LREN-MedNet:A Latent Representation and Vision Transformer Approach for Automated Cancer Detection

Mrs. D. Monica Seles
Assistant Professor
Department of Artificial Intelligence
Mepco Schlenk Engineering College
Sivakasi, India
monicaselesd@mepcoeng.ac.in

Prahanya Selvakumar
Student
Department of Artificial Intelligence
Mepco Schlenk Engineering College
Sivakasi, India
prahanya17_bai25@mepcoeng.ac.in

Kawena M
Student
Department of Artificial Intelligence
Mepco Schlenk Engineering College
Sivakasi, India
kawena12335_bai25@mepcoeng.ac.in

***Abstract*—Breast cancer is one of the most prevalent and life-threatening cancers, posing significant challenges in early detection and diagnosis. The clinical application of ultrasound imaging for evaluating breast cancer risk is often hindered by inter-observer variability and the labor-intensive nature of manual image labeling. This paper introduces LREN-MedNet, a novel architecture based on Latent Representation and Vision Transformer designed to enhance computer-aided diagnosis of cancer using ultrasound imaging. The proposed method tackles the inherent variability of ultrasound images obtained from different medical devices by employing an Adaptive ROI Transformation (ART) technique to standardize the extraction of regions of interest (ROIs) and eliminate non-relevant pixels. Furthermore, we present a transformer-based Latent Representation Extraction Network (LREN) that utilizes a substantial volume of unlabeled ultrasound images to pre-train the feature extraction model, enabling effective transfer to supervised tasks such as classification and segmentation. We have utilized the publicly available BUSI dataset, which includes 780 annotated breast ultrasound images across three categories: benign, malignant, and normal, along with corresponding segmentation masks. Experimental findings indicate that LREN-MedNet significantly improves diagnostic performance compared to state-of-the-art deep learning models, underscoring its potential for clinical application in computer-aided breast cancer diagnosis.**

***Index Terms*—Endoscopic Ultrasound (EUS), Breast Cancer, Dual Self-Supervised Learning, Adaptive ROI Transformation (ART), Deep Learning, Medical Imaging**

I. Introduction

Breast cancer is one of the most prevalent and life-threatening cancers, characterized by a high mortality rate and aggressive progression in advanced stages. The five-year survival rate varies significantly depending on early detection, with late-stage diagnoses leading to poor prognoses. To improve patient outcomes, early and accurate diagnosis is crucial, yet it remains a significant challenge in clinical practice. Compared to other imaging techniques like mammography or magnetic resonance imaging (MRI), ultrasound imaging is widely used for breast cancer detection due to its non-invasive nature and effectiveness in evaluating dense breast tissue [1].Ultrasound provides high-resolution images of breast lesions [2]; however, variations between devices and inconsistencies in image quality hinder its regular use in automated diagnostic systems.

Deep learning has achieved remarkable results in various diagnostic tasks within the realm of medical imaging. However, these models typically require a significant amount of labeled data for training, which can be time-consuming and labor-intensive, particularly in clinical settings. Additionally, variations in imaging devices can introduce non-standard data, negatively impacting the model's performance and generalizability. To address these challenges and enhance diagnostic flexibility and accuracy, there is a need for a robust framework capable of learning from both labeled and unlabeled data.

We propose LREN-MedNet, an innovative framework that integrates a *Latent Representation Extraction Network (LREN)* with *Adaptive ROI Transformation (ART)* to overcome these limitations. ART pre-processes multi-source endoscopic ultrasound (EUS) images to reduce inter-device variability by normalizing regions of interest (ROIs) and eliminating extraneous pixels. Meanwhile, LREN minimizes the dependence on labeled datasets by pre-training the model with unlabeled data through self-supervised learning techniques such as contrastive learning and masked autoencoding. The combination of these two components in LREN-MedNet facilitates improved feature extraction and generalization.

The Vision Transformer (ViT) [3], which has been pre-trained for the classification of benign, malignant, and normal breast ultrasound cases, serves as the cornerstone of the LREN-MedNet architecture. The effectiveness of our approach is validated through extensive testing using the *BUSI* dataset, which comprises 780 annotated ultrasound images with corresponding segmentation masks. In all instances, LREN-MedNet surpasses leading techniques in precision, robustness, and adaptability, effectively addressing critical challenges associated with ultrasound-based breast cancer diagnosis.

By integrating cutting-edge self-supervised learning with automated preprocessing [4], LREN-MedNet paves the way for potentially transformative advancements in computer-aided

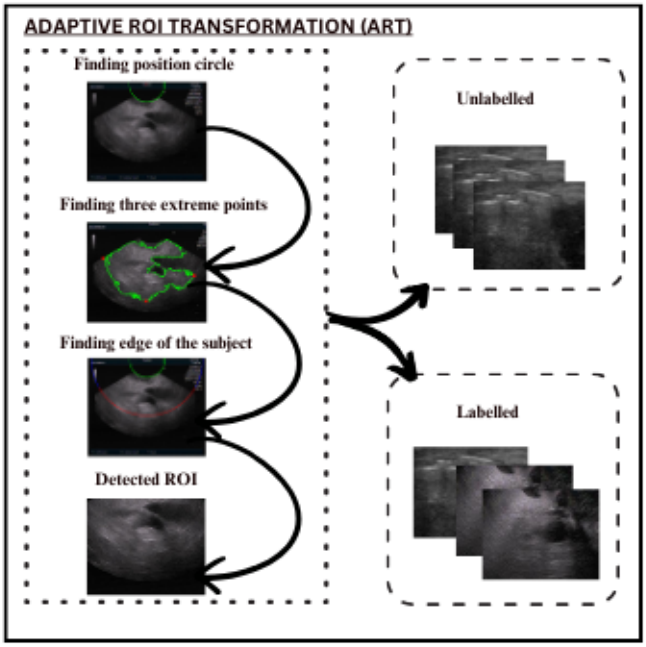

Fig. 1. ART Workflow

medical diagnosis. This framework presents an attractive solution for minimizing variability in medical imaging datasets, as it not only enhances diagnostic accuracy but also reduces dependence on manually annotated data. Future studies will concentrate on expanding the use of LREN-MedNet to 3D imaging and real-time clinical settings, as well as applying it to ROI-free designs.

II. Methodology

A. Adaptive ROI Transformation (ART)

Adaptive ROI Transformation (ART) is a preprocessing method aimed at normalizing endoscopic ultrasound (EUS) images by reducing inter-device variability and isolating diagnostically relevant regions of interest (ROIs) [5]. Normalization is necessary to ensure that subsequent tasks are invariant and robust to variations introduced by different imaging devices, including Pentax, Fujifilm, Olympus, and Aloka. The ART pipeline includes ROI detection [6], transformation, and normalization.

1) Device-Specific Variability: EUS images vary in resolution and quality across devices. Let $x \in \mathbb{R}^{H \times W}$ denote an input image of resolution $H \times W$, where:

- Pentax: 764×572,
- Fujifilm: 1916×1196,
- Aloka: 768×576,
- Olympus: 760×568.

To ensure uniformity, the ART pipeline first identifies the positioning circle, which defines the ROI, and transforms the circular ROI into a standardized rectangular format.

2) Obtaining the Boundaries: To accurately extract the Region of Interest (ROI) in Endoscopic Ultrasound (EUS) images, we follow a structured approach to boundary detection:

- **Position Circle Detection:** The black positioning circle at the top of the EUS image is detected using:
  - **Preprocessing:** The image is cropped to remove text interference, followed by median filtering to reduce noise.
  - **Grayscale Conversion:** The filtered image is converted to a single-channel grayscale image:

$$I_{\text{gray}}(x,y) = 0.3I_R(x,y) + 0.59I_G(x,y) + 0.11I_B(x,y) \quad (1)$$

  - **Edge Detection:** The Canny edge detection method is applied to extract boundaries.
  - **Circle Fitting:** The Hough Circle Transform is used to identify the positioning circle, selecting the topmost detected circle as the reference.
- **Edge of Subject Detection:** To extract the main subject:
  - **Median Filtering:** Further noise suppression on grayscale images.
  - **Binary Thresholding:** A thresholding technique is applied to segment the ultrasound signal.
  - **Contour Detection:** The largest connected contour is identified as the primary subject.
  - **Extreme Points Selection:** Three key points are extracted from the contour:
    - $P_1(x_1, y_1)$ - Leftmost point
    - $P_2(x_2, y_2)$ - Rightmost point
    - $P_3(x_3, y_3)$ - Bottommost point
- **Outer Circle Radius Calculation:** The ROI is defined with left-right symmetry, where the farthest extreme point from the top circle determines the lateral boundary. The outer circle radius R is computed as:

$$R = \frac{\max(|c_x - x_1|, |c_x - x_2|) + |c_y - y_3|}{1} \quad (2)$$

The final ROI parameters include:

  - Center: (c_x, c_y)
  - Radius range: $[r, R]$
  - Angular range: $[30^\circ, 150^\circ]$

3) Polar-to-Rectangle Transformation: To facilitate deep learning model training, the extracted sector-shaped ROI is transformed into a rectangular format using polar coordinate transformation.

  - **Polar Representation:** The center of the positioning circle (c_x, c_y) is set as the origin, and each point in the ROI is expressed in polar coordinates as:

$$O(\alpha, \theta) = f(c_x + \alpha\cos\theta, c_y + \alpha\sin\theta) \quad (3)$$

where $\alpha \in [r, R]$ and $\theta \in [30^\circ, 150^\circ]$.

– **Discretization and Resampling:** To avoid image distortion, a fine step size is used when converting the sector ROI into a rectangular form:

$$O(i,j) = f'\Big(c_x + (\alpha_{\min} + \alpha_{\text{step}} i)\cos\left(\theta_{\min} + \theta_{\text{step}} i\right), \\ c_y + (\alpha_{\min} + \alpha_{\text{step}} j)\sin\left(\theta_{\min} + \theta_{\text{step}} j\right)\Big) \tag{4}$$

– **Output Image Dimensions:** The width W and height H of the rectangular output image are calculated as:

$$W = \frac{\alpha_{\max} - \alpha_{\min}}{\alpha_{\text{step}}} + 1, \quad H = \frac{\theta_{\max} - \theta_{\min}}{\theta_{\text{step}}} + 1 \tag{5}$$

ensuring uniform spatial representation.

By eliminating the unpredictability associated with multi-source EUS devices, ART ensures that diagnostically significant traits are consistently represented. This conversion facilitates stable feature learning in downstream tasks, reducing inter-device bias. Additionally, ART enhances computational efficiency by concentrating on relevant anatomical regions and minimizing unnecessary data processing. The output from ART is provided as input to the Latent Representation Extraction Network (LREN), which guarantees that self-supervised pretraining is conducted on standardized, high-quality images. This seamless integration is crucial for the overall performance of LREN-MedNet.

B. Latent Representation Extraction Network (LREN)

Annotating medical imaging data, such as endoscopic ultrasound (EUS) images, demands significant time and resources. To address this challenge, the Latent Representation Extraction Network (LREN) [7] combines two complementary self-supervised techniques: masked autoencoding and contrastive learning [8]. This approach allows the model to learn stable feature representations from unlabeled data [9]. This section provides a comprehensive introduction to LREN, detailing its main components and associated equations.

1) Patch Generation: The input EUS image is partitioned into non-overlapping patches to enable local and global representation learning. A random subset of the patches is masked, and the rest of the visible patches are handled by the encoder [10]. In accordance with ViT [13], the masking ratio is 75%, as depicted in the input pipeline:

$$x_q = \text{Mask}(x), \quad x_k = \text{Mask}(x). \tag{6}$$

2) Encoders and Momentum Update: The LREN employs two encoders: a query encoder E_q and a momentum-based key encoder E_k, both based on the Vision Transformer (ViT) architecture. The query encoder generates feature representations q for the masked input, while the key encoder provides stable target features k. The momentum update for E_k ensures smoother parameter updates:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \tag{7}$$

where $m \in [0,1)$ is the momentum coefficient, empirically set to $m = 0.999$.

3) Contrastive Learning Loss: Contrastive learning aligns embeddings of positive pairs (augmented views of the same image) and separates embeddings of negative pairs. For each query feature q, a positive key k^+ and a set of negative keys $\{k_i\}$ are sampled. The contrastive loss is defined as:

$$L_q = -\log \frac{\exp(q \cdot k^+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)}, \tag{8}$$

where τ is a temperature hyperparameter, empirically set to 0.07, and K is the number of negative samples.

4) Masked Autoencoding Loss: Masked autoencoding complements contrastive learning by focusing on reconstructing the missing patches of the input image [11]. The encoder processes visible patches, generating latent representations z_q, which are passed to a lightweight decoder for reconstruction. The reconstruction loss is calculated as:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{x}_i - x_i\|^2, \tag{9}$$

where N is the number of pixels, and $\hat{x}_i$ and x_i denote the reconstructed and original pixel values.

5) Combined Unsupervised Loss: The overall unsupervised loss combines the contrastive and reconstruction losses:

$$L_u = w_1 L_{\text{MSE}} + w_2 L_q, \tag{10}$$

where w_1 and w_2 are learnable weights initialized to 0.4 and 0.6, respectively.

TABLE I
COMPARISON OF SELF-SUPERVISED LEARNING TECHNIQUES

Technique	**Type**	**Purpose**
Contrastive Learning	Instance Discrimination	Learn robust representations
Masked Autoencoding	Patch Reconstruction	Learn local/global features
Clustering-based SSL	Feature Grouping	Improve feature separation
Generative Pretraining	Data Augmentation	Generate diverse representations
Hybrid (LREN)	Contrastive + MAE	Combine best of both

The AdamW optimizer [12] is utilized to train the LREN with a learning rate of 10^{-4} and a weight decay of 5×10^{-2}. During the training process, the input images undergo various data augmentation techniques [13] , such as normalization, random cropping, and horizontal flipping. After pretraining [14], labeled datasets are

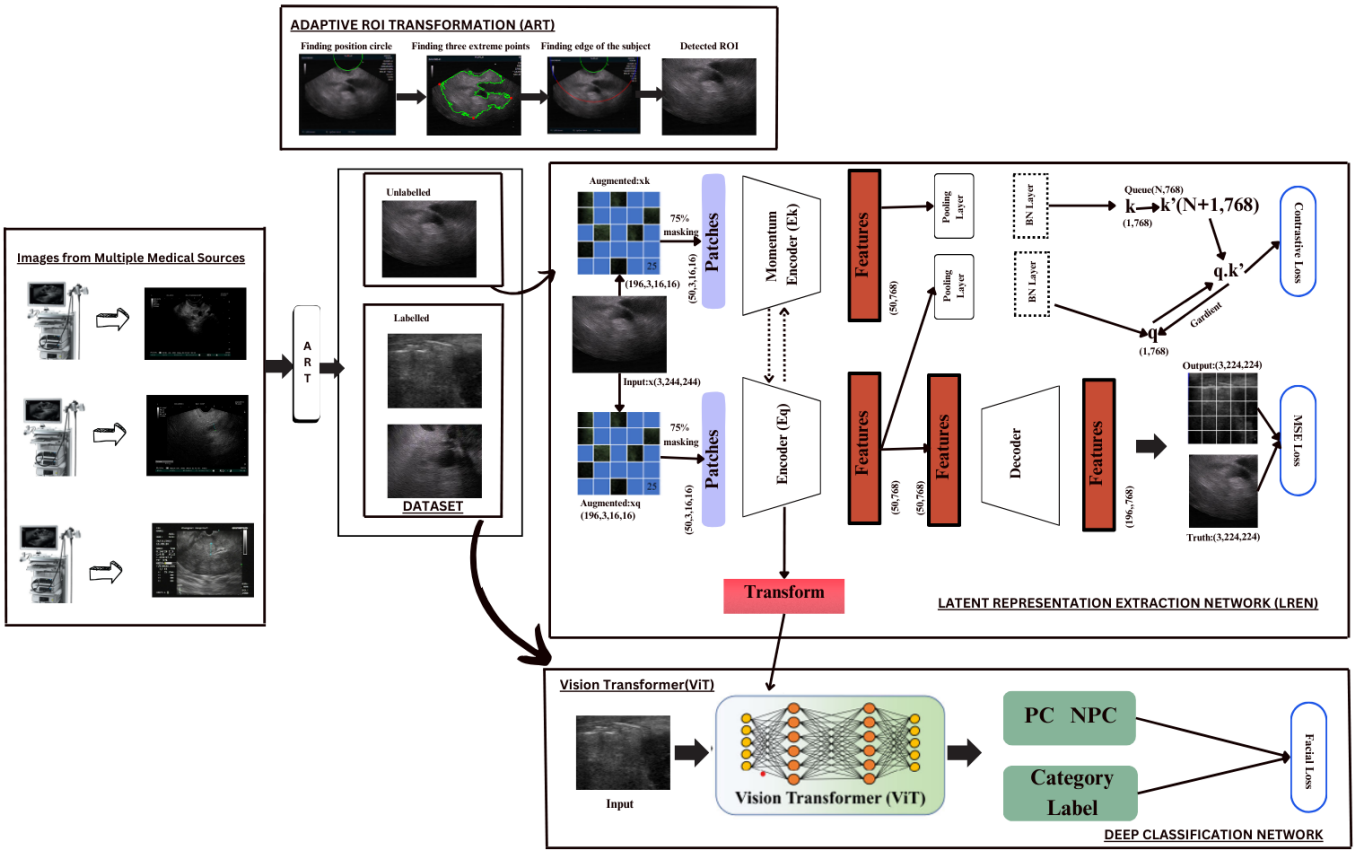

Fig. 2. Overview of LREN

employed to fine-tune the encoder for subsequent tasks, including segmentation and classification. By effectively leveraging unlabeled data, this dual self-supervised approach enables LREN-MedNet to develop reliable and transferable feature representations, significantly enhancing diagnostic performance on EUS datasets from diverse sources.

C. Vision Transformer (ViT)

Vision Transformer (ViT) represents a groundbreaking deep learning model that applies the transformer architecture—previously utilized in Natural Language Processing (NLP)—to image processing applications. Instead of using convolutional layers [15] , ViT segments an image into non-overlapping patches and treats each patch as a token, employing self-attention techniques to analyze the token sequence [16]. To preserve spatial information, positional embeddings are incorporated after each patch is linearly projected into a fixed-dimensional embedding space [17]. ViT demonstrates exceptional effectiveness for tasks such as object detection [18] and image classification [19], owing to its ability to capture global context and long-range dependencies. Focal loss is commonly employed to tackle challenges like class imbalance, which are prevalent in many vision tasks. Focal Loss is defined as follows:

$$\mathcal{L}_{\text{Focal}}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \tag{11}$$

where

$$p_t = \begin{cases} \hat{p}, & \text{if } y = 1, \\ 1 - \hat{p}, & \text{otherwise}, \end{cases} \tag{12}$$

helps prioritize hard-to-classify examples by reducing the impact of well-classified ones. Here, $\hat{p}$ represents the predicted probability, y is the true label, α_t is a balancing factor, and γ adjusts the focus on difficult samples. By integrating ViT with such loss functions, its robustness and performance are further enhanced for real-world applications.

III. Experimental Results

A. BUSI Dataset Overview

The Breast Ultrasound Image (BUSI) dataset is a well-curated collection designed for breast cancer detection and classification. It consists of ultrasound images categorized into three classes:

– **Benign:** Breast lesions that are non-cancerous.

– **Malignant:** Cancerous breast lesions requiring medical intervention.
– **Normal:** Healthy breast tissue without abnormalities.

The dataset comprises a total of 780 images, each having an associated segmentation mask for precise lesion localization. The distribution across classes is as follows:

– **Benign:** 437 images
– **Malignant:** 210 images
– **Normal:** 133 images

To enhance clinical applicability, the images were collected from real-world ultrasound scans, exhibiting variations in size, shape, and texture of breast lesions. Additionally, the dataset includes manually annotated segmentation masks that assist in lesion boundary detection, supporting both classification and segmentation tasks.

The BUSI dataset is widely used for training and evaluating deep learning models in medical image analysis, particularly for breast cancer detection, due to its diverse representation of breast tissue abnormalities. The Breast Ultrasound Images Dataset (BUSI) is publicly available on Kaggle and can be accessed for research purposes. It can be downloaded from the following link: https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset

B. Experimental Results on BUSI Dataset

This section presents the experimental results obtained from the classification of the BUSI dataset. The performance of the model is evaluated using the classification results table, testing loss curve, and testing accuracy curve.

1) Classification Results Table Analysis: The classification results are summarized in Table II, which provides insights into how well the model classifies benign, malignant, and normal cases.

TABLE II
CONFUSION MATRIX FOR BUSI DATASET CLASSIFICATION

True Label \ Predicted Label	Benign	Malignant	Normal
Benign	153	16	7
Malignant	33	49	2
Normal	13	4	37

– **Benign Cases:** Among 176 benign samples, the model correctly classified 153 cases. However, 16 benign cases were misclassified as malignant, and 7 were incorrectly labeled as normal.
– **Malignant Cases:** Out of 84 malignant samples, 49 were accurately classified. However, 33 were misclassified as benign, and 2 were predicted as normal. The misclassification between benign and malignant suggests some overlap in their features.
– **Normal Cases:** The model correctly identified 37 out of 54 normal cases. However, 13 normal cases were misclassified as benign, and 4 were predicted as malignant. This misclassification may be due to similarities between normal and benign structures in ultrasound images.

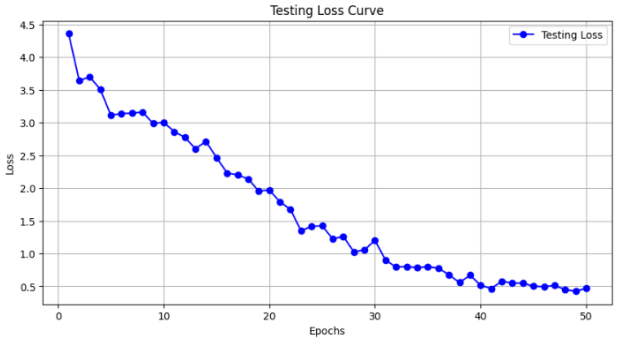

Fig. 3. Loss of the model

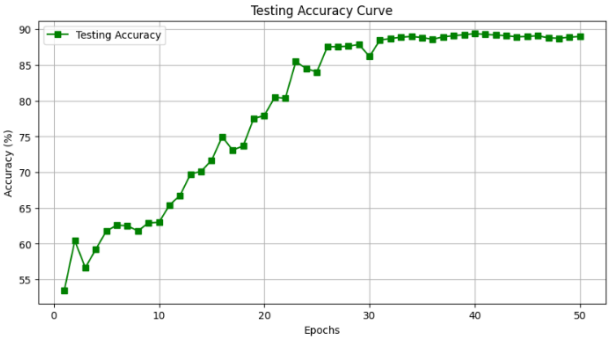

Fig. 4. Accuracy of the model

2) Testing Loss Analysis: The testing loss curve, shown in Figure 3, illustrates the decrease in loss over the training epochs. Initially, the loss is high, indicating that the model is still learning. As the training progresses, the loss steadily decreases, demonstrating that the model is effectively learning patterns in the data. Around epoch 30, the loss stabilizes, indicating convergence.

3) Testing Accuracy Analysis: The testing accuracy curve, shown in Figure 4, represents the improvement in accuracy over epochs. The accuracy starts at approximately 55% and increases consistently. Around epoch 25, there is a significant jump, reaching an accuracy of nearly 89%. After epoch 30, the accuracy stabilizes, indicating that the model has reached optimal learning and is no longer improving significantly.

4) Overall Performance Evaluation: The classification results table shows that the model performs well in identifying benign cases but has some misclassification issues when distinguishing between malignant and normal cases. The testing loss curve suggests effective learning with convergence around epoch 30, while the testing accuracy curve demonstrates a steady improvement, ultimately stabilizing at around 89% accuracy. These results indicate

that the model is robust but may require further fine-tuning or additional features to improve its ability to distinguish between malignant and normal cases.

IV. Conclusion

In conclusion, the integration of Adaptive ROI Transformation (ART) and the Latent Representation Extraction Network (LREN) within the LREN-MedNet model represents a significant advancement in the automated diagnosis of breast cancer. Utilizing the BUSI dataset, LREN-MedNet demonstrates superior diagnostic accuracy and robustness compared to existing state-of-the-art methods. The model effectively addresses challenges such as variations in ultrasound imaging and the limited availability of labeled medical data. By accurately classifying benign, malignant, and normal breast tissue cases, LREN-MedNet underscores its potential for real-world clinical applications. Furthermore, this study sets the foundation for future research in improving ultrasound-based breast cancer diagnosis, enhancing interpretability, and extending the approach to other medical imaging modalities.

Acknowledgment

The infrastructure and essential resources that facilitated this research were provided by Mepco Schlenk Engineering College, to which the authors are sincerely grateful. They would also like to extend their appreciation to the faculty and staff of the Department of Artificial Intelligence for their continued support and valuable insights throughout the project. Additionally, the authors acknowledge the utilization of publicly available datasets, which were integral to the development and evaluation of this framework.

References

[1] K. Park, W. Chen, M. A. Chekmareva, D. J. Foran, and J. P. Desai, "Electromechanical coupling factor of breast tissue as a biomarker for breast cancer," *IEEE*, 2023.

[2] A. Melek, S. Fakhry, and T. Basha, "Spatiotemporal mammography-based deep learning model for improved breast cancer risk prediction," *IEEE*, 2023.

[3] L. Gai, W. Chen, R. Gao, Y.-w. Chen, and X. Qiao, "Using vision transformers in 3-d medical image classifications," *School of Control Science and Engineering, Shandong University*, 2024.

[4] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi, "Big self-supervised models advance medical image classification," *Google Research and Health*, 2024.

[5] C. Mercan, B. Aygunes, S. Aksoy, L. G. Shapiro, E. Mercan, D. L. Weaver, and J. G. Elmore, "Deep feature representations for variable-sized regions of interest in breast histopathology," *IEEE*, June 2021.

[6] Q. Sun, Y. Liu, C. Sheng, H. Wang, and X. Lu, "Regions of interest extraction for hyperspectral small targets based on self-supervised learning," *IEEE*, 2024, member, IEEE.

[7] I. Xi Cen, Student Member, I. Yachao Li, Member, Z. Han, T. Gu, I. Peng Zhang, Member, and T. Cai, "Self-supervised learning method for sar multiinterference suppression," 2024.

[8] S. Kumar, A. Phukan, and A. Sur, "IPCL: Iterative Pseudo-Supervised Contrastive Learning to Improve Self-Supervised Feature Representation," *Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, India*, 2024, *Equal contribution.

[9] X. Sun, P. Chen, L. Chen, C. Li, T. H. Li, M. Tan, and C. Gan, "Masked Motion Encoding for Self-Supervised Video Representation Learning," 2024, *Equal contribution, †Corresponding author.

[10] J. Wang, Z. Yin, P. Hu, A. Liu, R. Tao, H. Qin, X. Liu, and D. Tao, "Defensive Patches for Robust Recognition in the Physical World," 2024, †Equal contribution, *Corresponding author.

[11] Z. Li, Z. Xue, M. Jia, X. Nie, H. Wu, M. Zhang, and H. Su, "DEMAE: Diffusion-Enhanced Masked Autoencoder for Hyperspectral Image Classification With Few Labeled Samples," *IEEE*, 2024, member, IEEE and Senior Member, IEEE.

[12] A. Javed, I. Rashid, S. Tahir, S. Saeed, A. M. Almuhaideb, and K. Alissa, "AdamW+: Machine Learning Framework to Detect Domain Generation Algorithms for Malware," *IEEE*, 2024, senior Member, IEEE.

[13] I. Nima Tajbakhsh, Member, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and I. Jianming Liang, Senior Member, "Convolutional neural networks for medical image analysis: Full training or fine tuning?"

[14] Y. Yao, B. Yu, I. Chen Gong, Member, and I. Tongliang Liu, Senior Member, "Understanding how pretraining regularizes deep learning algorithms," 2024.

[15] R. Wang, Z. Li, J. Cao, T. Chen, and L. Wang, "Convolutional Recurrent Neural Networks for Text Classification," 2024.

[16] G. J. Ferdous, K. A. Sathi, M. A. Hossain, and M. A. A. Dewan, "SPT-Swin: A Shifted Patch Tokenization Swin Transformer for Image Classification," 2024, member, IEEE.

[17] P. V. Arun and K. M. Buddhiraju, "A Deep Learning Based Spatial Dependency Modelling Approach Towards Super-Resolution," 2024, ph.D. student and Professor.

[18] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, cnns and vision transformers: A review," 2024.

[19] P. V. Arun and K. M. Buddhiraju, "A deep learning based spatial dependency modelling approach towards super-resolution," 2024, ph.D. student and Professor.