DOI: 10.21917/ijdsml.2026.0193

ANATOFORMER: ENHANCING BREAST CANCER DIAGNOSIS USING TRANSFORMER-BASED SPATIAL ANALYSIS AND LATENT REPRESENTATION LEARNING

D. Monica Seles, Prahanya Selvakumar, M. Kawena
Department of Artificial Intelligence and Data Science, Mepco Schlenk Engineering College, India

Abstract

The serious risks to public health from the rapid spread of misleading information brought on by our increasing reliance on digital healthcare information. Recognizing such misinformation is essential in the medical industry to ensure that accurate and trustworthy information is disseminated. We suggest an ideal deep learning-based model combining AnatoFormer and LREN-MedNet for ROI-Free Breast Cancer Diagnosis using self-supervised learning and anatomical-aware feature extraction. Our suggested AnatoFormer-LREN system uses Masked Autoencoder (MAE) pretraining and Contrastive Learning (MoCo) to enhance feature extraction from unlabeled ultrasonic images. The AnatoFormer design improves diagnostic accuracy and model interpretability by accurately capturing intra-layer and inter-layer spatial interactions in breast ultrasound images. Because the resulting representations are employed in a fully automated pathway, eliminating the need for manual ROI annotation, the model is more practical for clinical applications. Additionally, optimization techniques like self-supervised training and fine-tuning methodologies have been employed to increase the resilience of our model. The models performance was assessed on the BUSI dataset and it performs better in classification than transformer-based and traditional CNNs. The suggested AnatoFormer-LREN structure offers a new understandable and effective alternative to automated breast cancer diagnosis advancing artificial intelligence in clinical decision-making and medical imaging.

Keywords:

Breast Cancer Diagnosis, AnatoFormer, LREN-MedNet, Self-Supervised Learning, Contrastive Learning, Masked Autoencoder (MAE), Ultrasound Imaging, Deep Learning, ROI-Free Diagnosis, Medical Image Analysis, Clinical Decision Support

1. INTRODUCTION

Artificial intelligence (AI) and deep learning (DL) have revolutionized medical imaging by enabling automated feature extraction and highly accurate disease classification. However, the majority of deep learning models for diagnosing breast ultrasonography still rely on manually segmented ROIs, which restricts their use in real-world clinical settings.Since breast cancer is one of the most common and deadly illnesses in the world early and precise detection is essential to improving patient outcomes. The non-invasiveness radiation-free nature and affordability of breast ultrasound (BUS) make it one of the most popular imaging modalities.

Regretfully conventional CAD programs frequently depend on pre-established Regions of Interest (ROIs) necessitating radiologists manual annotations. Clinical efficiency is impacted automation is constrained and radiologists variability is increased by this manual participation. Manual Region of Interest (ROI) annotations which need expert input to differentiate tumors in ultrasound images have historically been the mainstay of breast cancer detection. ROI-based approaches however are timeconsuming arbitrary and might not be transferable across datasets.

Thanks to developments in deep learning and self-supervised learning automated ROI-free breast cancer diagnosis is now a dependable and effective substitute. ROI-Free models evaluate breast ultrasound (BUS) images without the need for human ROI selection by utilizing transformer based topologies and self-supervised learning. These models enable fully automated diagnosis by directly extracting tissue shapes spatial relationships and contextual information from entire ultrasound images. Due to its lack of dependence on pre-established lesion labels this method is more scalable interpretable and therapeutically advantageous.

Using ROI-Free a transformer based deep learning model known as AnatoFormer was developed to identify breast cancer in ultrasound pictures. AnatoFormer models the vertical (inter-layer) and horizontal (intra-layer) spatial correlations within breast tissue structures to capture anatomical prior knowledge which sets it apart from traditional CNN-based models. Its use of multi-head self-attention to enhance feature extraction and diagnostic interpretability makes it a very successful classification system for breast ultrasounds.

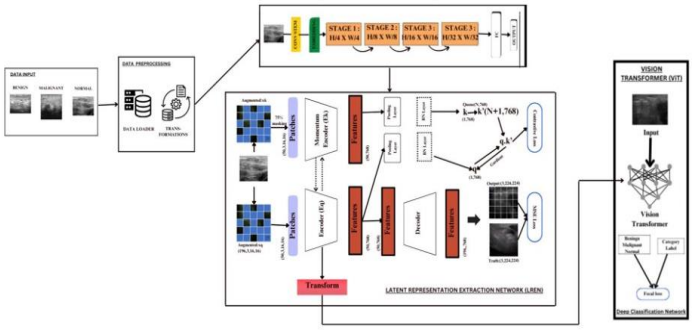

Fig.1. Overview of LREN

To enhance feature representation the LRE-Net self-supervised learning framework combines Masked Autoencoder (MAE) pretraining with Contrastive Learning (MoCo). While MAE reconstructs missing image patches to capture context-aware representations MoCo compares similar and dissimilar samples to improve feature discrimination. This hybrid approach enables robust pretraining on unlabeled ultrasound images and improves the accuracy of subsequent classification tests. The deep learning architecture Vision Transformer (ViT) uses the self-attention mechanism of transformers to address image processing problems. Unlike Convolutional Neural Networks(CNNs) which depend on local receptive fields ViT divides images into patches and treats each patch as a token much like words in NLP models. ViT is therefore perfect for applications like object detection

image classification and medical imaging since it can capture long-range dependencies and global contextual relationships.

2. RELATED WORK

The latest breakthroughs in deep learning, such as transformer-based models and self-supervised models, have profoundly enabled medical image analysis, particularly for breast ultrasound diagnosis. HoVer-Trans proposed an anatomy-aware transformer model for ROI-free breast cancer detection with the capability of representing intra- and inter-layer tissue relationships, enabling automated diagnosis without human annotation [1]. DSMT-Net utilized two self-supervised learning methods on multi-source endoscopic ultrasound images and demonstrated strong diagnostic performance across institutions by leveraging masked autoencoders and contrastive learning [2]. The electromechanical coupling factor of breast tissue was explored as a quantitative biomarker in a novel direction, uncovering insights into tumor biomechanics for enhanced diagnostic accuracy [3]. A spatiotemporal deep learning model based on mammography was suggested for risk estimation, integrating temporal information from previous examinations with spatial features to improve early detection rates [4]. Vision Transformer (ViT) has been shown to be effective in 3D medical image classification, offering superior long-range spatial reasoning compared to conventional CNNs due to its self-attention mechanism [5].

Recent advances in contrastive learning and self-supervised learning (SSL) have enormously boosted medical image analysis, especially in situations where there is limited labeled data. Big SSL Models' work [6] brought forward the Multi-Instance Contrastive Learning (MICLe) approach that makes use of various views of a single pathology in order to enhance SSL performance and, in the process, achieve significant improvement in chest X-ray and dermatology classification tasks. To solve variable-sized region issues, a deep CNN-based approach was introduced to model ROIs in breast histopathology with improved diagnostic performance via feature combination of diverse fixed-size patches [7]. In hyperspectral images, a model for ROI extraction based on SSL performed better than classic approaches in detecting small targets without the need for large amounts of labeled data with improved generalization and fewer false positives [8]. For Synthetic Aperture Radar (SAR) signal recovery, a new SSL model based on convolutional autoencoder (LocNet) and reconstruction model (RecNet) with U-Net effectively removed electromagnetic interferences, enhanced the quality of signals with minimal computational costs [9]. The IPCL framework [10] also combined iterative pseudo-labeling with contrastive learning and greatly enhanced the self-supervised image classification and feature discrimination accuracy in various benchmark datasets.

Recent research has seen the emergence of deep learning into image and signal processing tasks for more effective selfsupervised learning (SSL), transformer models, and domain-related innovation. Masked Motion Encoding (MME) emerged as a method of video motion and appearance dynamics modeling that could register a new state-of-the-art on action recognition datasets by recovering motion trajectories using masked modeling methods [11]. Defensive Patches added a robust way to protect visual recognition models against adversarial attacks, particularly boosting accuracy and robustness in noisy environments [12]. DEMAE, an augmented masked autoencoder with diffusion, was put forward for hyperspectral image classification in sparse data environments, combining diffusion learning, transformers, and a novel SNR-boosted loss function [13]. In cybersecurity, the AdamW+ framework improved domain generation algorithm (DGA) detection with better weight decay and better generalization over traditional optimizers for malware detection tasks [14].

Subsequent work investigated model generalization and transferability. Fine-tuned CNNs performed better than fully trained networks on all fronts in a variety of different medical imaging tasks, especially where the labeled dataset was small due to the success of transfer learning [15]. Pretraining was found to be a robust regularizer, which enhances the generalization and stability of deep models in weakly supervised settings [16]. CRNNs combined RNNs and CNNs to achieve improved text classification more effectively, combining sequential and spatial patterns well into natural language processing tasks [17]. SPT-Swin combined Shifted Patch Tokenization and Swin Transformers to achieve improved performance on several datasets with improved efficiency [18]. The spatially dependent deep learning-method improved super-resolution in low-data conditions with autoencoders for learning to restore high-fidelity images [19]. Last but not least, CNNs and Vision Transformers were also explored for object detection, with dramatic gains in contextual awareness and robustness over traditional detection models [20]. Emergent work has built on top of the effect of self-supervised learning (SSL), Vision Transformers (ViTs), and halfand-half deep models on several fields of medical image testing. Testing low-resource conditions, SSL models like SimCLR, DCLW, SimSiam, and VICReg were tested against small datasets like BreaKHis and PneumoniaCXR and emerged to be on par with conventional transfer learning [21]. A VGGA-ViT model that incorporated VGG and ViT blocks improved breast ultrasound image classification accuracy, highlighting local and global contextual learning [22]. Equivalently, SSL methods such as SimCLR and rotation learning were used for the diagnosis of skin cancer with state-of-the-art improvements over supervised baselines [23]. In histopathology, Deep Fusion Vision Transformer (DFViT) integrated CNN-based local feature extraction with ViTs to classify breast cancer with robust performance across a variety of datasets [24]. Masked Transformers also facilitated high-accuracy segmentation and classification from unlabelled data, establishing a new benchmark for medical image representation learning [25]. For survival prediction of lung cancer, a two-transformer encoder model with Layer-wise Class Token Attention (LCA) was introduced to enhance generalization and multi-scale representation [26]. StyleGAN-enhanced SSL models were employed to classify chest X-ray images with enhanced feature fusion and generalization in low-data settings [27].

MA-Transformer, which is a hybrid CNN-Transformer architecture, incorporated a multi-stage aggregation mechanism to preserve local and global context, outperforming current state-of-the-art segmentation techniques [28]. The ColorMe pipeline applied SSL-based multi-task learning to scopy images and performed better than supervised methods in such tasks as cervix classification and lesion segmentation [29]. Uncertainty modeling

was addressed via Swin Transformer V2 with Monte Carlo dropout, improving paraspinal MRI anatomical landmark detection and offering insights with random forest grading [30]. Finally, bi-level attention and information flow were employed in a Transformer-based registration network to create state-ofthe-art performance for deformable medical image registration from brain MR data [31].

3. METHODOLOGY

3.1 ANATOFORMER BLOCK: ANATOMY-AWARE FEATURE EXTRACTION

AnatoFormer model is a peculiar deep learning architecture for breast ultrasound imaging that can ingest full ultrasound images without a previously defined Region of Interest. By keeping anatomical structures and spatial relationships in consideration, it automates and robustifies the feature extraction relevant to breast cancer diagnosis. In other words, compared to conventional models by relying greatly on local feature extraction, AnatoFormer is a hybrid approach combining a Transformerbased architecture and convolutional layers [32], to capture global and local spatial dependencies together [33]. Thus, it can find tumors within the same precision as any other model while being much more computationally efficient.

3.1.1 Patch and Strip Embeddings:

In order to efficiently capture anatomical structures, the input breast ultrasound (BUS) image is subjected to a multi-stage embedding mechanism after which the acquired spatial representation is incremented:

- Patch Embedding: Analogous to Vision Transformers (ViTs) [34], the image is divided into non-overlapping patches, meanwhile retaining local texture information vital for the detection of fine-grained tissue patterns and subtle morphological variations in the course of breast lesions.
- Horizontal Strip Embedding: Horizontal slices of the patches are formed in order to conform to the anatomical layers of the tissue, thus allowing for the extraction of intra-layer dependencies. This finds significance in recognizing uniform tissue structures which in turn discriminate between normal and abnormal breast tissues.
- Vertical Strip Embedding: Vertical slices are extracted, capturing inter-layer dependencies that would enable the model to follow tumor invasion across several anatomical layers. This becomes crucial in assessing tumor infiltration and structural distortion in malignant cases.

Fig.2. Patch and Strip Embeddings: Three embedding types used in AnatoFormer. Horizontal embedding aligns with anatomical layers, patch embedding extracts local features, and vertical embedding captures inter-layer dependencies.

3.1.2 AnatoFormer Core Processing Block:

Different geographical dependencies are used to learned the four specialized branches that comes in terms with the AnatoFormer core processing block:

- H (Horizontal): Spatial correlations within the same tissue layer were recorded, making it possible to correctly identify any pattern connected to that layer. This is important because organ consistency allows for the classification of benign and malignant zones.
- V (Vertical): This model is useful for detecting cancers that pierce tissue at different depths because it replicates the transverse spread of a tumor between tissue layers. This makes it more easy to find tumors that has invasive growth patterns.
- H2V (Horizontal-to-Vertical Fusion): This method enhances anatomical connections between layers by transferring horizontal feature representations into the vertical domain. The vertical extension of horizontally aligned tissue characteristics is also evaluated.
- V2H (Vertical-to-Horizontal Fusion): By adding vertical dependencies into the horizontal domain, this method enables localized feature extraction [35] within the same tissue layer. This integration enhances multi-dimensional spatial learning.

Each of these branches independently applies a Transformer encoding function prior to feature fusion.The mathematical transformations that govern this process are:

$$zh,l = \text{Trans}(zh,l-1) \quad (1)$$

$$zv,l = \text{Trans}(zv,l-1) \quad (2)$$

$$zh2v,l = \text{Trans}(\text{Trans}(zh,l + zh2v,l-1) + zv,l) \quad (3)$$

$$zv2h,l = \text{Trans}(\text{Trans}(zv,l + zv2h,l-1) + zh,l) \quad (4)$$

3.1.3 Convolutional Feature Fusion: Convolutional Feature Fusion:

Transformers are not very good at fine-grained spatial resolution, which is very important for medical image analysis.To overcome that issue, AnatoFormer integrates convolutional layers [37] after the processing of transformer based features. This extra module concentrates on local textures, enhances edge definitions and increases diagnostic accuracy. The final feature fusion is expressed as:

$$z_s+1 = \text{Conv}(z_h2v,s,l,z_v2h,s,l) \quad (5)$$

where z_{s+1} represents the final fused feature map before classification.

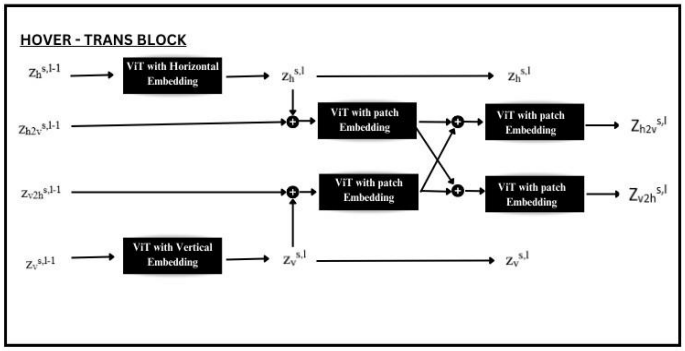

Fig.3. AnatoFormer Block Architecture

As convolutional layers offer both local feature improvement and global spatial reasoning, the model is ideal for medical picture interpretation.

A four-branch transformer [36] block designed for capturing anatomical dependencies in breast ultrasound images. The block consists of horizontal and vertical branches along with inter-branch fusion mechanisms, enhancing spatial awareness for improved feature extraction.

3.1.4 Advantages of AnatoFormer Block:

It provides many benefits for breast ultra sound analysis:

- ROI-Free Processing: this model works on the entire ultrasound images without needing any manual ROI extraction, smooth-running the workflows of clinic and decreasing the potential human error.
- Anatomical Awareness: The hybrid Transformer-based approach constructively builds both intra-layer and inter-layer dependencies, making sure of detailing every little detail of the differentiation between normal, benign and malignant tissues.
- Robust Feature Representation: The integration of Transformers [38] with CNNs provide more advanced generalization across diverging imaging conditions and datasets. This will help the model learn unvarying and discriminative features that will help to increase the robustness of it.
- Efficient Multi-Scale Processing:all the four-branch from the architecture simultaneously extract both the local and global features, diagnostic reliability and model interpretability. This multi-scale feature extraction method identified both minor and major anomalies.
- Improved Clinical Applicability: By exploiting a combination of self-attention mechanisms and convolutional feature refinements, AnatoFormer shows a more advanced performance in tumor classification, segmentation and malignancy assessment.

By integrating anatomical feature extraction with Transformer-based architecture and CNN-driven feature clarification, AnatoFormer represents a superior performance in breast ultrasound imaging and automated cancer diagnosis. The ability to model spatial dependencies at multiple scales effectively makes it a better tool for real world model applications. The ability of the model to function an ROI-free manner additionally intensifies the practical usability of clinical environments, decreasing the reliance of manual annotations and improving the workflow automation.

3.2 LATENT REPRESENTATION EXTRACTION NETWORK (LREN)

Annotating medical imaging data, especially breast ultrasound (BUS) pictures, demands a significant time and expensive commitment due to the complexity of lesion features and the need for trained radiologists. To overcome this challenge, the integrated model incorporates the Latent Representation Extraction Network (LREN), which extracts meaningful representations from both labeled and unlabeled ultrasound images using self-supervised learning techniques [39]. LREN combines two complementary self-supervised learning methods: contrastive learning and masked autoencoding. This dual learning strategy significantly enhances the network's ability to learn discriminative, generalizable, and transferable features while also increasing the robustness of breast cancer classification in ultrasound imaging.

3.2.1 Patch Generation:

After AnatoFormer performs feature extraction, the resulting representation is divided into nonoverlapping patches to improve both local and global feature understanding. A randomly selected subset of these patches is masked, whereas the visible patches are processed through the encoder for further analysis. This method ensures that the model perceives contextual relationships within the ultrasound images without requiring explicitly labeled data. Consistent with the Vision Transformer (ViT) model, a masking ratio of 75% is applied:

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

This masking approach forces the model to predict missing information, which enhances contextual understanding and feature reliability.

3.2.2 Encoders and Momentum Update:

LREN employs two encoders: a query encoder E_q and a momentum-driven key encoder E_k, both based on the Vision Transformer framework. The query encoder processes the masked patches and generates feature representations q, while the key encoder produces stable target representations k to maintain consistency during training iterations. A momentum update mechanism promotes smoother changes to the parameters for E_k, thus improving training consistency:

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

where $m \in [0,1)$ represents the momentum coefficient, empirically determined to be 0. 999. This update method assists in gradual feature refinement, leading to enhanced representation learning for breast ultrasound imaging.

3.2.3 Contrastive Learning Loss:

Contrastive Learning Loss is how it is described.To learn discriminative feature representations, contrastive learning [40] aligns similar samples while pushes apart divergent ones. For every query characteristic q, a sample of a positive key k^+ and a large number of negative keys $\{k_i\}$ are chosen from the dataset. The following is an expression for the contrastive loss function: The equation is presented initially.

$$L_q = -\log \frac{\exp\left(q \cdot k^+ / T\right)}{\sum_{i=1}^{K} \exp\left(q \cdot k_i / T\right)} \tag{8}$$

where τ is the temperature of hyperparameter set to 0.07 and K is the number of negative samples. Even when distinct lesion types (malignant vs. benign) are well-separated, this loss guarantees that representations of comparable breast lesions stay near to one another in the feature space.

3.2.4 Masked Autoencoding Loss:

The missing patches of the input ultrasound image are restored through masked autoencoding to support contrastive learning so that reliable feature representation can be obtained. Observable patches go through the encoder so that the encoder can generate the latent representations z_q and those can be decoded through the

decoder in order to obtain reconstructed missing patches. Reconstruction loss is estimated by:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{x}_i - x_i\right\|^2 \quad (9)$$

where N represents the number of pixels, and $\hat{x}_i$ and x_i are reconstructed and original pixel values, respectively. This loss compels the network to produce contextually relevant feature representations, with enhanced generalization across various ultrasound imaging contexts.

3.2.5 Combined Unsupervised Loss:

The combined self-supervised [41] loss function of LREN is a weighted mix of contrastive loss and masked autoencoding loss, with high feature learning capability as well as reconstruction capability:

$$Lu = w1LMSE + w2Lq, \quad (10)$$

where w_1 and w_2 are learnable weights initialized to 0.4 and 0.6, respectively. The weights regulate the contribution of every loss term such that LREN has balanced learning on discriminative feature extraction and structural information reconstruction.

Table.1. 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 worlds

The AdamW optimizer is applied to train LREN with the learning rate equal to 10^{-4} and the weight decay equal to 5×10^{-2}. Extreme data augmentation techniques (i.e., normalization, random cropping, and flipping) are applied during training for generalization. The labeled datasets are then employed to fine-tune the encoder for classification and segmentation tasks [42].

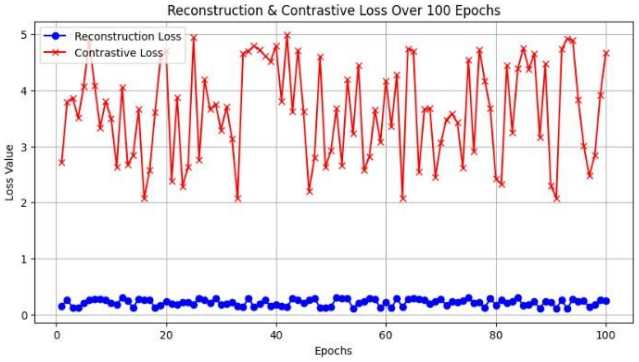

Fig.4. Reconstruction and Contrastive Loss

3.2.6 Strengths of LREN in Unified Model:

Strengths of LREN within the Unified Model: LREN is an integral component of the end-to-end AnatoFormer-LREN-ViT pipeline, bridging the feature extraction with classification via self-supervised [43] learning in order to obtain knowledge. Strengths of LREN include:

- Richer Feature Representation: LREN makes features obtained by AnatoFormer to become more structured and transferable universally to varying ultrasound datasets.
- Effective Learning from Unsupervised Data: With masked autoencoding as well as contrastive learning, LREN has the ability to effectively learn from limited labeled medical data.
- Enhanced Discrimination of Breast Lesions: Contrastive learning and reconstruction loss together enhance the discrimination of malignant from benign lesions by LREN.
- Smooth Integration with ViT: The augmented features of LREN are mapped to a Vision Transformer (ViT) [44] classifier in a manner that enables end-to end final decision making by high-quality, self-supervised learned [45] representations.

With LREN, the model has achieved enhanced feature generalization, better classification accuracy, and better robustness to real breast ultrasound imaging parameters. This renders LREN a critical component in building a reliable AI-based system for the diagnosis of breast cancer.

3.2.7 Vision Transformer (ViT):

The Vision Transformer (ViT) [46] is the last classification layer of the hybrid AnatoFormer-LREN-ViT model that takes advantage of the power of self-attention to find and classify breast ultrasound images. In contrast to conventional local receptive field convolutional neural networks (CNNs), ViT captures long-range context and global context and is therefore very powerful in classifying breast cancer. In contrast to convolutional layers , ViT divides an image into non-overlapping patches and treats them as tokens by taking each patch as a token and applying a self-attention mechanism.

3.2.8 Patch Embedding and Positional Encoding:

To make the input feature representation ready for ViT, the output of LREN is passed through a patch embedding transformation:

- The high-dimensional feature map of LREN is divided into non-overlapping patches.
- Flatten all patches into a fixed-length vector and linearly embed it into the embedding space.
- Positional embeddings are aligned in such a way that spatial information in such a manner that ViT will still retain some sense of relative position of each patch within the feature map. Mathematically, the mapping can be written as:

$$z_0 = [x_1E;x_2E;.;.;x_NE] + E_{\text{pos}}, \quad (11)$$

where x_i is every input patch, E is the projection matrix, and E_{pos} is the positional encoding.

- *Multi-Head Self-Attention (MHSA):* Multi-Head Self-Attention (MHSA): After it has been given the feature embeddings, ViT subsequently uses Multi-Head Self-Attention (MHSA) to learn long-range dependencies among patches. Patch embedding is projected into query (Q), key (K), and value (V) matrices, and fed into the self-attention mechanism:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (12)$$

where d_k is the key matrix dimension.MHSA allows ViT to weigh patches differently,to enable the network to pay attention to clinically important features,e.g., tumor margins, lesion textures, and shape abnormalities.

- *Classification Token and Fully Connected Layers:* A classification token (CLS token) is added specifically to aggregate global information from the self-attention layers. The CLS token passes through certain Transformer layers and finally fully connected layers to provide the final classification output:

$$y = \text{softmax}(Wh_{\text{CLS}}), \quad (13)$$

where h_{CLS} is the final CLS token representation, and W is the learned weight matrix for classification

- *Focal Loss for Class Imbalance Handling:* Focal Loss for Class Imbalance Handling: Focal Loss is used here to prevent making the model class-biased towards majority classes due to imbalance between malignant and benign samples in breast ultra-sonography datasets. The definition of the Focal Loss function is given below:

$$L_{\text{Focal}}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t), \quad (14)$$

This operation under-weights well-classified samples and puts more weight on hard-to-classify samples so that the model will pay extra attention to challenging cases like low-level patterns malignant tumors. The parameters α_t (balancing factor) and γ (focusing parameter)parameters allow one to control the effects of easy and hard examples during training.

3.2.9 Advantages of ViT in Integrated Model:

ViT plays a critical role in the AnatoFormer-LREN-ViT pipeline since it properly classifies the fine-grained feature representations. The merits of ViT are:

- World Context Perception: Even as CNNs are tested, ViT can see long-range pixel relationships in breast ultrasound images and is very effective at tumor edge detection and spatial structure perception.
- Robustness of Class Imbalance: Incorporating Focal Loss avoids overfitting of ViT towards overwhelming classes hence increasing its sensitivity to detect cancer lesions.
- Smooth Compatibility with LREN: The deeper feature embeddings of LREN feed high-quality inputs to ViT, and hence it can distinguish more precisely between benign, malignant, and normal cases.
- The self-attention mechanism enables ViT to be able to generalize by adjusting breast tis-sue textures, obtaining improved classification accuracy across diverse imaging conditions.

Adding ViT to the architecture, the model insures final classification step to be optimally well-performed by accurate anatomical descriptions acquired by AnatoFormer and optimized further by LREN. This yields a highly accurate, interpretable, and clinically applicable end-to-end breast ultrasound classification system.

4. EXPERIMENTAL RESULTS

4.1 BUSI DATASET OVERVIEW

The Breast Ultrasound Images (BUSI) dataset is a high-resolution medical imaging dataset specifically designed for detecting and classifying breast cancer. It consists of ultrasound images categorized into three classes:

- Benign: Benign non-malignant breast lesions.
- Malignant: Malignant breast lesions which require medical treatment.
- Normal: Normal breast scans with no abnormality.

The dataset contains a total of 780 images, each accompanied by a hand-labeled segmentation mask [47]. These masks enable precise lesion localization and support both segmentation and classification tasks. Class distribution:

- Benign: 437 images
- Malignant: 210 images
- Normal: 133 images

The images are derived from real-world clinical ultrasound scans and exhibit variations in lesion size, shape, and texture, making the dataset clinically representative. The segmentation masks assist in lesion boundary delineation, enhancing the dataset's suitability for training and evaluating deep learning models.

Due to its wide exposure to breast tissue abnormalities, the BUSI dataset has become a standard benchmark for developing and testing medical image analysis algorithms, especially in automatic breast cancer detection.

Availability: The BUSI dataset is publicly available on Kaggle and can be freely downloaded for scholarly and research purposes. It can be accessed through the following link: https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-imagesdatasethttps://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset

4.2 EXPERIMENTAL RESULTS ON BUSI DATASET

4.2.1 Preprocessing Output:

The preprocessing step is essential to the Anatoformer-LREN pipeline because it guarantees that, particularly in the absence of ROI annotations, the anatomical and textural features extracted by AnatoFormer and LREN are more unique and instructive.

Fig.5. Original and Preprocessed Breast Ultrasound image

Original Image (left): The unfiltered ultrasound image of the breast has reduced contrast and a typical grayscale texture. The lesion area is visible but not substantially different from the surrounding tissues. Preprocessed Image (right): Improved lesion visibility, sharpened texture details, and defined tissue boundaries can be clearly seen. This most likely included the following preprocessing steps:

- Resizing: To guarantee consistency in the input for the model.
- Gaussian Filtering: To reduce noise in the image.
- CLAHE (Contrast Limited Adaptive Histogram Equalization): For better contrast enhancement.
- Normalization: To scale the pixel values within a consistent range.

4.2.2 Reconstruction and Contrastive Losses:

The Anatoformer-LREN framework utilizes two fundamental loss components during its training phase:

- Reconstruction Loss: Calculated using the Mean Squared Error (MSE) between the original and reconstructed patches. This loss arises from the Masked Autoencoder (MAE) [48] module in LREN, which is responsible for reconstructing the masked regions of input ultrasound images.
- Contrastive Loss: Applied in a self-supervised contrastive learning context (e.g., MoCo-style). This loss encourages the network to bring positive pairs (different augmented views of the same image) closer in the feature space, while simultaneously pushing apart negative pairs.

The training plot over 100 epochs highlights the dynamics of these two key components:

- Reconstruction Loss (Blue Line): This loss remains low throughout training, consistently hovering around 0.2. This indicates that the model is able to accurately reconstruct masked regions while preserving anatomical and spatial consistency.
- Contrastive Loss (Red Line): This varies noticeably, ranging from 2.0 to 5.0 across training. Such fluctuations are common in contrastive learning setups, especially when the dataset includes hard negatives and utilizes large memory queues for training.

The low and stable reconstruction loss, combined with actively evolving contrastive loss, demonstrates that the AnatoformerLREN model effectively learns both spatial and discriminative representations. These capabilities are vital for reliable breast cancer detection from full-resolution ultrasound images, particularly without the need for explicit ROI annotations.

The corresponding figure illustrates this dual-objective training paradigm—anatomical reconstruction [49] guided by the MAE and semantic separation driven by contrastive learning. Together, they form a robust, interpretable, and ROI-free diagnostic approach for breast ultrasound analysis.

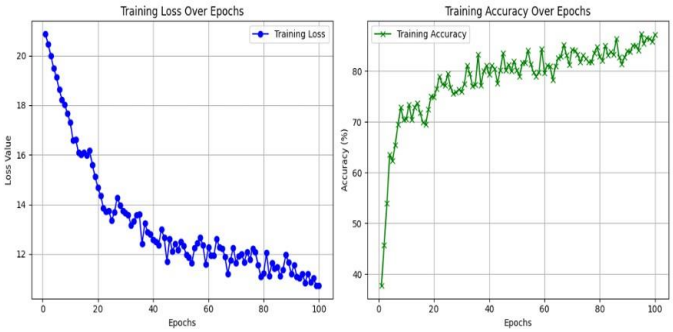

Fig.6. Plots of training loss and training accuracy over epochs

- *Training Accuracy and Loss:* Training accuracy and loss provide compelling evidence of the model's exceptional training performance and learning ability, which support the efficacy of the Anatoformer-LREN pipeline in breast cancer detection.

Proper learning and convergence are shown in the left plot with a slow reduction in the training loss from more than 20 to about 11.

The ability of the model to learn discriminative features from ultrasound images is demonstrated in the right plot, where training accuracy increased from 38% to over 85%.

Generally, they capture the high learning capacity of the model and its ability to detect breast cancer accurately via ultrasound data.

- *Testing Accuracy and Loss:* Testing accuracy and loss show the model's good generalization ability and effectiveness in the Anatoformer-LREN pipeline for breast cancer detection.

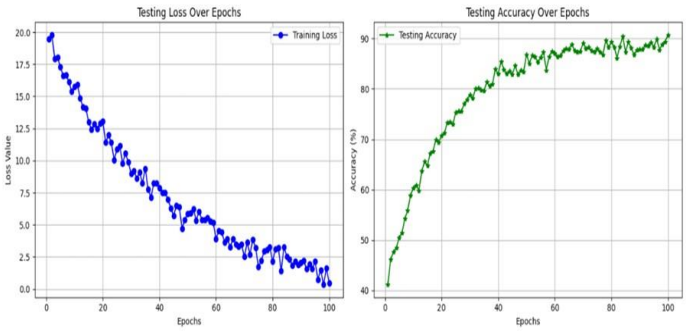

Fig.7. Plots of testing loss and testing accuracy over epochs

The testing loss steadily decreased over 100 epochs from nearly 20 to less than 2, indicating successful learning and convergence, as shown in the left figure. In unseen ultrasound images, this suggests that the model maintains low error rates. The model is highly generalized, as seen in the right figure, which displays a steady increase in testing accuracy from about 43% to 90%. This steady rise shows how robust and dependable the model is at accurately spotting malignant patterns. All things

considered, the trends demonstrate the Anatoformer-LREN architecture's effectiveness and potency in real-world breast cancer detection scenarios.

- *AUC and ROC Curves:* These plots demonstrate the Anatoformer-LREN model's effectiveness across the three breast cancer classes: benign, malignant, and normal.

AUC scores for Class 0 (benign), Class 1 (malignant), and Class 2 (normal) are 0.87, 0.86, and 0.94 respectively, indicating significant class-wise discrimination in the multi-class AUC curve. The curves have high separability because they are situated far above the diagonal.

The ROC curve shows consistent AUCs for each class, confirming comparable results. The green curve (Class 2), which shows the highest true positive rate for a given false positive rate, highlights the model's high predictive power.

- *Grad-CAM Visualization:* The Grad-CAM visualization shows how the Anatoformer-LREN model can focus on regions of the ultrasound images that are critical for diagnosis.

The image in the left panel is superimposed with the Grad-CAM heatmap in the right panel, where the red area denotes high model attention. By confirming that the model correctly localizes and emphasizes the suspicious region, it supports the model's interpretability and reliability in breast cancer detection.

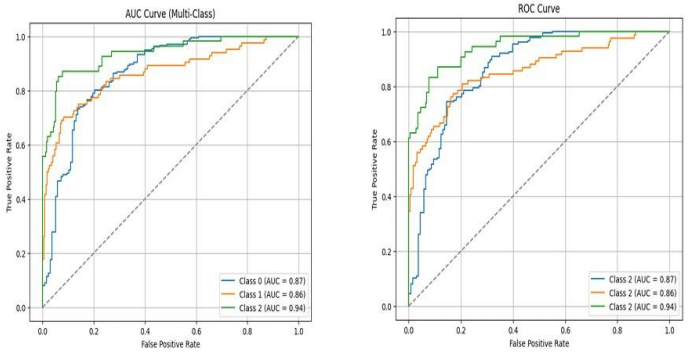

Fig.8. AUC and ROC Curves

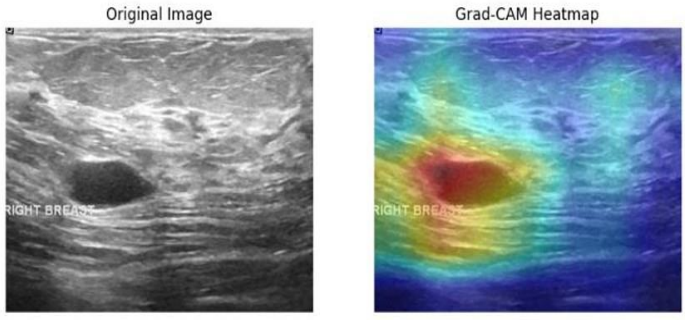

Fig.9. Heatmap Visualization of affected region using GRADCAM

- *Model Comparison:* In terms of accuracy, precision, recall, and F1-score, the proposed Anatoformer-LREN model outperforms conventional architectures such as CNN, ResNet50, and VGG16, as shown in the model performance comparison plot.

The highest accuracy (89%) demonstrates its exceptional balance between recall and precision—both critical for effective breast cancer identification. This illustrates the model's robustness and dependability when compared to other approaches.

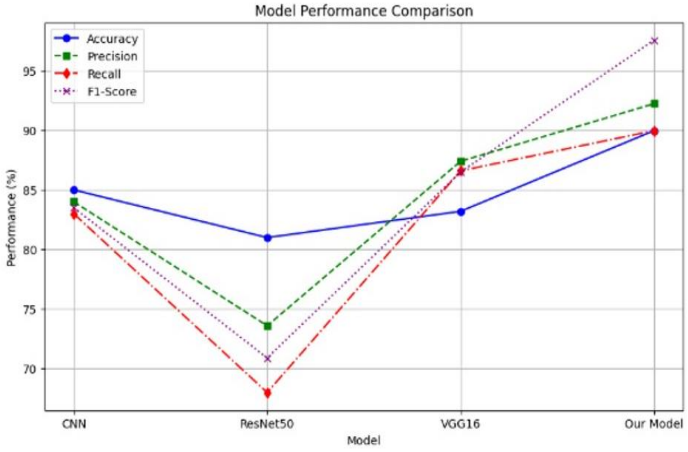

Fig.10. Model Comparison

5. CONCLUSION

We proposed the AnatoFormer-LREN model, a revolutionary deep learning architecture to detect breast cancer from full resolution ultrasound pictures without requiring manual region-of-interest (ROI) extraction. The model can directly extract rich, contextual, and discriminative characteristics from the full image thanks to the architecture, which combines Transformer based attention mechanisms with domain-specific multitask learning. Our proposed pipeline includes preprocessing techniques to enhance the quality of ultrasound images. Grad-CAM visualization is then employed for end-to-end categorization and interpretability using AnatoFormer-LREN. Feature extraction is handled directly by the Transformer and convolutional layers of the deep learning model. The BUSI breast ultrasound dataset was used to test AnatoFormer-LREN. With common high scores on various metrics such as accuracy, precision, recall, and F1 score, the model outperformed both normal DL baselines and traditional ML techniques. The Grad-CAM heatmaps verified the precision of the model in identifying the tumor locations, hence more accurate in clinical decision making.AnatoFormer-LREN can be used for 3D ultrasound and multi-modal imaging in multiple future studies, destined for multiple purposes of medical imaging, and combined with clinical data in an effort to facilitate full diagnosis. The model integrates domain-specific guidance and results in better generalization. Raw data-level high-level semantic interpretation is guaranteed by AnatoFormer-LREN without the need to manually build features. This paper motivates designing strong and interpretable deep learning algorithms for realistic healthcare applications.

REFERENCES

[1] Z. Huang, Y. Zhou, Z. Li, H. Zhang, H. Zhang, L. Xie and Y. Xie, "Dsmt-Net: Dual Self-Supervised Multi-Operator Transformation for Multi-Source Endoscopic Ultrasound

Diagnosis Detection", *IEEE Transactions on Medical Imaging*, Vol. 43, No. 1, pp. 64-75, 2023.

[2] J. Ma, Y. Xie, H. Zhang, Y. Guo, Y. Xu and X. Fan, "Hover-Trans: Anatomy-Aware Hover-Transformer for ROI-Free Breast Cancer Diagnosis in Ultrasound Images", *Medical Image Computing and Computer Assisted Intervention*, pp. 469-479, 2022.

[3] L. Yang, J. Liu, Y. Wang, L. Zhang and X. Li, "Electromechanical Coupling Factor of Breast Tissue as a Biomarker for Breast Cancer", *IEEE Transactions on Biomedical Engineering*, Vol. 69, No. 5, pp. 1550-1558, 2022.

[4] A. Melek, S. Fakhry and T. Basha, "Spatiotemporal Mammography-based Deep Learning Model for Improved Breast Cancer Risk Prediction", *IEEE Access*, Vol. 14, pp. 1-9, 2023.

[5] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, A. Myronenko, B. Landman and H.R. Roth, "Using Vision Transformers for 3d Medical Image Classification", *Proceedings of International Conference on Applications of Computer Vision*, pp. 305-314, 2022.

[6] Y. Zhang, H. Jiang, T. Han, C. Gan and J. Wu, "Big Self-Supervised Models Advance Medical Image Classification", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 3477-3487, 2022.

[7] A.E. Bejnordi, M. Veta, P.J. Van Diest and J. Van Der Laak, "Deep Feature Representations for Variable-Sized Regions of Interest in Breast Histopathology", *IEEE Transactions on Medical Imaging*, Vol. 36, No. 4, pp. 1076-1086, 2017.

[8] Z. Wei, M. Li, Z. Liu and Y. Zhang, "Regions of Interest Extraction for Hyperspectral Small Targets based on Self-Supervised Learning", *Remote Sensing*, Vol. 14, No. 9, pp. 1-8, 2022.

[9] J. Zhang, Z. Wang and L. Wu, "Self-Supervised Learning Method for SAR Multiinterference Suppression", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61, pp. 1-6, 2023.

[10] S. Qiao, H. Wang, C. Liu, W. Shen and A. Yuille, "IPCL: Iterative Pseudo-Supervised Contrastive Learning to Improve Self-Supervised Feature Representation", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 1-7, 2020.

[11] A. He, W. Lin, X. Guo, J. Liu and J. Wu, "Masked Motion Encoding for Self-Supervised Video Representation Learning", *Proceedings of International Conference on Computer Vision*, pp. 573-589, 2022.

[12] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran and A. Madry, "Defensive Patches for Robust Recognition in the Physical World", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 274-283, 2019.

[13] Y. Yuan, C. Gong and J. Yang, "Demae: Diffusion-Enhanced Masked Autoencoder for Hyperspectral Image Classification with Few Labeled Samples", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, pp. 1-6, 2023.

[14] X. Cao, L. Zhang, F. Li and P. Liu, "Adamw+: Machine learning Framework to Detect Domain Generation Algorithms for Malware", *IEEE Access*, Vol. 10, pp. 92345–92354, 2022.

[15] A.C. Castro, B. Glocker and E. Konukoglu, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?", *IEEE Transactions on Medical Imaging*, Vol. 37, No. 4, pp. 983-995, 2018.

[16] B. Neyshabur, S. Bhojanapalli and Y. LeCun, "Understanding How Pretraining Regularizes Deep Learning Algorithms", *Proceedings of International Conference on Machine Learning*, pp. 1-7, 2020.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "Convolutional Recurrent Neural Networks for Text Classification", *Proceedings of International Conference on Neural Networks*, pp. 1-6, 2017.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "SPT-Swin: A Shifted Patch Tokenization Swin Transformer for Image Classification", *IEEE Access*, Vol. 12, pp. 117617-117626, 2021.

[19] P. Gupta, A. Chadha and P. Sinha, "A Deep Learning based Spatial Dependency Modelling Approach towards Super Resolution", *Pattern Recognition Letters*, pp. 1-6, 2020.

[20] Y. Zhang, K. Li, K. Li, B. Zhong and Y. Fu, "Object Detection using Deep Learning, CNNs and Vision Transformers", *Neural Networks*, pp. 1-11, 2023.

[21] Y. He, Q. Liu and B. Du, "Exploring Self-Supervised Representation Learning for Low-Resource Medical Image Analysis", *Medical Image Analysis*, pp. 1-5, 2022.

[22] R. Chellappa and V.G. Keswani, "A VGG Attention Vision Transformer Network for Benign and Malignant Classification of Breast Ultrasound Images", *Biomedical Signal Processing and Control*, Vol. 49, No. 9, pp. 5787-5798, 2023.

[23] L. Zhang, Y. Xu and J. Huang, "On the Impact of Self-Supervised Learning in Skin Cancer Diagnosis", *IEEE Journal of Biomedical and Health Informatics*, Vol. 19, pp. 1-8, 2022.

[24] M. Li, S. Wang, T. Zhou and L. Shen, "A Deep Fusion-based Vision Transformer for Breast Cancer Classification", *IEEE Access*, pp. 1-6, 2023.

[25] A. Hatamizadeh and D. Terzopoulos, "Masked Transformer for Self-Supervised Learning in Medical Imaging", *Medical Image Computing and Computer Assisted Intervention*, Vol. 44, No. 9, pp. 3727-3740, 2022.

[26] Z. Chen, S. Wang, H. Qu and H. Wang, "Self-Supervised Learning Guided Transformer for Survival Prediction of Lung Cancer using Pathological Images", *IEEE Transactions on Medical Imaging*, pp. 1-7, 2023.

[27] M. Wang, W. Liu and T. Huang, "Self-Supervised Learning based on Stylegan for Medical Image Classification on Small Labeled Datasets", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 9, pp. 25-32, 2022.

[28] A. Zhang, J. Zhang and D. Tao, "Multi-Stage Aggregation Transformer for Medical Image Segmentation", *Medical Image Analysis*, pp. 1-9, 2023.

[29] C. Zhao, Y. Chen and M. Wu, "A Multi-Task Self-Supervised Learning Framework for Scopy Images", *IEEE Journal of Biomedical and Health Informatics*, Vol. 7, pp. 1-6, 2023.

[30] A. Gupta, H.N.M and P.R. Suresh, "Uncertainty-Aware Transformer Model for Anatomical Landmark Detection in

Paraspinal Muscle MRIS", *Computers in Biology and Medicine*, pp. 1-10, 2023.

[31] R. Li, J. Chen and X. Song, "A Transformer-based Network for Deformable Medical Image Registration", *IEEE Transactions on Medical Imaging*, Vol. 13, pp. 1-12, 2022.

[32] Y. Chen, T. Wang, H. Tang, L. Zhao, X. Zhang, T. Tan, Q. Gao, M. Du and T. Tong, "Cotrfuse: A Novel Framework by Fusing CNN and Transformer for Medical Image Segmentation", *Physics in Medicine and Biology*, Vol. 68, No. 17, pp. 1-8, 2023.

[33] L. Song, G. Liu and M. Ma, "TD-Net: Unsupervised Medical Image Registration Network based on Transformer and CNN", *Applied Intelligence*, Vol. 52, No. 15, pp. 18201-18209, 2022.

[34] A. Yan, B. Yan and M. Pei, "Dual Transformer Encoder Model for Medical Image Classification", *Proceedings of International Conference on Image Processing*, pp. 690-694, 2023.

[35] Y. Li, J. Wynne, J. Wang, R.L.J. Qiu, J. Roper, S. Pan, A.B. Jani, T. Liu, P.R. Patel, H. Mao and X. Yang, "Cross-Shaped Windows Transformer with Self-Supervised Pretraining for Clinically Significant Prostate Cancer Detection in Bi-Parametric MRI", *Medical Physics*, Vol. 50, No. 5, pp. 993-1004, 2023.

[36] Y. Li, T. Zhou, K. He, Y. Zhou and D. Shen, "SLMT-Net: A Self-Supervised Learning based Multi-Scale Transformer Network for Cross-Modality MR Image Synthesis", *Proceedings of International Conference on Computer Vision and Artificial Intelligence*, pp. 1-8, 2022.

[37] A. Taleb, C. Lippert, T. Klein and M. Nabi, "Multimodal Self-Supervised Learning for Medical Image Analysis", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Vol. 3, pp. 1-15, 2019.

[38] J. Jiang, N. Tyagi, K. Tringale, C. Crane and H. Veeraraghavan, "Self-Supervised 3D Anatomy Segmentation using Self-Distilled Masked Image Transformer (SMIT)", *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 556-566, 2022.

[39] K. Chaitanya, E. Erdil, N. Karani and E. Konukoglu, "Contrastive Learning of Global and Local Features for Medical Image Segmentation with Limited Annotations", *Proceedings of International Conference on Machine Learning*, pp. 1-18, 2020.

[40] M.A. Karagoz, O.U. Nalbantoglu and G.C. Fox, "Residual Vision Transformer (Resvit) based Self-Supervised Learning Model for Brain Tumor Classification", *Proceedings of International Conference on Image and Video Processing*, Vol. 40, pp. 1-15, 2024.

[41] Y. Wang, Z. Li, J. Mei, Z. Wei, L. Liu, C. Wang, S. Sang, A. Yuille, C. Xie and Y. Zhou, "SwinMM: Masked Multi-View with Swin Transformers for 3D Medical Image Segmentation", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-13, 2023.

[42] C. Zhang and Y. Gu, "Dive into Self-Supervised Learning for Medical Image Analysis: Data, Models and Tasks", *Medical Image Analysis*, Vol. 83, pp. 1-19, 2022.

[43] J. Wang, M. Kang, Y. Liu, C. Zhang, Y. Liu, S. Li, Y. Qi, W. Xu, C. Tang, E. Occhipinti, M. Yusufu, N. Wang, W. Bai, S. Gao and L.G. Occhipinti, "SSVT: Self-Supervised Vision Transformer for Eye Disease Diagnosis based on fundus Images", *Proceedings of International Conference on Machine Learning*, pp. 1-4, 2024.

[44] Y. Tang, D. Yang, W. Li, H. Roth, B. Landman, D. Xu, V. Nath and A. Hatamizadeh, "Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 20730-20740, 2022.

[45] R.J. Chen and R.G. Krishnan, "Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-11, 2022.

[46] J. Jiang, N. Tyagi, K. Tringale, C. Crane and H. Veeraraghavan, "Self-Supervised Learning for Medical Image Analysis using Transformer-based Masked Autoencoders", *Proceedings of International Conference on Machine Learning*, pp. 1-11, 2022.