 . https://physionet.org/content/vindr-mammo/1.0.0/

[48] Nguyen, H.T., Nguyen, H.Q., Pham, H.H., Lam, K., Le, L.T., Dao, M., Vu, V.: VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Scientific Data **10**, 277 (2023) https://doi.org/10.1038/s41597-023-02100-7

[49] Sarker, S., Sarker, P., Bebis, G., Tavakkoli, A.: Mv-swin-t: Mammogram classification with multi-view swin transformer. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2024). IEEE

[50] Van Tulder, G., Tong, Y., Marchiori, E.: Multi-view analysis of unregistered medical images using cross-view transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 104–113 (2021). Springer

[51] Wei, T., Aviles-Rivero, A.I., Wang, S., Huang, Y., Gilbert, F.J., Schönlieb, C.-B., Chen, C.W.: Beyond fine-tuning: Classifying high resolution mammograms using function-preserving transformations. Medical image analysis **82**, 102618 (2022)

[52] Bharanidharan, N., SR, S.C., et al.: Optimized extreme learning machine for breast tumor classification using mammograms. IEEE Open Journal of the Computer Society (2025)

[53] Jiang, R., Su, T., Si, J., Wang, C., Feng, W., Zhu, J., Yuan, P.: Teacher assistant-based knowledge distillation bridging architecture differences on heterogeneous models. Neurocomputing **703**, 134697 (2026) https://doi.org/10.1016/j.neucom.2026.134697