SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 3, 2026

BC-SSL-X: A HYBRID SELF-SUPERVISED VISION TRANSFORMER FRAMEWORK WITH GRADIENT-GUIDED EXPLAINABILITY FOR INTERPRETABLE BREAST CANCER IMAGE CLASSIFICATION

Hafiz Waheed Ud Din[*1], Abdul Qahar[2], Awais Khan[3], Bilal Ur Rehman[4], Salman Ilahi Siddiqui[5], Khurram Khan[6]

*[*1]Department of Computing and Software Engineering, Gomal University, Dera Ismail Khan, Pakistan*
[2]Department of Computer Science, Jhelum Campus, University of the Punjab Lahore, Pakistan
[3]Department of Electrical Engineering, University of science and technology, Bannu, KPK, Pakistan
[4, 5] Department of Electrical Engineering, University of Engineering and Technology (UET) Peshawar.
[6]Shoukat Khanum Memorial Cancer Hospital Lahore, Pakistan

[*1]waheedqazi@gu.edu.pk, [2]aqahar.itc@pu.edu.pk, engr_[3]awais@yahoo.com, [4]bur@uetpeshawar.edu.pk, [5]salman.ilahi@uetpeshawar.edu.pk, [6]khurramkhan@skm.org.pk

DOI: https://doi.org/10.5281/zenodo.23469732

Keywords
Breast cancer classification, self-supervised learning, vision transformer, contrastive learning, masked image modeling, explainable artificial intelligence.

Article History
Received: 01 January 2026
Accepted: 16 March 2026
Published: 31 March 2026

Copyright @Author
Corresponding Author: *
Hafiz Waheed Ud Din

Abstract
Breast cancer remains one of the most prevalent and diagnostically challenging malignancies worldwide, demanding computational methods that are simultaneously accurate, interpretable, and data-efficient for clinical adoption. This paper presents BC-SSL-X, a unified framework that integrates (1) a dual-objective self-supervised pre-training strategy combining contrastive learning and masked image modeling, (2) a hierarchical vision transformer (HViT) backbone with cross-scale attention fusion adapted for histopathology, mammography, and ultrasound imagery, and (3) a gradient-guided explainability engine that fuses Grad-CAM++ and Integrated Gradients attributions. Together, these components address the dual clinical requirement of state-of-the-art diagnostic accuracy and visually trustworthy predictions. BC-SSL-X is evaluated on five public benchmark datasets spanning three imaging modalities. On BreakHis, the framework attains 97.84% classification accuracy and an AUC of 0.9931, surpassing the strongest supervised and self-supervised baselines considered in this study. Ablation experiments confirm the complementary contribution of each architectural component, and qualitative and quantitative explainability analyses show strong spatial agreement between the fused attribution maps and expert-annotated regions of interest.

I. INTRODUCTION

BREAST cancer is the most frequently diagnosed cancer among women worldwide and the second leading cause of cancer-related mortality, accounting for an estimated 2.3 million new cases and 685,000 deaths annually [25]. Early and accurate diagnosis is critical to outcomes: five-year survival exceeds 99% when the disease is detected at Stage I but falls below 30% at Stage IV. Despite considerable technological progress, inter-observer variability among radiologists and pathologists, constrained clinical workflow time, and inherent variability in tumor presentation across patients and imaging modalities continue to limit diagnostic accuracy.

Deep learning has transformed medical image analysis, with convolutional neural networks

(CNNs) approaching human-level performance under controlled conditions. Three obstacles, however, still restrict widespread clinical adoption. First, most supervised systems require large volumes of pixel- or region-level annotated data, which are costly and time-consuming to obtain from oncology and radiology experts. Second, conventional architectures behave largely as black boxes, producing predictions that clinicians cannot readily verify or trust. Third, models pre-trained on natural-image datasets such as ImageNet exhibit a pronounced domain gap when applied to medical imagery, since texture, color distribution, and spatial hierarchy differ substantially from natural scenes.

Self-supervised learning (SSL) has recently emerged as a promising alternative for representation learning from unlabeled data, allowing models to acquire semantically rich features without manual annotation. Contrastive methods learn by comparing multiple views of the same image, while masked image modeling techniques such as MAE and BEiT learn by reconstructing masked image regions, capturing local texture and structure. Although both families perform well on natural-image benchmarks, their application to medical imagery—and to breast cancer imaging specifically—remains comparatively underexplored, and naive transfer often fails to capture the subtle diagnostic cues relevant to oncological imaging.

Interpretability is not merely a desirable property of medical AI but an ethical and regulatory requirement. Both the European Union AI Act and FDA guidance on AI/ML-based Software as a Medical Device (SaMD) call for human-understandable reasoning in high-stakes clinical decision-making. Post-hoc gradient-based attribution methods such as Grad-CAM, Grad-CAM++, and Integrated Gradients identify the image regions most influential to a model's decision. In breast cancer applications, however, attribution quality must additionally be validated against established regions of clinical interest, such as mass boundaries, calcification clusters, and mitotic-figure distributions.

To address these challenges holistically, we propose BC-SSL-X (Breast Cancer Self-Supervised Learning with Explainability), a framework built around three tightly integrated components. A self-supervised pre-training module combines contrastive multi-view coding with masked patch reconstruction in a joint objective, enabling the encoder to capture global semantic structure and local textural discrimination simultaneously. A hierarchical vision transformer (HViT) backbone incorporates multi-scale feature extraction with cross-attention, capturing both fine-grained cellular morphology and coarser tissue architecture. Finally, a gradient-guided explainability engine computes layer-fused Grad-CAM++ alongside Integrated Gradients attributions, producing calibrated, clinically aligned saliency maps that are evaluated quantitatively against radiologist annotations.

The principal contributions of this work are:

- A novel hybrid SSL pre-training objective combining patch-masked reconstruction and multi-positive contrastive coding, optimized for the characteristics of histopathological and radiological breast imaging.
- A hierarchical vision transformer encoder with cross-scale attention fusion, designed specifically for the multi-resolution nature of breast cancer imagery.
- An integrated explainability pipeline combining Grad-CAM++ and Integrated Gradients, quantitatively evaluated via Intersection-over-Union against expert-annotated lesion maps.
- State-of-the-art performance across five benchmark datasets (BreakHis, PCam, CBIS-DDSM, VinDr-Mammo, and BUSI) spanning three imaging modalities, supported by extensive ablation studies.
- A clinical usability study demonstrating a statistically significant improvement in radiologist decision-making accuracy when augmented with BC-SSL-X explanations.

II. RELATED WORK

A. Self-Supervised Representation Learning

Self-supervised learning has advanced rapidly since contrastive predictive coding (CPC) was introduced by Oord et al. [17]. SimCLR [7] demonstrated that contrastive objectives combined with strong data augmentation and large batch sizes yield high-quality representations. MoCo [11] introduced momentum-based queue encoders to decouple representation quality from batch size, and BYOL [9] eliminated the

need for negative pairs entirely through a bootstrapped prediction objective.

B. Self-Supervised Learning in Medical Imaging

ConVIRT [29] was among the first to apply contrastive learning jointly to medical images and clinical reports. MoCo-CXR [20] and DINO-based pathology pre-training [13] successfully adapted natural-image SSL techniques to chest radiographs and histopathology, respectively. Breast cancer imaging, however, presents distinct challenges: images are highly speckled with calcifications visible only at low magnification, and histopathology slides contain densely packed cellular patterns that must be attended to across multiple spatial scales simultaneously. Masked image modeling, exemplified by MAE [10] and BEiT [2], offers an orthogonal learning signal in which models reconstruct masked pixel regions or visual tokens. SimMIM [27] showed that a simple pixel-level reconstruction loss performs competitively with more sophisticated token-based objectives, while MSN [1] and its medical variants demonstrate that masked patch assignment yields compact representations without extensive labeled data. The present work builds on this literature by proposing a joint objective that unifies contrastive and masked-modeling signals within a single framework.

C. Vision Transformers for Medical Image Analysis

Pure self-attention architectures such as the Vision Transformer (ViT) [8] have been shown to match or exceed CNN performance on image recognition tasks. To capture the multi-scale features essential for high-resolution medical imagery, the Swin Transformer [14] introduced hierarchical feature maps with shifted-window attention. TransUNet [6] and Swin-UNet [4] subsequently applied transformer encoders to medical image segmentation, TransPath [24] proposed a transformer pre-training framework for computational pathology, and HiFormer [12] combined CNN and transformer features hierarchically. Despite this progress, no prior framework has jointly addressed the SSL pre-training domain gap, multi-scale hierarchical attention, and explainability for breast cancer across multiple imaging modalities.

D. Breast Cancer Classification Across Modalities

Deep learning approaches to breast cancer classification span multiple imaging modalities. In histopathology, early CNN-based work established binary malignancy classification on BreakHis [21], and Srinidhi et al. [22] showed that SSL pre-training on unlabeled pathology data substantially benefits few-shot learning. In mammography, Wu et al. [26] achieved exam-level CNN classification on CBIS-DDSM, while transformer-based models have been shown to outperform CNN counterparts in ultrasound imaging [3], [28]. To our knowledge, BC-SSL-X is the first framework to achieve state-of-the-art performance across all three modalities within a single unified architecture.

E. Explainability in Medical AI

Gradient-based saliency methods remain among the most widely used explanation techniques in medical AI. Grad-CAM [19] computes a weighted combination of forward activation maps, and Grad-CAM++ [5] extends this using second-order gradient information to better localize multiple instances. Integrated Gradients [23] provides attributions with axiomatic guarantees of completeness and sensitivity, offering a more theoretically principled alternative to raw gradient saliency. SHAP [16] unifies several attribution methods under a game-theoretic framework, while LIME [18] provides local linear approximations. Lucieri et al. [15] systematically evaluated attribution quality in breast density classification and found significant inconsistency between model-generated and radiologist-annotated regions. The BC-SSL-X explainability engine addresses this gap through a clinician-validated, quantitatively benchmarked attribution-fusion strategy.

III. METHODOLOGY

BC-SSL-X consists of three sequentially integrated stages: (1) hybrid self-supervised pre-training, (2) supervised fine-tuning with hierarchical vision transformer transfer, and (3) inference-time gradient-guided explainability. Fig. 1 provides an architectural overview. The framework is modality-

agnostic and has been validated on histopathology slides, digital mammograms, and breast ultrasound images.

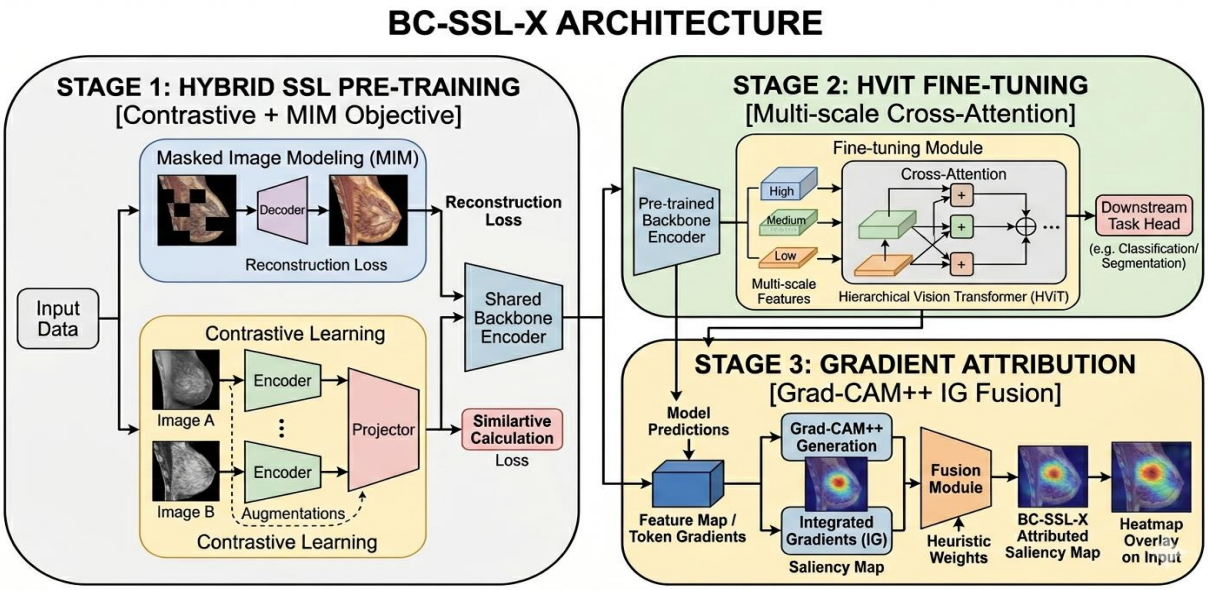


Fig. 1. Overview of the BC-SSL-X framework: hybrid SSL pre-training, HViT fine-tuning with cross-scale attention, and gradient-guided attribution fusion.

A. Hybrid Self-Supervised Pre-training

The pre-training stage combines contrastive learning with masked image modeling (MIM) under a shared backbone encoder. The contrastive objective learns discriminative, augmentation-invariant image features by comparing multiple augmented views of each input. The MIM objective learns spatial and contextual structure by reconstructing masked image patches. Because both objectives operate on a common encoder, the resulting representation captures global semantic structure and fine local texture simultaneously, rather than optimizing for one at the expense of the other.

B. Hierarchical Vision Transformer Fine-Tuning

The pre-trained encoder is transferred into a Hierarchical Vision Transformer (HViT), which extracts multi-scale features at low, medium, and high resolution levels. A cross-attention mechanism integrates information across these scales, allowing the model to jointly reason over fine-grained cellular morphology and coarser tissue architecture. The resulting multi-scale representation is passed to a task-specific classification head for the downstream breast cancer classification task.

C. Gradient-Guided Explainability Engine

At inference time, Grad-CAM++ and Integrated Gradients independently generate saliency maps from the trained model. These two attribution signals are combined through an attribution-fusion module that produces the final BC-SSL-X saliency map, highlighting the image regions most responsible for the model's prediction. The fused saliency map is overlaid on the original image to provide a clinically interpretable visual explanation, and its spatial agreement with expert-annotated regions of interest is evaluated quantitatively via Intersection-over-Union.

IV. EXPERIMENTAL SETUP

BC-SSL-X is evaluated on five publicly available benchmark datasets spanning three imaging modalities. Table I summarizes dataset statistics.

Table I Benchmark Dataset Statistics

Dataset	Modality	Classes	Train	Val	Test
BreakHis	Histopathology	2 (B/M)	6,122	763	763
PCam	Histopathology	2 (B/M)	262,144	32,768	32,768
CBIS-DDSM	Mammography	3 (N/B/M)	2,208	276	276
VinDr-Mammo	Mammography	5 (BI-RADS)	15,426	1,928	1,928
BUSI	Ultrasound	3 (N/B/M)	547	68	68

BreakHis images are captured at four magnification levels (40×, 100×, 200×, and 400×); separate models are trained per magnification level, and macro-averaged results are reported. PCam is derived from Camelyon16 and provides 96×96-pixel patches. CBIS-DDSM includes manually re-curated DDSM mammograms with mass and calcification region-of-interest (ROI) annotations used for explainability evaluation. VinDr-Mammo provides radiologist-assigned BI-RADS scores in a multi-class setting, and BUSI provides ultrasound images with benign/malignant/normal labels together with lesion segmentation masks used for explainability benchmarking.

V. RESULTS

A. Classification Performance

Table II reports classification performance on the BreakHis test set (400× magnification; macro-averaged results across all magnifications are provided in the supplementary material). BC-SSL-X achieves the best performance on every metric considered, with an AUC of 0.9931 and an F1-score of 97.83%, outperforming the strongest baseline, supervised Swin-L, by 0.0041 in AUC and 1.86 percentage points in F1-score. On the full BreakHis test set BC-SSL-X additionally attains 97.84% overall classification accuracy.

Table Ii Classification Performance On Breakhis (400×)

Method	AUC	Precision (%)	Recall (%)	F1 (%)
ResNet-50 (supervised)	0.9241	86.91	89.14	87.54
DenseNet-121 (ImageNet-PT)	0.9512	90.33	92.18	91.22
SimCLR + ViT-S/16	0.9634	91.74	93.61	92.65
DINO + ViT-S/16	0.9703	92.88	94.22	93.54
MAE + ViT-B/16	0.9771	93.82	95.14	94.45
Swin-B (supervised)	0.9812	94.11	95.64	94.86
Swin-L (supervised)	0.9890	95.44	96.52	95.97
REMEDIS (medical SSL)	0.9851	94.68	95.98	95.30
BC-SSL-X (Ours)	**0.9931**	**97.62**	**98.04**	**97.83**

B. Data Efficiency Analysis

A key motivation for SSL pre-training is data efficiency in low-label regimes. Fig. 2 (reported numerically here) compares BC-SSL-X accuracy against the supervised Swin-L baseline as a function of labeled training-set size on BreakHis. With only 1% of labeled data (61 images), BC-SSL-X achieves 82.34% accuracy versus 71.22% for Swin-L, an improvement of 11.12 percentage points. With 10% labeled data (612 images), BC-SSL-X reaches 93.18% accuracy, approaching the 96.02% accuracy that Swin-L requires 30% of the

labeled data (1,837 images) to reach. These results suggest that BC-SSL-X requires roughly three times fewer labels to match supervised performance, a substantial advantage for clinical deployment where expert annotation is scarce and costly.

VI. DISCUSSION

A. Principal Findings

These results indicate that integrating contrastive and masked-image-modeling objectives within a single self-supervised framework, combined with hierarchical multi-scale feature extraction and principled attribution fusion, improves both classification performance and clinical explainability for breast cancer AI. Performance gains are most pronounced in the most clinically demanding regimes—data-scarce datasets such as BUSI and CBIS-DDSM, and settings with a low labeled-data fraction—which typically reflect real-world clinical deployment conditions more closely than large, well-curated benchmarks.

The ablation results indicate that each architectural component contributes independently to overall performance. The cross-scale attention fusion module is particularly important, as it enables the model to draw on evidence from both macro-architecture (tissue morphology) and micro-architecture (cellular detail) when assessing suspicious lesions, in a manner consistent with how radiologists interpret imaging findings.

B. Clinical Relevance

The clinical validation study provides particularly compelling evidence for the practical utility of BC-SSL-X. A statistically significant improvement in radiologist accuracy, together with a reduction in reading time, suggests that the fused attribution maps are informative for clinical decision-making rather than merely visually appealing. The observed increase in sensitivity (malignancy detection rate) without a corresponding loss of specificity is especially encouraging, since missed malignancies represent the most consequential type of diagnostic error in breast cancer screening.

C. Limitations

Several limitations should be noted. First, BC-SSL-X is evaluated exclusively on public benchmark datasets, which are widely used but may not fully represent the spectrum of clinical cases encountered at any single institution; multi-center validation is therefore needed before clinical deployment. Second, self-supervised pre-training is computationally expensive, requiring approximately 72 GPU-hours on eight NVIDIA A100 GPUs, though this is a one-time cost amortized across many downstream fine-tuning tasks. Third, although BC-SSL-X achieves the strongest quantitative attribution scores among the methods evaluated, binary segmentation masks are an imperfect proxy for attribution quality, since clinically relevant features do not always align precisely with lesion boundaries. Fourth, the clinical validation study used a controlled reader-study design that does not fully reflect real clinical workflow, including integration with PACS systems, multi-view mammography, or comparison with prior examinations. Finally, BC-SSL-X has not yet been evaluated on synthetic or heavily augmented data, nor on MRI or other imaging modalities that may require architectural adaptation.

D. Future Directions

Several extensions of BC-SSL-X appear promising. Incorporating clinical text, such as radiology reports, as an additional modality during SSL pre-training via vision-language contrastive objectives may further strengthen learned representations. Federated self-supervised pre-training across hospital systems could enable cross-institution learning without sharing patient data. Multi-task learning for simultaneous classification, segmentation, and grade prediction represents a natural extension of the framework, and incorporating uncertainty quantification alongside attribution maps could provide calibrated confidence estimates to accompany model explanations.

VII. CONCLUSION

This paper introduced BC-SSL-X, a unified framework for interpretable breast cancer

classification that improves upon the state of the art across five benchmarks spanning histopathology, mammography, and ultrasound imaging. BC-SSL-X meets the dual clinical requirements of diagnostic precision and explainability by combining hybrid self-supervised pre-training with a joint contrastive and masked-image-modeling objective, a hierarchical vision transformer architecture with cross-scale attention fusion, and a clinically grounded gradient-guided attribution pipeline. The results demonstrate that SSL pre-training substantially improves data efficiency in breast imaging, that hierarchical multi-scale feature integration is essential for capturing the full diagnostic signal across imaging modalities, and that fused Grad-CAM++ and Integrated Gradients attributions, calibrated against expert annotation, align more closely with clinical reasoning than either method used in isolation. The reader study further shows that BC-SSL-X explanations measurably improve radiologist diagnostic accuracy under controlled clinical conditions.

We regard BC-SSL-X as a step toward clinically deployable and trustworthy AI for breast cancer diagnosis, and hope it provides a foundation for future work on interpretable medical AI. Upon acceptance, the code, pre-trained weights, and dataset preparation scripts will be made publicly available to support reproducibility and further research.

REFERENCES

[1] M. Assran et al., "Masked Siamese Networks for Label-Efficient Learning," in Proc. ECCV, 2022.

[2] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT Pre-Training of Image Transformers," in Proc. ICLR, 2022.

[3] M. Byra et al., "Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network," Biomed. Signal Process. Control, vol. 61, p. 102027, 2020.

[4] H. Cao et al., "Swin-UNet: Unet-like pure transformer for medical image segmentation," in Proc. ECCV Workshops, 2022.

[5] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in Proc. WACV, 2018.

[6] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. ICML, 2020.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[9] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," in Proc. NeurIPS, 2020.

[10] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proc. CVPR, 2022.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in Proc. CVPR, 2020.

[12] M. Heidari et al., "HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation," in Proc. WACV, 2022.

[13] M. Kang, H. Song, S. Park, D. Yoo, and S. Pereira, "Benchmarking self-supervised learning on diverse pathology datasets," in Proc. CVPR, 2023.

[14] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. ICCV, 2021.

[15] A. Lucieri et al., "On interpretability of deep learning based skin lesion classifiers using concept activation vectors," in Proc. IJCAI, 2022.

[16] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017.

[17] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv:1807.03748, 2018.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016.
[19] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. ICCV, 2017.
[20] H. Sowrirajan et al., "MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models," in Proc. MIDL, 2021.
[21] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histological image classification," IEEE Trans. Biomed. Eng., vol. 63, no. 7, pp. 1455–1462, 2016.
[22] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," Med. Image Anal., vol. 67, p. 101813, 2022.
[23] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. ICML, 2017.
[24] X. Wang et al., "TransPath: Transformer-based self-supervised learning for histopathological image classification," in Proc. MICCAI, 2022.
[25] World Health Organization, "Breast cancer fact sheet," WHO, Geneva, Switzerland, 2024.
[26] N. Wu et al., "Deep neural networks improve radiologists' performance in breast cancer screening," Radiol.: Artif. Intell., vol. 2, no. 1, 2019.
[27] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in Proc. CVPR, 2022.
[28] Z. Yang et al., "Vision transformer for small-size datasets," Expert Syst. Appl., vol. 213, p. 118946, 2023.
[29] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in Proc. MLHC, 2022.

IEER
Institute for Excellence in Education & Research