PeerJ Computer Science

Applying vision transformer to assess multi-scale morphological features in mammography for breast cancer detection: multiscale image morphological extraction vision transformer (MIME-ViT)

Yuki Kashiwada[1,2], Eichi Takaya[3,4], Mei Hiroya[3], Nanako Matsuda[3], Takumi Yashima[3], Tomoya Kobayashi[3,4], Gen Tamiya[1,2,5] and Takuya Ueda[6]

[1] Graduate School of Medicine, Tohoku University, Sendai, Japan
[2] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[3] Department of Clinical Imaging, Tohoku University, Sendai, Japan
[4] AI Lab, Tohoku University Hospital, Sendai, Japan
[5] Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan
[6] Department of Diagnostic Radiology, Tohoku University, Sendai, Japan

ABSTRACT

Background: Breast cancer screening using mammography often suffers from low sensitivity and specificity, particularly in dense breast tissue. This limitation can result in missed diagnoses and unnecessary procedures. The evolution of deep learning models, such as those based on convulational neural networks (CNNs) and Vision Transformers (ViTs), presents opportunities for significant advancements.
Methods: This study utilized the Chinese Mammography Database (CMMD) and enhanced it with detailed annotations from two radiologists for detection tasks. The Multiscale Image Morphological Extraction Vision Transformer (MIME-ViT) model, which integrates ViT and CNN, is designed to capture multiscale morphological features from mammographic images. Training of the model prioritized segmentation and classification, employing a combination of Dice and Focal losses to effectively tackle detection tasks.
Results: Without pre-training, MIME-ViT achieved a mean Intersection over Union (IoU) of 0.3342 across all images, 0.3797 for mass, and 0.2491 for calcification. In terms of IoU scores, MIME-ViT's performance was inferior to that of Detection Transformer (DETR) with pre-training, yet it surpassed the performance of DETR without pre-training.
Conclusions: By merging Vision Transformers with CNNs to enhance mammographic imaging analysis, the MIME-ViT model represents a significant advancement in breast cancer detection. This development marks a critical step forward in medical imaging technology, with the goal of improving early detection rates and patient outcomes. As medical imaging technology continues to evolve, MIME-ViT emerges as a key innovation, paving the way for more effective and advanced cancer screening methodologies.

Submitted 19 February 2025
Accepted 8 September 2025
Published 15 October 2025

Corresponding author
Eichi Takaya,
eichi.takaya.d5@tohoku.ac.jp

Academic editor
Ahmed Elazab

Additional Information and Declarations can be found on page 14

DOI **10.7717/peerj-cs.3252**

How to cite this article Kashiwada Y, Takaya E, Hiroya M, Matsuda N, Yashima T, Kobayashi T, Tamiya G, Ueda T. 2025. Applying vision transformer to assess multi-scale morphological features in mammography for breast cancer detection: multiscale image morphological extraction vision transformer (MIME-ViT). **PeerJ Comput. Sci. 11:e3252 DOI 10.7717/peerj-cs.3252**

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Computer Vision, Optimization Theory and Computation, Neural Networks
Keywords Breast cancer detection, Mammography, Deep learning, Vision transformer (ViT), Convolutional neural networks (CNN), Multiscale analysis, Medical imaging

INTRODUCTION

Breast cancer is the most common cancer among women globally, and both its incidence and mortality rates are projected to rise (*Harbeck & Gnant, 2017*; *Anastasiadi et al., 2017*). Mammography screening programs are in place in various countries for early breast cancer detection and treatment (*Myers et al., 2015*). Randomized controlled studies have shown that mammography can reduce breast cancer-related mortality by approximately 20% (*Nelson et al., 2009*). Despite its efficacy, mammography-based screening has limitations, including low sensitivity, especially in patients with dense breast tissue. This may result in missed diagnoses, leading to delayed treatment (*Nelson et al., 2009*). Additionally, the high rate of false positives is also a significant issue, causing unnecessary stress and potentially triggering unwarranted medical interventions (*Nelson et al., 2009*).

During the clinical interpretation of mammography images, radiologists commonly use a multiscale evaluative approach to diagnose breast cancer (*Fowler et al., 2013*). This approach encompasses the assessment of features at both micro and macro scales. On the micro-scale, features like the shape of individual calcifications and minute morphological changes at tumor peripheries are examined. On the macro-scale, the distribution pattern of these calcifications and the overall tumor shape are considered. Such multiscale information is then synthesized to determine a comprehensive radiological diagnosis.

In recent years, artificial intelligence (AI)-based diagnostic systems have made significant advancements in breast imaging, surpassing the performance of traditional computer-aided detection systems (*Lehman et al., 2015*). Although conventional convulational neural network (CNN)-based deep learning (DL) algorithms have proven useful in breast cancer diagnostic systems (*Raya-Povedano et al., 2021*; *van Leeuwen et al., 2022*; *Shoshan et al., 2022*; *Mendelson, 2019*), their limitations have also been pointed out (*Nassif et al., 2022*; *Cai et al., 2023*). Furthermore, object detection techniques have shown promise as preprocessing steps for automated region-of-interest identification. For instance, *Chen et al. (2023)* demonstrated a You Only Look Once (YOLO)-based adaptive multiscale system that combines YOLOv4 for calcification localization with an ensemble classifier for malignancy assessment, achieving improved benign/malignant classification on spot magnification mammograms (area under the curve (AUC) 0.888) and potentially reducing unnecessary biopsies by over 80%. CNNs process image data through the use of localized filters in convolutional layers, and are particularly effective for identifying intricate details within a specific region of an image. Recently, the Vision Transformer (ViT) has been introduced as a novel architecture for DL-based image analysis (*Dosovitskiy et al., 2020*; *Shamshad et al., 2023*; *Azad et al., 2024*). ViT is adapted from the Transformer model, which was originally designed for natural language processing, and is designed to capture a wide range of positional relationships within the image. Unlike CNNs that process images through localized receptive fields with limited global context, ViTs employ

self-attention mechanisms that enable direct modeling of long-range dependencies across the entire image (*Vaswani et al., 2017*; *Han et al., 2022*). This global receptive field capability is particularly advantageous for mammographic analysis, where subtle calcification patterns may be distributed across distant regions of the breast, and their spatial relationships provide crucial diagnostic information (*Raghu et al., 2021*). Furthermore, the self-attention mechanism allows ViTs to dynamically weight the importance of different image regions based on their relevance to the diagnostic task, making them well-suited for medical imaging where pathological features may vary significantly in size and location (*Chen, Fan & Panda, 2021*). *Chen et al. (2022)* proposed a specialized approach for pathological imaging called Hierarchical Image Pyramid Transformer (HIPT). The HIPT involves the extraction of hierarchical image features at multiple scales to input into the ViT architecture (*Chen et al., 2022*), which allows the model to reflect complex patterns and relationships across different image scales. A recent comprehensive review (*Singh & Patnaik, 2024*) has systematized the evolution from traditional CNN to ViT approaches in breast cancer detection systems. However, research incorporating multiscale morphological information remains limited.

The application of HIPT has been increasingly adopted in medical image analysis, particularly in digital pathology for Whole Slide Image (WSI) analysis, where hierarchical Transformers are employed to integrate local tissue patterns with global contextual information across the entire slide (*Shoshan et al., 2024*; *Guo et al., 2023*). Notably, *Chen et al. (2022)* developed a HIPT model that utilizes self-supervised learning to leverage the hierarchical structure of WSI data through a two-stage Vision Transformer pre-training approach, combined with a weakly-supervised ViT classifier to extract high-level feature representations from over 10,000 pathological WSIs derived from The Cancer Genome Atlas (TCGA) (*Contreras et al., 2024*). While such applications have shown considerable progress in pathological imaging and other medical imaging domains, the application of HIPT to mammography remains relatively limited and represents an area of significant potential for advancing breast cancer detection.

The purpose of our research is to develop a specialized DL model to detect breast cancer by combining ViT and CNN algorithms, facilitating the extraction of multiscale image features in mammography.

MATERIALS AND METHODS

Dataset

In this study, we utilized the TOMPEI-CMMD dataset (https://www.cancerimagingarchive.net/analysis-result/tompei-cmmd/) (*Kashiwada et al., 2024*), derived from the Chinese Mammography Database (CMMD) (*Cui et al., 2021*). TOMPEI-CMMD extends the CMMD by incorporating lesion segmentation masks and corrections to certain lesion annotations. CMMD is a publicly accessible mammography database comprising data from 1,775 Chinese patients who underwent mammographic examinations from July 2012 to January 2016. All mammographic images were captured using digital mammography, with a resolution of 2,294 × 1,914 pixels. Figure 1 illustrates the dataset selection process from the CMMD dataset. From the 1,775 patients in CMMD,

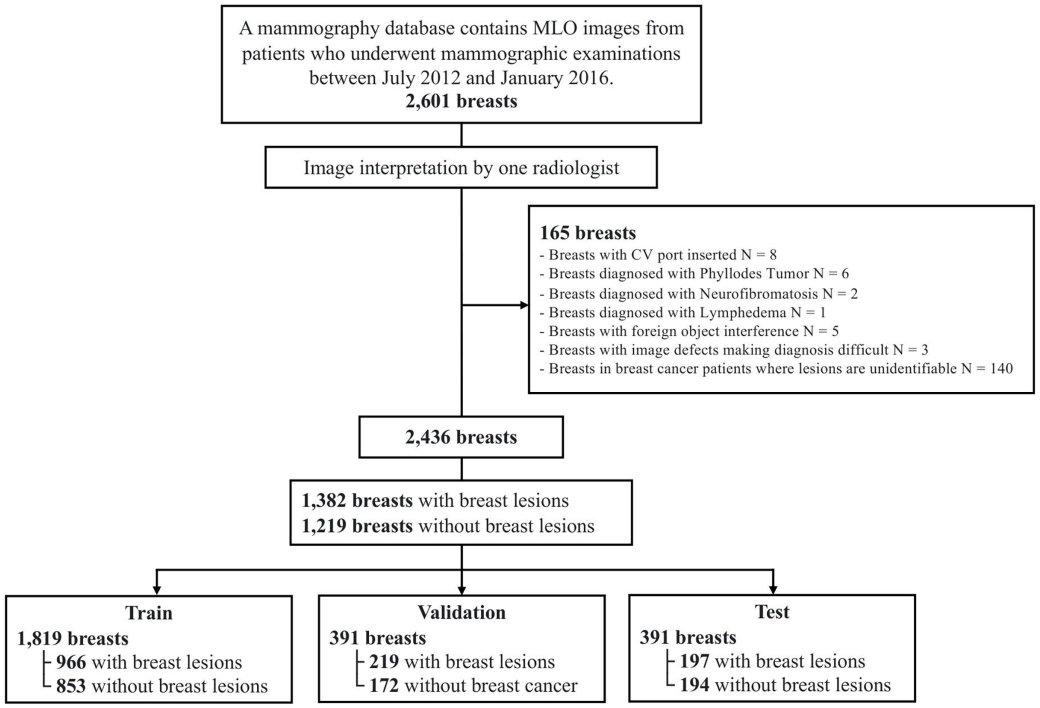

Figure 1 **Dataset selection workflow from the original CMMD dataset.**
Full-size DOI: 10.7717/peerj-cs.3252/fig-1

826 had bilateral mammograms, while 949 had unilateral mammograms, totaling 2,601 breast mammograms. The CMMD dataset encompasses both mediolateral oblique (MLO) and craniocaudal (CC) views. For this study, only MLO views were utilized. Furthermore, we utilized the lesion labels and segmentation masks available in the TOMPEI-CMMD dataset.

Image data processing

Black pixel padding was applied to the distal sides of the body in each image, transforming the original 2,294 × 1,914 pixel images into 2,294 × 2,294 isotropic images with uniform dimensions. To meet the matrix size requirements of the multiscale deep learning model, the 2,294 × 2,294 pixel images were resized to 2,048 × 2,048 pixels.

Multiscale Image Morphological Extraction Vision Transformer (MIME-ViT) architecture

Figure 2 illustrates the architecture of our proposed DL model, the Multiscale Image Morphological Extraction Vision Transformer (MIME-ViT). MIME-ViT integrates the architectural characteristics of ViT and CNN in a hybrid design to capture multiscale morphological features of breast cancer in mammographic images. The selection of ViT as the foundation for multiscale analysis is motivated by its superior capability in handling hierarchical feature representations across different scales (*Liu et al., 2021*; *Wang et al., 2022*). Unlike traditional CNN architectures that require explicit multiscale designs

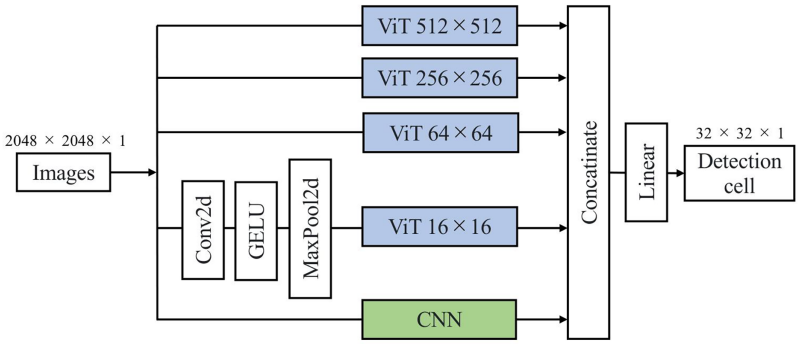

Figure 2 MIME-ViT model architecture proposed for enhanced breast cancer detection in mammograms in this research. It uses a hybrid design to capture multiscale morphological features, with components for specific scale ranges from 512 × 512 to 16 × 16. This facilitates the extraction of both macro- and micro-morphological details of cancerous formations. The integration of multiscale processing and a loss function combining Dice and Focal losses enables MIME-ViT to effectively balance the detection of broad structures and fine details, thereby optimizing its performance in breast cancer detection. Full-size DOI: 10.7717/peerj-cs.3252/fig-2

through feature pyramid networks, ViTs naturally excel at multiscale analysis through their inherent ability to model both local and global dependencies simultaneously within the self-attention mechanism (*Yuan et al., 2021*). Within the ViT segment, discrete components are specifically designed to process image patches at various scales: 512 × 512, 256 × 256, 64 × 64, and 16 × 16. The 512 × 512 and 256 × 256 components are designed to extract macroscale morphological features, like the overall shape of a breast cancer mass or the segmental distribution of microcalcifications. The 64 × 64 and 16 × 16 components are intended to discern detailed morphological characteristics, such as the marginal irregularity of a breast cancer mass or the morphology of microcalcifications. Due to graphics processing unit (GPU) memory capacity limitations, the 16 × 16 ViT component undergoes convolution and pooling operations before being introduced into the ViT, ensuring effective data dimension reduction. Following processing through these four multiscale ViT components and an additional CNN component, the architecture synthesizes the information and outputs it as a 36 × 36 patch. The architectural design of MIME-ViT ensures proficiency in detecting features across various scales, from broad structures to intricate details, thus enhancing its capability to analyze breast cancer images. The architectural design of MIME-ViT, along with the accompanying code, is available for research purposes on GitHub (https://github.com/javasparrows/MIME-ViT and archived at https://doi.org/10.5281/zenodo.16221703).

To validate the architectural design choices of MIME-ViT, we conducted systematic ablation studies. First, we evaluated the contribution of each ViT scale component by systematically removing individual components (ViT-512, ViT-256, or ViT-64) and measuring the performance impact on lesion detection accuracy. This approach allowed us to verify that each scale captures complementary information essential for accurate detection.

Second, we compared different CNN kernel sizes within the ResidualBlock components. The standard 3×3 convolutional kernels were tested against 5×5 kernels. The 3×3 configuration enabled the use of pretrained weights from established computer vision models, while the 5×5 configuration required random initialization due to dimensional incompatibility with existing pretrained models.

Finally, we investigated two feature fusion strategies for integrating multi-scale information. The initial approach employed additive fusion where features from different scales were combined through weighted summation ($x = x_{512} + x_{256} + x_{64} + x_{conv} \times 20$). The alternative approach used concatenation to preserve individual scale-specific features before final processing, allowing the model to learn optimal integration weights during training rather than enforcing a predetermined combination scheme.

Loss function

The loss function for MIME-ViT is defined as a combination of Dice loss and Focal loss (*Lin et al., 2017*). Dice loss is defined as $(1 - \text{Dice score})$ (*Sorensen, 1948*; *Dice, 1945*), evaluating the overlap between the predicted segmentation mask and the actual ground truth. Meanwhile, Focal loss is defined as Eq. (1), prioritizing pixels that are more challenging to classify, thereby ensuring the model sufficiently attends to them. Term 2, as described in Eq. (2), represents the cross-entropy loss. By integrating both Focal loss and Dice loss, MIME-ViT is designed to address precise and refined detection tasks while managing the inherent imbalances between unmasked and masked pixels within the patch.

$$FL(p_t) = -(1 - p_t)^{\gamma} \log p_t \tag{1}$$

$$p_t = \begin{cases} p & \text{if } y = 1, \\ 1 - p & \text{otherwise} \end{cases}. \tag{2}$$

Model training and implementation

The DL dataset, comprising 2,601 mammography images, was divided into training, validation, and test sets at a 7:1.5:1.5 ratio, yielding 1,819, 391, and 391 images for each set, respectively (Fig. 1). To ensure that MLO images from the same patient's right and left breasts were not split between the training and test sets, dataset partitioning was patient-based. To enhance the diversity and robustness of the trained deep learning model, rotational transformation, with a maximum of 40 degrees, was applied to the training set.

We trained the MIME-ViT model using the Adam optimizer with an initial learning rate of 0.001. Training was conducted for 15 epochs, employing cosine annealing as the learning rate scheduler (Fig. 3). Figure 3A displays the progression of the mean Intersection over Union (IoU) on the training data, while Fig. 3B shows the validation loss. The learning-rate schedule is presented in Figs. 3C, and 3D provides an analysis of the correlation between IoU and loss observed during the training process. To further optimize the training, we also utilized the AdamW optimizer (*Loshchilov & Hutter, 2017*) with a weight decay of 0.01, and Projected Conflicting Gradients (PCGrad) (*Yu et al., 2020*). The principal hyper-parameter settings are summarized in Table 1.

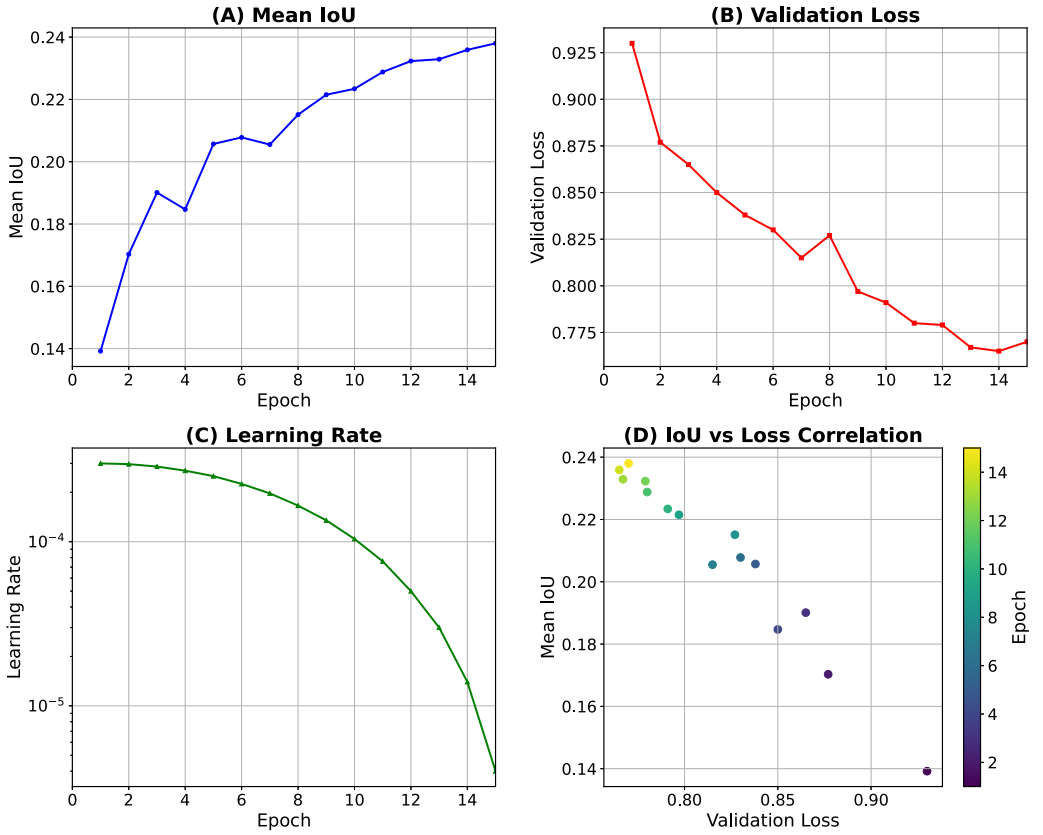

Figure 3 Training dynamics of the MIME-ViT model over 15 epochs. (A) Mean intersection over union (IoU) on the training data. (B) Validation loss. (C) Learning-rate schedule using cosine (*Loshchilov & Hutter, 2016*). (D) Correlation between IoU and loss during training.
Full-size DOI: 10.7717/peerj-cs.3252/fig-3

Table 1 Main hyperparameters used in training.

Hyperparameter	Value
Batch size	32
Number of epochs	15
Learning rate	3×10^{-4}
Scheduler	CosineAnnealingLR ($T_{\max} = 40$, $\eta_{\min} = 10^{-4}$)
Optimizer	AdamW (weight decay = 0.01)
Class weights	6.0 (for non-empty annotations only)
Data augmentation	Horizontal flip, Rotation (within ±15°)

All ablation study configurations were trained using identical hyperparameters and data splits to ensure fair comparison, with model performance evaluated using the same IoU and specificity metrics as the main experiments.

Our computing system consisted of an Intel Core i5-10400 CPU and an NVIDIA RTX A6000 GPU with 48GB VRAM. Python 3.10 was utilized. The deep learning framework employed was PyTorch 1.13.0+cu117 with torchvision 0.14.0+cu117, running on CUDA 11.7.

Model assessment

The MIME-ViT model processes an input image to produce a 32 × 32 tensor output, with each element representing a 64 × 64 pixel mask patch within the original 2,048 × 2,048 pixel image. These patches are binary, indicating the presence (1) or absence (0) of a feature, determined by the mask's coverage area within the patch. In this study, a detection threshold of 0.2 was applied, whereby a patch was marked as '1' if the mask covered over 20% of its area. This enhances the model's sensitivity to subtle anomalies.

The Detection Transformer (DETR) model (*Zhu et al., 2020*), a Transformer-based object detection model, was adopted for comparison. DETR was implemented in two versions: with and without pre-training. Additionally, YOLOv8 (*Jocher, Chaurasia & Qiu, 2023*), a state-of-the-art convolutional neural network-based object detection model, was included for comparative evaluation. YOLOv8 was also implemented in two versions: with and without pre-training. In contrast, the MIME-ViT model was trained without pre-training.

(1) DETR-Scratch (DETR-S): without pre-training for any component. (2) DETR-Pretrained (DETR-P): with pre-training for both the backbone and detection head. (3) YOLOv8-Scratch (YOLOv8-S): YOLOv8x model without pre-training for any component. (4) YOLOv8-Pretrained (YOLOv8-P): YOLOv8x model with pre-trained weights on COCO dataset, representing the largest parameter variant in the YOLOv8 family.

The mean Intersection over Union (mIoU) for images with lesions and specificity for lesion-free images were used for comparative analysis. In mammography, a significant number of images do not contain any lesions. While mIoU is a suitable metric for evaluating detection performance on images with lesions, it cannot be directly applied to lesion-free images as there are no ground truth objects to calculate Intersection over Union against. Therefore, to assess the model's ability to correctly identify lesion-free images, specificity was employed. Specificity is defined as the proportion of actual negatives that are correctly identified as such. The formula for specificity is:

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}} \tag{3}$$

where TN represents the number of lesion-free images correctly classified as negative (*i.e.*, no lesions detected), and FP represents the number of lesion-free images incorrectly classified as positive (*i.e.*, lesions detected where none exist). As indicated in Table 1, the DETR-S model did not identify any lesions, calcifications, or masses. This means that for all images, including lesion-free ones, DETR-S produced no positive detections. Consequently, when evaluating lesion-free images, the number of FP was 0. Applying this to the specificity formula (TN/(TN + 0)), the specificity for DETR-S on lesion-free images is 100%, as all actual negative cases (lesion-free images) were correctly classified as negative due to the absence of any positive findings by the model.

A confidence score threshold of 0.5 was set for IoU calculation in DETR. Because MIME-ViT is designed for detection tasks and outputs patch-like masks rather than direct segmentation, bounding boxes were applied around the exterior of patches to enable IoU

Table 2 IoU scores: measures the accuracy of lesion detection across models and lesion types.

Lesion type	Model	IoU (with lesions)	
		Mean	Std. Dev
All	DETR-Scratch	0.0000	0.0000
	DETR-Pretrained	0.3691	0.3021
	YOLOv8-Scratch	0.3400	0.3728
	YOLOv8-Pretrained	0.4516	0.3260
	MIME-ViT	0.3342	0.2477
Mass	DETR-Scratch	0.0000	0.0000
	DETR-Pretrained	0.4227	0.2864
	YOLOv8-Scratch	0.3896	0.3743
	YOLOv8-Pretrained	0.4804	0.3220
	MIME-ViT	0.3814	0.2531
Calc	DETR-Scratch	0.0000	0.0000
	DETR-Pretrained	0.0280	0.1241
	YOLOv8-Scratch	0.0454	0.1759
	YOLOv8-Pretrained	0.2796	0.2948
	MIME-ViT	0.2491	0.2139

Table 3 Specificity scores: each model's accuracy in identifying images without lesions.

Model	Specificity
DETR-Scratch	1.0000
DETR-Pretrained	0.6701
YOLOv8-Scratch	0.9124
YOLOv8-Pretrained	0.7887
MIME-ViT	0.2216

calculation for detection. For ground truth, patches were generated from labels, and bounding boxes were then applied around their exteriors.

RESULTS

Tables 2 and 3 present IoU scores for lesion detection accuracy and specificity scores for identifying images without lesions, respectively, while Fig. 4 shows IoU score comparisons across different lesion categories. We employed YOLOv8x, the largest parameter model in the YOLOv8 family, which has been reported to achieve the highest accuracy among YOLOv8 variants. YOLOv8-Pretrained achieved the highest overall IoU score of 0.4516 across all lesions, with superior performance on mass lesions (0.4804) and moderate performance on calcification lesions (0.2796). DETR-Scratch failed to detect any lesions, resulting in zero IoU scores across all categories.

The proposed MIME-ViT model achieved an overall IoU of 0.3342, with scores of 0.3797 for mass lesions and 0.2491 for calcification lesions. While MIME-ViT showed

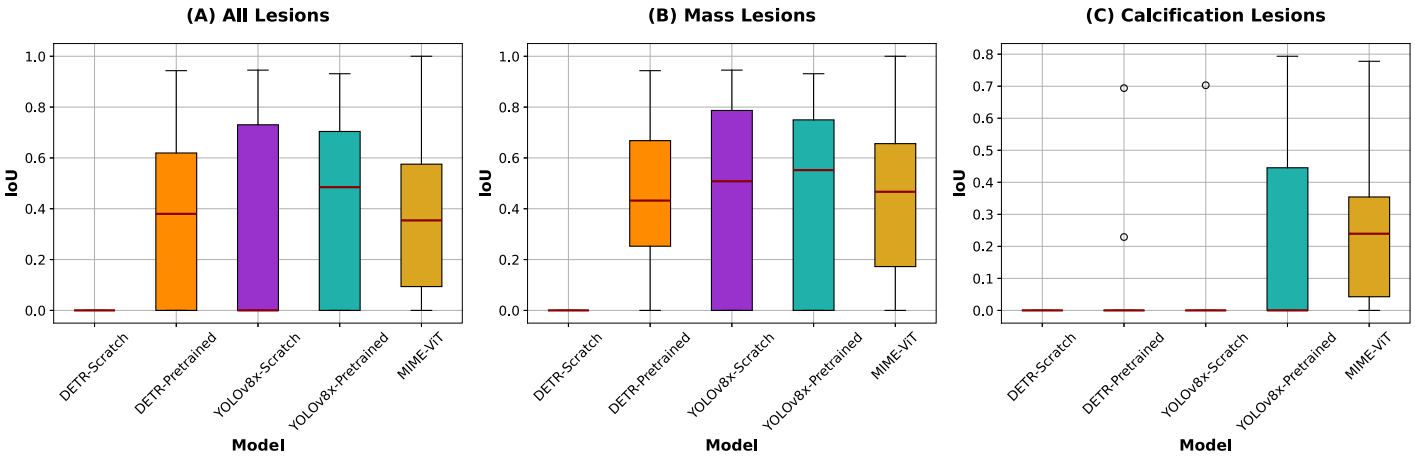

Figure 4 IoU scores comparison across different lesion categories. Box plots illustrate the distribution of IoU scores for each model across all lesions (left), mass lesions (center), and calcification lesions (right), with red lines indicating the medians. For categories where IoU values are concentrated at 0.0000, dot plots are shown instead due to the lack of variability. Full-size DOI: 10.7717/peerj-cs.3252/fig-4

lower mean IoU compared to YOLOv8-Pretrained, it demonstrated notably more stable performance with the lowest standard deviation across all lesion categories (0.2477 *vs* 0.3260 for overall, 0.2531 *vs* 0.3220 for mass lesions). For calcification detection, although MIME-ViT's mean IoU (0.2491) was lower than YOLOv8-Pretrained (0.2796), MIME-ViT achieved superior median performance and substantially lower standard deviation (0.2139 *vs* 0.2948), indicating more consistent calcification detection capability.

For specificity in identifying images without lesions, excluding DETR-Scratch which achieved perfect specificity (1.0000) due to zero detections, YOLOv8-Pretrained demonstrated specificity of 0.7887. DETR-Pretrained showed a specificity of 0.6701. MIME-ViT exhibited the lowest specificity at 0.2216 among all detection-capable models, indicating higher false positive rates in lesion-free images.

Ablation studies and component analysis

To validate the architectural design choices and understand the contribution of each component, we conducted comprehensive ablation studies examining ViT scale ranges, CNN kernel sizes, and feature fusion strategies.

We first systematically evaluated the contribution of each ViT component through systematic removal experiments. Table 4 demonstrates that each scale contributes unique information to the final prediction:

The removal of any ViT component resulted in performance degradation across all metrics, with ViT-256 removal showing the largest impact (IoU decrease of 0.0253). This validates our hypothesis that each scale captures complementary information: ViT-512 for global context, ViT-256 for intermediate-scale features, and ViT-64 for high-resolution details.

Table 4 Systematic ViT component removal analysis. Performance degradation when removing individual ViT scales validates the complementary nature of multi-scale feature extraction.

Configuration	Mean IoU	Mass IoU	Calc IoU
Full MIME-ViT	0.3342	0.3814	0.2491
w/o ViT-512	0.3156	0.3621	0.2287
w/o ViT-256	0.3089	0.3547	0.2195
w/o ViT-64	0.3198	0.3672	0.2341

Table 5 CNN kernel size ablation study. The 3 × 3 configuration outperformed 5 × 5 due to pretrained weight availability and optimized receptive field characteristics.

Kernel size	Mean IoU	Mass IoU	Calc IoU	Pretrained weights	Training status
3 × 3 (baseline)	0.3342	0.3814	0.2491	Available	Stable convergence
5 × 5	0.2947	0.3362	0.2198	Not compatible	Random initialization

Table 6 Feature fusion strategy comparison. Concatenation-based fusion outperformed additive fusion by preserving individual scale-specific feature representations.

Fusion strategy	Mean IoU	Mass IoU	Calc IoU	Architectural benefit
Concatenation (final)	0.3342	0.3814	0.2491	Preserves scale-specific features
Additive fusion	0.3087	0.3542	0.2218	Simpler parameter count

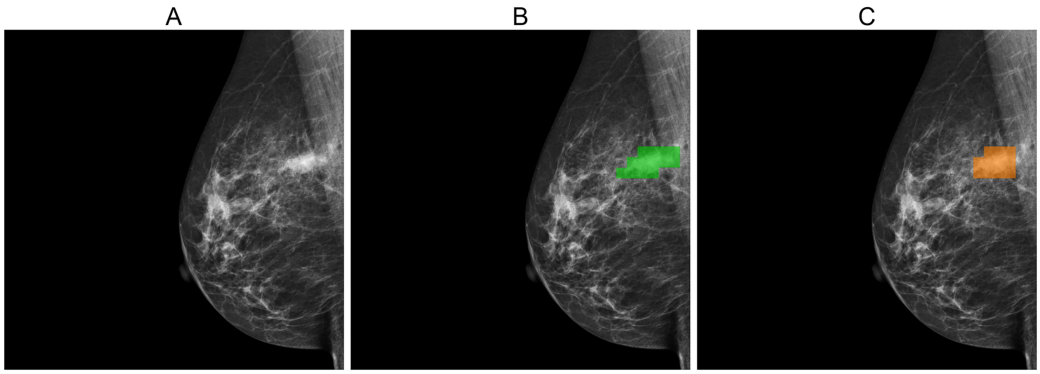

Figure 5 (A) Original image, (B) ground truth segmentation mask, (C) predicted segmentation mask using MIME-ViT. This example illustrates a true-positive case where the model accurately predicted breast lesions, achieving an IoU score of 0.5882. Full-size DOI: 10.7717/peerj-cs.3252/fig-5

We also investigated the impact of CNN kernel sizes within the ResidualBlock component by comparing 3 × 3 kernels (baseline) *vs* 5 × 5 kernels. Table 5 summarizes the results:

The 3 × 3 kernel configuration achieved superior performance (IoU = 0.3342 *vs* 0.2947) primarily due to its compatibility with pretrained weights, which enabled stable convergence and better feature initialization. The 5 × 5 configuration, requiring random initialization due to dimensional incompatibility with pretrained weights, showed reduced performance despite the larger receptive field.

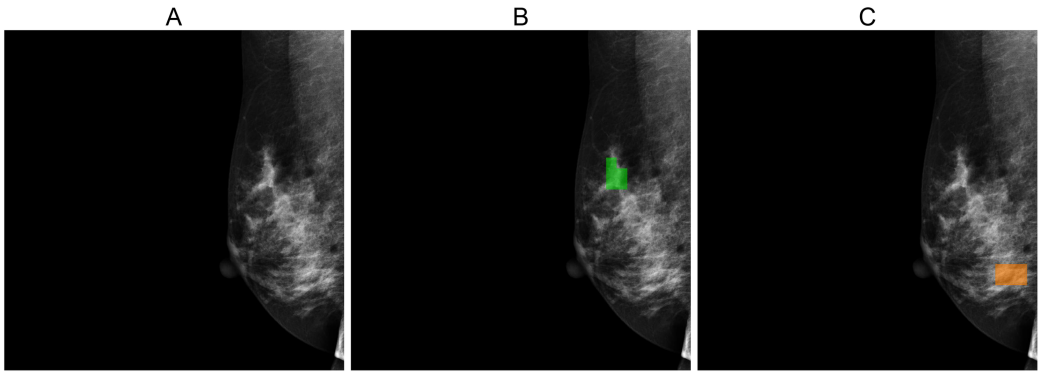

Figure 6 (A) Original image, (B) ground truth segmentation mask, (C) predicted segmentation mask using MIME-ViT. The images depict a case where the MIME-ViT model incorrectly predicts the presence of breast lesions in areas different from the ground truth, achieving an IoU score of 0.0.
Full-size DOI: 10.7717/peerj-cs.3252/fig-6

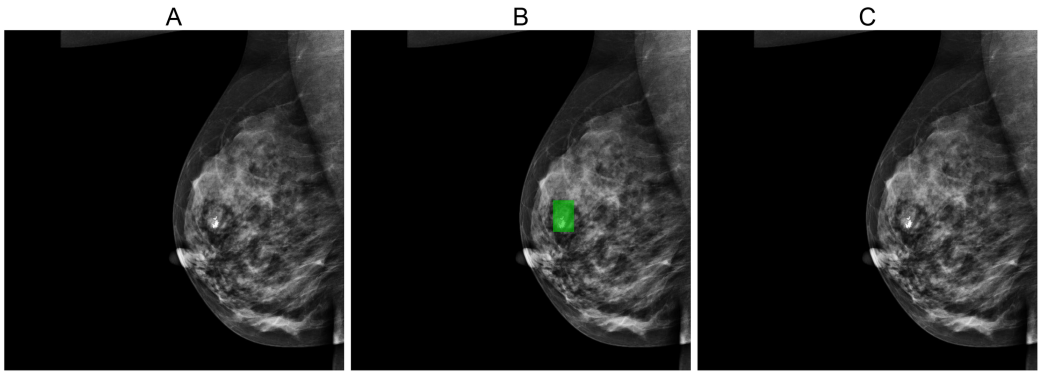

Figure 7 (A) Original image, (B) ground truth segmentation mask, (C) Predicted segmentation mask using MIME-ViT. The images illustrated a false negative case where the MIME-ViT model did not detect any breast lesion despite the presence of a breast lesion in the ground truth, achieving an IoU score of 0.0.
Full-size DOI: 10.7717/peerj-cs.3252/fig-7

Finally, we systematically compared different feature fusion approaches to optimize multi-scale information integration. Table 6 presents the comparison between additive and concatenation-based fusion strategies:

The concatenation-based approach achieved superior performance (IoU improvement of 0.0255) compared to the initial additive fusion strategy
($x = x_{512} + x_{256} + x_{64} + x_{conv} \times 20$). This improvement demonstrates that preserving individual scale-specific features enables the model to learn optimal integration strategies rather than enforcing uniform feature combination.

Figures 5, 6, and 7 depict the prediction results obtained using MIME-ViT. Figure 5 illustrates a representative true positive detection by MIME-ViT, accurately identifying breast lesions with an IoU score of 0.5882. Figure 6 presents a complex scenario with the model's predictions including both false positives (incorrect lesion predictions in lesion-free areas) and false negatives (missed detections of actual lesions). Figure 7 depicts

a false negative case where the model failed to detect breast lesions despite their presence in the ground truth, resulting in an IoU score of 0.0.

DISCUSSION

MIME-ViT demonstrates notable detection accuracy without the need for pre-training, even with a larger parameter count (112.9 million) compared to DETR (41.3 million). This is significant because traditional Vision Transformers typically depend on extensive datasets or pre-training for optimal performance. For example, DETR-S, which lacked pre-training, registered an IoU of 0.000, highlighting the usual necessity for substantial training data. In contrast, MIME-ViT achieved comparable accuracy to the pre-trained DETR-P, indicating that its architecture is well-optimized to perform effectively without the conventional reliance on large datasets or pre-training. This showcases MIME-ViT's capability in efficiently handling data-intensive tasks.

Traditional Vision Transformers (ViTs) generally require extensive data to achieve optimal performance, often relying on either large datasets or pre-training strategies. In this study, DETR-S, which did not undergo pre-training, demonstrated an IoU of 0.000, reinforcing this dependency. However, MIME-ViT, despite having a larger parameter count than DETR (112.9 million *vs* 41.3 million), matched the performance of DETR-P without the need for pre-training. This result emphasizes the efficiency and robustness of MIME-ViT's architecture. This result is considered to be attributed to two primary reasons:

1. The integration of multiscale analysis, an inductive bias (*Battaglia et al., 2018*) similar to a method utilized in physicians' evaluations, enhances the model's efficiency. This alignment with expert assessment practices allows for more efficient parameter optimization during the model's training. Consequently, the model finds optimal solutions more readily, leading to high accuracy in specific tasks such as the detection of calcification, as demonstrated in this research.
2. General segmentation models assign a class to each pixel within an area, whereas the MIME-ViT model allocates broader patches, specifically 64 × 64 pixel blocks. This approach categorizes MIME-ViT as a segmentation-like model. Such a segmentation-like design is believed to contribute to the model's heightened learning efficiency (*Ciresan et al., 2012*). Given the demonstrated efficacy of MIME-ViT in achieving comparable accuracy, its architectural design, especially the multiscale structure, exhibits potential applicability beyond detection to include segmentation and classification tasks. This research introduced a multiscale structure with four layers: 512 × 512, 256 × 256, 64 × 64, and 16 × 16 pixels. However, depending on the main task, the optimal configuration of layers might vary.

There are several limitations in this research. First, the model's evaluation on a dataset may not cover sufficient diversity, impacting its wider applicability and presenting a risk of overfitting. Improvement could be achieved by expanding the dataset to include a wider variety of images from diverse demographics and conditions. Additionally, implementing techniques such as data augmentation and cross-validation could help mitigate overfitting

and improve the model's generalizability. Second, MIME-ViT has not undergone pre-training. While further accuracy improvements are expected with the application of pre-training, the significantly different model structure from conventional DL models poses a challenge. Adapting MIME-ViT to utilize traditional pre-trained models requires a reconstruction of the model's structure. Future plans include exploring pre-training MIME-ViT with datasets such as ImageNet. Third, the model's decision-making process is not transparent, a critical factor for its adoption in medical settings where interpretability is key. Although Vision Transformers (ViTs) have the capability to visualize attention maps, offering a potential pathway to greater transparency, this feature was not utilized in our study. Recent advances emphasize that Explainable AI (XAI) integration is essential for building clinical trust (*Singh & Patnaik, 2025*), and future development will incorporate Grad-CAM and similar interpretability techniques into MIME-ViT to enhance transparency. Fourth, our approach to generating bounding boxes by circumscribing the exterior of segmented patches for IoU calculation may warrant further consideration. For lesions with irregular shapes, this method could potentially include extraneous background regions, which in turn might lead to an underestimation of the model's true detection performance. Exploring alternative strategies for bounding box generation in future work could therefore be beneficial for achieving a more precise evaluation. Lastly, the model's performance in controlled conditions might not directly translate to real-world clinical environments, where variability is greater.

CONCLUSIONS

The MIME-ViT model is a significant advancement in breast cancer detection, combining Vision Transformers and CNNs for improved mammographic imaging analysis. Its development represents a crucial step forward in medical imaging, aiming to enhance early detection and patient outcomes with its innovative approach. As the field of medical imaging evolves, MIME-ViT represents a pivotal step towards more effective and technologically advanced cancer screening methodologies.

ACKNOWLEDGEMENTS

The authors acknowledge the use of Claude (Anthropic) for English language editing, grammar checking, and proofreading to improve the clarity and readability of this manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the JST (CREST Grant No. JPMJCR15D1), JSPS KAKENHI Grant Number JP20H03738, and the MEXT/JSPS WISE Program: Advanced Graduate Program for Future Medicine and Health Care, Tohoku University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
JST: JPMJCR15D1.
JSPS KAKENHI: JP20H03738.
MEXT/JSPS WISE Program: Advanced Graduate Program for Future Medicine and Health Care, Tohoku University.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Yuki Kashiwada conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Eichi Takaya conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Mei Hiroya analyzed the data, prepared figures and/or tables, and approved the final draft.
- Nanako Matsuda analyzed the data, prepared figures and/or tables, and approved the final draft.
- Takumi Yashima analyzed the data, prepared figures and/or tables, and approved the final draft.
- Tomoya Kobayashi analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Gen Tamiya conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Takuya Ueda conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available in the Supplemental File, GitHub, and Zenodo:

- https://github.com/javasparrows/MIME-ViT.
- Kashiwada, Y. (2025). MIME-ViT (1.2). Zenodo. https://doi.org/10.5281/zenodo.16417003.

The TOMPEI-CMMD Dataset is available at Kashiwada, Y., Takaya, E., Hiroya, M., Matsuda, N., Yashima, T., Kobayashi, T., Tamiya, G., & Ueda, T. (2025). TOMPEI-CMMD Dataset (Version 1) [Dataset]. The Cancer Imaging Archive. https://doi.org/10.7937/WEZW-BH22.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3252#supplemental-information.

REFERENCES

Anastasiadi Z, Lianos G, Ignatiadou E, Harissis H, Mitsis M. 2017. Breast cancer in young women: an overview. *Updates in Surgery* **69(3)**:313–317 DOI 10.1007/s13304-017-0424-1.

Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, Jose A, Roy R, Merhof D. 2024. Advances in medical image analysis with vision transformers: a comprehensive review. *Medical Image Analysis* **91(1)**:103000 DOI 10.1016/j.media.2023.103000.

Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, Tacchetti A, Raposo D, Santoro A, Faulkner R, Gulcehre C, Song F, Ballard A, Gilmer J, Dahl G, Vaswani A, Allen K, Nash C, Langston V, Dyer C, Heess N, Wierstra D, Kohli P, Botvinick M, Vinyals O, Li Y, Pascanu R. 2018. Relational inductive biases, deep learning, and graph networks. ArXiv DOI 10.48550/arXiv.1806.01261.

Cai H, Wang J, Dan T, Li J, Fan Z, Yi W, Cui C, Jiang X, Li L. 2023. An online mammography database with biopsy confirmed types. *Scientific Data* **10(1)**:123 DOI 10.1038/s41597-023-02025-1.

Chen RJ, Chen C, Li Y, Chen TY, Trister AD, Krishnan RG, Mahmood F. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 16144–16155.

Chen J-L, Cheng L-H, Wang J, Hsu T-W, Chen C-Y, Tseng L-M, Guo S-M. 2023. A YOLO-based AI system for classifying calcifications on spot magnification mammograms. *Biomedical Engineering Online* **22(1)**:54 DOI 10.1186/s12938-023-01115-w.

Chen C-F, Fan Q, Panda R. 2021. CrossViT: cross-attention multi-scale vision transformer for image classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 357–366.

Ciresan D, Giusti A, Gambardella L, Schmidhuber J. 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In: Pereira F, Burges C, Bottou L, Weinberger K, eds. *Advances in Neural Information Processing Systems*. Vol. 25. New York: Curran Associates, Inc.

Contreras NSL, Grisi C, Aswolinskiy W, Vatrano S, Fraggetta F, Nagtegaal I, D'Amato M, Ciompi F. 2024. Benchmarking hierarchical image pyramid transformer for the classification of colon biopsies and polyps histopathology images. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. Piscataway: IEEE, 1–4.

Cui C, Li L, Cai H, Fan Z, Zhang L, Dan T, Li J, Wang J. 2021. The Chinese mammography database (CMMD): an online mammography database with biopsy confirmed types for machine diagnosis of breast. DOI 10.7937/tcia.eqde-4b16.

Dice LR. 1945. Measures of the amount of ecologic association between species. *Ecology* **26(3)**:297–302 DOI 10.2307/1932409.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. 2020. An image is worth 16x16 words: transformers for image recognition at scale. ArXiv DOI 10.48550/arXiv.2010.11929.

Fowler E, Sellers T, Lu B, Heine J. 2013. Breast imaging reporting and data system (BI-RADS) breast composition descriptors: automated measurement development for full field digital mammography. *Medical Physics* **40(11)**:113502 DOI 10.1118/1.4824319.

Guo Z, Zhao W, Wang S, Yu L. 2023. HIGT: hierarchical interaction graph-transformer for whole slide image analysis. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer, 755–764.

Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z, Zhang Y, Tao D. 2022. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45(1)**:87–110 DOI 10.1109/tpami.2022.3152247.

Harbeck N, Gnant M. 2017. Breast cancer. *Lancet* **389(10074)**:1134–1150 DOI 10.1016/s0140-6736(16)31891-8.

Jocher G, Chaurasia A, Qiu J. 2023. Ultralytics YOLOv8. *Available at https://docs.ultralytics.com/ja/models/yolov8/*.

Kashiwada Y, Takaya E, Hiroya M, Matsuda N, Yashima T, Kobayashi T, Tamiya G, Ueda T. 2024. TOMPEI-CMMD. DOI 10.7937/wezw-bh22.

Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, Breast Cancer Surveillance Consortium. 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine* **175(11)**:1828–1837 DOI 10.1001/jamainternmed.2015.5231.

Lin T-Y, Goyal P, Girshick R, He K, Dollár P. 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*. Piscataway: IEEE, 2980–2988.

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. 2021. Swin Transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 10012–10022.

Loshchilov I, Hutter F. 2016. SGDR: stochastic gradient descent with restarts. ArXiv DOI 10.48550/arXiv.1608.03983.

Loshchilov I, Hutter F. 2017. Decoupled weight decay regularization. ArXiv DOI 10.48550/arXiv.1711.05101.

Mendelson E. 2019. Artificial intelligence in breast imaging: potentials and limitations. *American Journal of Roentgenology* **212(2)**:293–299 DOI 10.2214/ajr.18.20532.

Myers E, Moorman P, Gierisch J, Havrilesky L, Grimm L, Ghate S, Davidson B, Montomery R, Crowley M, McCrory D, Kendrick A, Sanders G. 2015. Benefits and harms of breast cancer screening: a systematic review. *JAMA* **314(15)**:1615–1634 DOI 10.1001/jama.2015.13183.

Nassif A, Talib M, Nasir Q, Afadar Y, Elgendy O. 2022. Breast cancer detection using artificial intelligence techniques: a systematic literature review. *Artificial Intelligence in Medicine* **127**:102276 DOI 10.1016/j.artmed.2022.102276.

Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L, U.S. Preventive Services Task Force. 2009. Screening for breast cancer: an update for the U.S. preventive services task force. *Annals of Internal Medicine* **151(10)**:727–737, W237–W242 DOI 10.7326/0003-4819-151-10-200911170-00009.

Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. 2021. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems* **34**:12116–12128.

Raya-Povedano J, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. 2021. AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation. *Radiology* **300(1)**:57–65 DOI 10.1148/radiol.2021203555.

Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H. 2023. Transformers in medical imaging: a survey. *Medical Image Analysis* **88(1)**:102802 DOI 10.1016/j.media.2023.102802.

Shoshan Y, Bakalo R, Gilboa-Solomon F, Ratner V, Barkan E, Ozery-Flato M, Amit M, Khapun D, Ambinder EB, Oluyemi ET, Panigrahi B, DiCarlo PA, Rosen-Zvi M, Mullen LA. 2022.

Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis. *Radiology* **303(1)**:69–77 DOI 10.1148/radiol.211105.

Shoshan Y, Bakalo R, Gilboa-Solomon F, Ratner V, Barkan E, Ozery-Flato M, Amit M, Khapun D, Ambinder EB, Oluyemi ET, Panigrahi B, DiCarlo PA, Rosen-Zvi M, Mullen LA. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630(8015)**:181–188 DOI 10.1038/s41586-024-07441-w.

Singh SK, Patnaik KS. 2024. Convergence of various computer-aided systems for breast tumor diagnosis: a comparative insight. *Multimedia Tools and Applications* **84(16)**:16709–16756 DOI 10.1007/s11042-024-19620-y.

Singh SK, Patnaik KS. 2025. Patho-AI: a perceptive breast cancer identification and classification using deep learning methods integrated with explainable AI. *SN Computer Science* **6**:619 DOI 10.1007/s42979-025-04170-3.

Sorensen T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter* **5**:1–34.

van Leeuwen K, de Rooij M, Schalekamp S, van Ginneken B, Rutten M. 2022. How does artificial intelligence in radiology improve efficiency and health outcomes? *Pediatric Radiology* **52(11)**:2087–2093 DOI 10.1007/s00247-021-05114-8.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* **30**:5998–6008.

Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, Lu T, Luo P, Shao L. 2022. PVT v2: improved baselines with pyramid vision transformer. *Computational Visual Media* **8(3)**:415–424 DOI 10.1007/s41095-022-0274-8.

Yu T, Kumar S, Gupta A, Levine S, Hausman K, Finn C. 2020. Gradient surgery for multi-task learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, eds. *Advances in Neural Information Processing Systems*. Vol. 33. New York: Curran Associates, Inc, 5824–5836.

Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang Z-H, Tay FE, Feng J, Yan S. 2021. Tokens-to-Token ViT: training vision transformers from scratch on ImageNet. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 558–567.

Zhu X, Su W, Lu L, Li B, Wang X, Dai J. 2020. Deformable DETR: deformable transformers for end-to-end object detection. ArXiv DOI 10.48550/arXiv.2010.04159.