Image Anal Stereol 2026;45:1-10 doi: 10.5566/ias.3695
Original Research Paper

DEEP LEARNING DRIVEN BREAST CANCER MALIGNANCY PREDICTION IN ULTRASOUND LEVERAGING MULTISCALE FEATURE FUSION AND SELF SUPERVISED LEARNING

MEICHEN WANG, WEN LI, HUIHUI ZHU AND ZHAOXI LI✉

Department of Physical Diagnosis, Shanghai Health Medical Center, (formerly Huadong Sanatorium) China
e-mail: wangmeichen1212@outlook.com, li158048150@163.com, moonicar@sina.com,
lizhaoxi3596@outlook.com

(Received July 10, 2025; revised November 20, 2025; accepted November 27, 2025)

ABSTRACT

Accurate malignancy prediction in breast ultrasound imaging is challenged by limited annotated data, high inter-observer variability, and inherent noise in sonographic textures. To address these limitations, we propose a deep learning framework that synergistically integrates multiscale feature fusion and self-supervised learning (SSL) to improve diagnostic performance while minimizing reliance on labeled datasets. The architecture employs a hierarchical convolutional backbone with multiscale feature extractors that capture both coarse contextual semantics and fine-grained morphological cues of lesions. Features across multiple receptive fields are fused via a top-down multiscale fusion strategy using bilinear upsampling and channel concatenation, enhancing the model's ability to localize and characterize malignant regions. We applied a self-supervised contrastive learning approach tailored for medical ultrasound, incorporating spatial transformation invariance and anatomical context preservation to learn domain-relevant representations from unlabeled data. The pretrained encoder is fine-tuned with a supervised classification head using a limited set of annotated images. Extensive experiments on two publicly available breast ultrasound datasets demonstrate that our model achieves higher performance over state-of-the-art baselines, yielding significant improvements in AUC, F1-score, and sensitivity. Ablation studies confirm the individual and combined efficacy of the multiscale fusion and SSL modules. This work establishes a scalable and label-efficient pipeline for ultrasound-based malignancy prediction, with implications for real-time clinical decision support.

Keywords: Breast cancer, Convolutional neural networks, Self supervised learning, Ultrasound..

INTRODUCTION

Breast cancer continues to represent a major global health burden, accounting for a significant proportion of cancer diagnoses and mortality among women (Arnold *et al.*, 2022). According to the World Health Organization, breast cancer has surpassed lung cancer as the most commonly diagnosed cancer worldwide, with millions of new cases reported annually[1]. Early detection and timely intervention are essential for improving survival rates, yet many women, particularly in low-resource settings, still face barriers to accurate and accessible diagnostic services (Anderson *et al.*, 2003). Medical imaging plays a central role in the early detection pipeline, guiding clinical decisions regarding biopsy, treatment planning, and follow-up care. Among the available modalities, such as mammography, magnetic resonance imaging (MRI), and computed tomography (CT), ultrasound imaging stands out due to its safety, affordability, portability, and lack of ionizing radiation (Anandhamala *et al.*, 2018). These attributes make ultrasound especially valuable for breast cancer screening in younger women and in regions where access to advanced imaging technologies is limited.

However, despite its advantages, breast ultrasound imaging presents several inherent challenges (Madjar, 2018). The quality and diagnostic value of ultrasound images are highly dependent on the skill and experience of the operator, leading to variability in acquisition and interpretation. Moreover, ultrasound images typically exhibit low contrast, speckle noise, and subtle texture variations, which can obscure lesion boundaries and make it difficult to distinguish between benign and malignant masses. Manual interpretation is not only time-consuming and subjective, but also prone to intra- and inter-observer inconsistencies, contributing to both false positives and false negatives. These limitations underscore the need for intelligent, automated systems that can assist radiologists by offering consistent, accurate, and reproducible malignancy assessments.

In recent years, deep learning has emerged as

[1] www.who.int/news-room/fact-sheets/detail/breast-cancer

a transformative approach to medical image analysis, demonstrating impressive results across various diagnostic tasks, including disease classification, organ segmentation, and anomaly detection (Suzuki, 2017). Convolutional neural networks (CNNs), in particular, have proven effective in learning hierarchical feature representations directly from raw image data, often surpassing traditional handcrafted features in both accuracy and robustness. In the context of breast ultrasound, deep learning models have shown promise in improving lesion classification by capturing complex patterns related to shape, margin, and echotexture. However, a major obstacle in the implementation of these models in clinical practice is their data hunger and most supervised deep learning models require large volumes of annotated data to achieve generalizable performance. Given that medical annotations often require domain expertise and are expensive to obtain, this constraint limits the scalability and adaptability of deep learning solutions in real-world healthcare environments.

To overcome this bottleneck, self-supervised learning (SSL) has gained traction as a viable alternative to traditional supervised training. SSL strategies enable models to learn useful and transferable representations of unlabeled data by solving pretext tasks, such as predicting image rotations, solving jigsaw puzzles, or contrasting image embeddings (Shurrab and Duwairi, 2022). These tasks encourage the network to extract meaningful structures from the data without requiring explicit labels. In medical imaging, SSL is especially appealing due to the abundance of unlabeled clinical scans stored in hospital databases. By leveraging SSL, models can be pretrained on large-scale unlabeled ultrasound datasets to acquire a strong initialization, which can then be fine-tuned on small labeled datasets for downstream classification. This label-efficient learning paradigm not only alleviates the burden of manual annotation but also improves the model's robustness and generalization to unseen data distributions.

Complementing the benefits of SSL is the concept of multiscale feature fusion, a technique that allows models to integrate information across multiple spatial resolutions. In ultrasound imaging, lesions can exhibit both macrolevel contextual features, such as surrounding tissue heterogeneity, and microlevel attributes, such as edge sharpness and internal echogenicity. By designing architectures that capture and fuse features at different scales, models can learn more effectively the nuanced visual cues necessary for accurate malignancy prediction. Multiscale fusion strategies have been widely adopted in computer vision tasks such as object detection and semantic segmentation, but their application in medical ultrasound, especially in conjunction with SSL, remains underexplored.

In this study, we propose a novel deep learning framework that unifies multiscale feature fusion with self-supervised learning to address the limitations of current approaches to the prediction of breast cancer malignancy in ultrasound imaging (see Fig. 1). Our model architecture includes a multiscale backbone that extracts hierarchical features across varying receptive fields and integrates them through a multiscale feature fusion via top-down upsampling and channel concatenation. To enable robust pretraining, we introduce a self-supervised contrastive learning strategy tailored to ultrasound data, incorporating transformation-based invariance and anatomical consistency. We evaluated our framework on publicly available breast ultrasound datasets, demonstrating its higher performance compared to state-of-the-art supervised and SSL baselines. Through extensive quantitative analysis and ablation studies, we validate the individual and combined contributions of the proposed multiscale and SSL components. This work aims to bridge the gap between data-efficient learning and high-performance diagnosis, contributing toward more accessible and reliable deep learning assisted breast cancer screening solutions.

MATERIAL AND METHOD

In this section, we describe our proposed deep learning framework for predicting breast cancer malignancy in ultrasound images, which integrates a multiscale feature fusion strategy with a self-supervised learning paradigm. The framework is designed to address the dual challenges of limited labeled data and the complex and multiscale nature of sonographic breast lesions. It consists of three primary components: (1) a preprocessing pipeline that standardizes ultrasound inputs and increases variability; (2) a self-supervised contrastive learning phase that pretrains a multiscale encoder using large volumes of unlabeled data; and (3) a supervised fine-tuning phase in which the pre-trained encoder is adapted to the malignancy classification task using a relatively small labeled dataset.

DATASET

To develop the SSL (semi-supervised learning) approach, we used five breast ultrasound data sets: BUS-BRA (1,875 images) (Gómez-Flores *et al.*,

[2] https://www.kaggle.com/datasets/orvile/bus-uc-breast-ultrasound

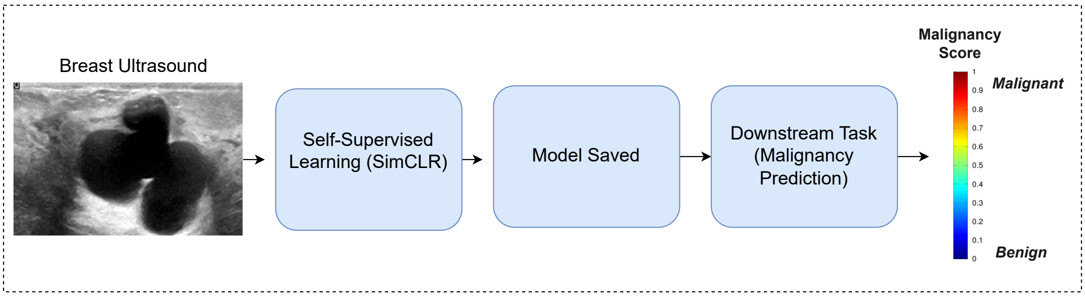

Fig. 1. *Illustration of the proposed method pipeline.*

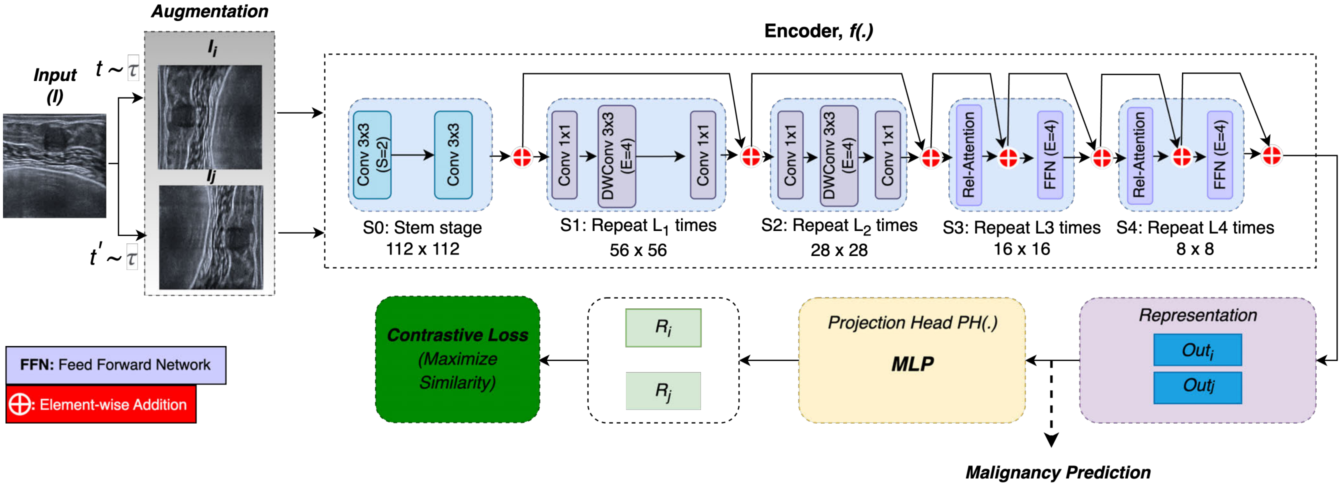

Fig. 2. *Illustration of the proposed self-supervised learning framework for predicting breast tumor malignancy. The BUS image undergoes preprocessing and augmentation before being input into the deep learning model, which generates a malignancy score reflecting the likelihood of cancer.*

2024), BUS-UCLM (683 images) (Vallez *et al.*, 2025), BUS-UC (811 images)[2], QAMEBI (232 images) (Ardakani *et al.*, 2023) and US3M (1,532 images) (Yan *et al.*, 2024). These datasets served as the foundation for pretraining our model, enabling it to learn rich, generalizable feature representations from a diverse collection of unlabeled and partially labeled breast ultrasound images.

For the downstream task of classification of breast cancer malignancy, we employed two publicly available data sets that reflect the diversity of the real world in imaging and pathological characteristics. These datasets were selected to evaluate the effectiveness of our proposed deep learning framework in a clinically meaningful binary classification setting (benign vs. malignant), while also testing its robustness across different geographic and demographic populations.

The first evaluation data set, known as the UDIAT BUS dataset (Yap *et al.*, 2017), originates from the UDIAT Diagnostic Center at Parc Taul Health Corporation in Sabadell, Spain. It contains 163 grayscale ultrasound images depicting breast lesions, with histopathologically confirmed annotations comprising 109 benign and 54 malignant cases. Each image has an average resolution of approximately 760 × 570 pixels and follows standardized ultrasound imaging protocols.

The second dataset comes from Baheya Hospital (Al-Dhabyani *et al.*, 2020) in Egypt, a renowned center for breast cancer screening and treatment. It includes 780 grayscale ultrasound images originally labeled normal (133), benign (487), and malignant (210). For the purposes of our study, only the benign and malignant cases were considered, resulting in 697 images used for binary classification. All images have a consistent resolution of 500 × 500 pixels.

Using these two evaluation datasets, one from a European clinical setting and the other from a Middle Eastern healthcare context, we were able to assess the performance, generalizability and clinical relevance of our malignancy prediction framework across heterogeneous and geographically diverse ultrasound imaging data.

MULTISCALE FEATURE FUSION ARCHITECTURE

To effectively capture multiscale features of breast lesions, including fine details such as margin irregularities, as well as a broader tissue context, we developed a multiscale hybrid backbone based on the CoAtNet (Dai *et al.*, 2021) architecture. The input ultrasound image is processed through a series of convolutional stages that progressively extract hierarchical feature maps. Each stage captures increasingly abstract and spatially compressed representations of the input, with feature maps downsampled by factors of 2, 4, and 8 relative to the original image size.

To enhance the ability of the model to capture wider contextual information without further reducing spatial resolution, we incorporate dilated convolutions into the final two stages. This allows the deeper layers to expand their receptive field while preserving important spatial details.

To integrate information across multiple scales, we apply a top-down fusion strategy. Feature maps from the deeper layers are upsampled using bilinear interpolation to match the spatial resolution of the shallowest feature map. These upsampled maps are then concatenated along the channel dimension, resulting in a unified representation that combines both detailed local features and high-level contextual information.

SELF-SUPERVISED CONTRASTIVE LEARNING

Given the scarcity of labeled medical data, we adopt a self-supervised learning approach (Abdel-Nasser *et al.*, 2022) to pretrain the encoder using unlabeled ultrasound images. We formulate the task as instance discrimination using a contrastive learning objective. For each image x_i in a batch of size N, we apply two random augmentations to generate two correlated views $x_i^{(1)}$ and $x_i^{(2)}$. The augmented images are passed through a shared encoder $f(\cdot)$, followed by a projection head $g(\cdot)$, to obtain 128-dimensional embeddings.

As illustrated in Fig. 2, the network is designed to learn robust feature representations from pairs of enhanced ultrasound (US) images. These image pairs are created by applying a set of data augmentation techniques to each original image, resulting in two distinct but semantically related views, denoted as I_i and I_j. This approach allows the network to learn invariances and correspondences between different visual appearances of the same underlying anatomical structures. To generate these image pairs, we utilize a combination of augmentation strategies that simulate common variations in ultrasound imaging. Specifically, we apply horizontal and vertical flipping, 30-degree rotations, Gaussian blurring, and random changes in brightness and contrast, with the latter applied probabilistically (with a 0.1 chance). In addition, we introduce color jitter to simulate subtle variations in intensity and tone that can occur across scans. These augmentations enrich the dataset and encourage the network to focus on the intrinsic features of the anatomical content rather than superficial differences. The learning process is guided by a contrastive loss function, which is optimized to bring the feature representations of the augmented views (I_i and I_j) closer together in the embedding space. By maximizing the similarity between these different representations of the same original image, the model learns to extract stable and discriminative visual features. This process is crucial for enhancing the model's generalization ability across varying ultrasound appearances and improving downstream performance on related tasks.

Furthermore, we use a multiscale input technique to effectively capture characteristics across varying degrees of spatial resolution, which is especially crucial in ultrasonic imaging, where structures of interest might range significantly in size, shape, and appearance. Instead of depending on a single fixed input resolution, we generate numerous iterations of each ultrasound picture at three distinct spatial scales: 224×224, 112×112, and 28×28 pixels. Each scaled images is analyzed individually to extract features at its specific resolution. This methodology allows the model to focus on both broad contextual patterns (derived from lower-resolution images) and detailed texture or boundary information (obtained from higher-resolution images). Following feature extraction at each scale, the resultant feature maps are concatenated along the channel dimension to provide a full multiscale representation, then input into the network for further processing. This method enhances the model's ability to generalize across various anatomical differences and noise conditions often seen in clinical ultrasound data.

The upper branch seen in Fig. 2 illustrates the primary feature extractor, designated as f, which is founded on CoAtNet (Dai *et al.*, 2021) which is a hybrid architecture that incorporates the advantages of CNNs and transformers. CoAtNet is explicitly designed to integrate the local inductive bias and efficiency of convolutional layers with the global modeling capabilities of self-attention mechanisms in transformer architectures. Our method has five successive stages: S0, S1, S2, S3, and S4. The first three stages (S0–S2) use convolutional blocks, which are adept at collecting local texture and spatial information often seen in ultrasound images. The latter two stages (S3 and S4) use transformer blocks to facilitate long-range dependency modeling, hence enhancing the larger anatomical context and improving semantic consistency in feature maps. Moreover, Stage S0 includes a down-sampling process with a stride of 2, which reduces the spatial dimensions of the input image, enhances computing efficiency, and retains critical structural characteristics. This hierarchical architecture enables the encoder to incrementally abstract characteristics from low-level textures to high-level semantics, making it highly appropriate for intricate medical imaging applications.

The second stage of the encoder network, designated as S1, integrates a Mobile Inverted Bottleneck Convolution (MBConv) block, sometimes referred to as an inverted residual block, as represented by CoAtNet (Dai *et al.*, 2021). MBConv, in contrast to conventional convolutional layers, utilizes depthwise separable convolutions instead of regular convolutions, therefore significantly lowering computational complexity while preserving representational capacity. This makes it particularly appropriate for medical imaging applications that need great efficiency without compromising performance.

The design of Stage S2 replicates that of S1, using an identical MBConv block for feature extraction. The successive stages, S3 and S4, shift from convolutional to transformer-based processes. Each stage comprises a transformer block that has a 2D relative attention mechanism, a feed-forward network (FFN), and a self-attention module. These components allow the network to simulate long-range relationships within the ultrasound image, essential for recognizing larger anatomical context. To avoid overfitting and ensure efficacy, both S3 and S4 include max-pooling processes with a stride of 2 inside the self-attention submodules, therefore diminishing spatial dimensions while preserving critical semantic information. As a result, the final result feature map at the bottleneck of Stage S4 has a spatial dimension of 8×8, as seen at the intersection of the upper and lower branches in Fig. 2.

$$Out_i = f\left(\tilde{I}_i\right) \tag{1}$$

$$Out_j = f\left(\tilde{I}_j\right) \tag{2}$$

Here, Out_i and $Out_j \in R^d$ represent the output feature embeddings of the augmented input image pairs $\tilde{I}_i$ and $\tilde{I}_j$, respectively. These embeddings are high-dimensional representations capturing the visual characteristics learned from the network.

The projection head $PH(\cdot)$ is implemented as a two-layer MLP that maps the encoder output into a 128-dimensional contrastive embedding space. The first linear layer expands the feature dimensionality to 2048 units, followed by Batch Normalization and a ReLU activation function. The second linear layer reduces the dimensionality to 128. Given encoder outputs Out_i and Out_j, the projection is defined as:

$$R_i = W^{(2)}\sigma\left(\mathrm{BN}\left(W^{(1)}Out_i\right)\right), \tag{3}$$

$$R_j = W^{(2)}\sigma\left(\mathrm{BN}\left(W^{(1)}Out_j\right)\right), \tag{4}$$

where $W^{(1)}$ and $W^{(2)}$ denote the weights of the MLP layers, σ is the ReLU activation, and BN$(\cdot)$ represents Batch Normalization.

The contrastive loss can be formulated as

$$\mathscr{L}_{FINAL} = \frac{1}{2N}\sum_{k=1}^{N}\left[\ell(2k-1,2k)+\ell(2k,2k-1)\right] \tag{5}$$

Here, N represents the mini-batch of ultrasound images, while the contrastive prediction yields $2N$ data points derived from pairs of data-augmented ultrasound samples. The ℓ may be calculated as

$$\ell_{i,j} = -\log\frac{\exp\left(CM\left(R_i,R_j\right)/\tau\right)}{\sum_{k=1}^{2N}1_{[k\neq i]}\exp\left(CM\left(R_i,R_k\right)/\tau\right)} \tag{6}$$

where τ is a temperature parameter fixed at 0.07 in our trials; $1_{[k\neq i]} \in \{0,1\}$ represents an indicator function to ascertain if $k \neq i$; CM is the cosine similarity function, which can be represented as

$$CM(R_i,R_j) = R_i^T R_j/\|R_i\|\|R_j\| \tag{7}$$

Contrastive loss reduces when projections from the same image demonstrate similarity; whereas, the error rate rises.

PERFORMANCE MEASUREMENT

To comprehensively evaluate the performance of the proposed deep learning approach for classifying breast cancer in ultrasound imagery, we employ four commonly used evaluation metrics: accuracy, precision, recall, and F1-score. These metrics provide a balanced assessment of the model's ability to correctly identify both benign and malignant lesions, especially in the presence of class imbalance.

The metrics are mathematically defined as follows:

Accuracy reflects the proportion of total correct predictions (both benign and malignant) over all predictions:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

Precision quantifies the proportion of correctly predicted malignant cases among all cases predicted as malignant:

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

Recall (also known as sensitivity) measures the ability of the model to correctly identify actual malignant cases:

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

F1-score provides the harmonic mean of precision and recall, offering a single performance measure that balances both:

$$F1-score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP+FP+FN} \tag{11}$$

Here, the terms are defined as:

TP: True Positives — malignant cases correctly classified as malignant.

TN: True Negatives — benign cases correctly classified as benign.

FP: False Positives — benign cases incorrectly classified as malignant.

FN: False Negatives — malignant cases incorrectly classified as benign.

These metrics allow us to assess not just the overall correctness of the model (via accuracy), but also its clinical reliability in correctly identifying important malignant cases (via recall) and minimizing false alarms (via precision). The F1-score provides a single aggregated measure to summarize this trade-off, which is particularly valuable in medical diagnosis scenarios.

EXPERIMENTAL RESULTS AND DISCUSSION

IMPLEMENTATION DETAILS

We trained the SSL network for 1000 epochs with a batch size of 128. The training was performed on unlabeled breast ultrasound datasets. This setup enabled the model to learn robust feature representations without relying on manual annotations. For the downstream task, prior to model training, all breast ultrasound images were uniformly resized to 224 × 224 pixels using aspect ratio-preserving padding to maintain spatial consistency. The intensity of the images was standardized by normalizing the score by subtracting the specific mean of the data set and dividing by the standard deviation, ensuring consistent intensity distributions between samples. To enhance the robustness and generalization of the model, a comprehensive data augmentation pipeline was applied. During the supervised fine-tuning phase, augmentations included random horizontal and vertical flips, rotations of up to 30 degrees, scaling transformations with probability 50%, elastic deformations and brightness or contrast adjustments. For the self-supervised learning phase, a distinct set of stochastic augmentations was used to generate positive view pairs necessary for contrastive learning, allowing the model to learn meaningful feature representations without reliance on labeled data. Model optimization used the ADAM optimizer with an initial learning rate of 0.0001, training over 200 epochs with mini-batches of four samples to balance computational efficiency and gradient stability. Consistent hyperparameter settings were maintained in all models to ensure a fair comparison. All baseline models used the same train/validation/test splits, identical preprocessing, identical hyperparameters (optimizer, batch size, learning rate), and were pretrained on the same unlabeled ultrasound images for the same number of epochs. The data sets were independently partitioned into training subsets (70%), validation (10%), and testing (20%) to address the characteristics specific to the data sets and ensure an impartial evaluation. Note that all dataset splits were performed strictly at the patient or case level to prevent any form of data leakage across training, validation, and test sets.

STATE-OF-THE-ART RESULTS COMPARISON

The quantitative results presented in Table 1 provide a comprehensive comparison between recent convolutional neural networks (CNN), advanced SSL frameworks, and the proposed SSL-driven model for the classification of breast cancer malignancies using ultrasound images. In the UDIAT dataset, the proposed model achieves a highest accuracy of 96.45%, significantly surpassing both conventional CNN architectures and state-of-the-art SSL models. This superior performance is supported by a precision of 95. 20% and a recall of 96. 80%, indicating that the model effectively balances false positives and false negatives, crucial for clinical reliability. The high F1-score of 96.00% confirms this equilibrium, reflecting

Table 1. *Performance comparison of recent CNN and self-supervised learning models with the proposed SSL-based method.*

Model	**UDIAT dataset**					**Baheya dataset**				
	Accuracy	Precision	Recall	F1-Score	AUC	Accuracy	Precision	Recall	F1-Score	AUC
ConvNeXt-Tiny	88.75	89.10	87.95	88.52	90.20	89.40	88.85	86.90	87.86	90.85
EfficientNetV2-S	86.30	85.65	86.50	86.07	88.30	87.50	86.95	85.80	86.37	89.00
Swin Transformer (SSL pre-trained)	91.40	90.85	91.10	90.97	92.75	91.20	90.55	90.95	90.75	92.60
DINO-ViT (Self-Supervised)	92.85	92.10	93.05	92.57	93.90	90.85	90.10	91.30	90.70	92.90
SimCLR (SSL CNN backbone)	89.75	89.20	88.85	89.02	90.80	88.90	88.40	87.90	88.15	90.20
Proposed Model	**96.45**	**95.20**	**96.80**	**96.00**	**98.10**	**95.80**	**94.30**	**95.10**	**94.70**	**97.25**

consistent performance across both malignant and benign class predictions. In addition, an AUC of 98. 10% demonstrates an exceptional discriminative ability to differentiate malignancies, highlighting the model's strong robustness in handling ultrasound image variability. Compared to self-supervised models such as DINO-ViT and Swin Transformer, which leverage transformer-based architectures pre-trained on large unlabeled datasets, show markedly improved feature representation learning. For example, DINO-ViT achieved an accuracy of 92.85% and an AUC of 93.90%, outperforming traditional CNNs like ConvNeXt-Tiny and EfficientNetV2-S by a notable margin. This improvement is attributable to the SSL methods' capability to learn semantically rich and invariant features, facilitating better generalization when fine-tuned on relatively small medical datasets.

On the Baheya dataset, which contains more diverse ultrasound samples, the proposed model maintains consistent excellence with an accuracy of 95.80%, along with precision and recall values of 94.30% and 95.10%, respectively. The F1-score of 94.70% and AUC of 97.25% indicate that the model effectively generalizes to different data distributions and imaging conditions, further confirming its robustness. SSL-based transformers continue to outperform CNN-based SSL methods like SimCLR, reflecting the advantage of hierarchical and attention-driven feature extraction in complex ultrasound images. The conventional CNNs, although optimized for image classification, show relatively lower performance due to their limited ability to capture long-range dependencies and complex texture patterns prevalent in ultrasound imaging. In contrast, the transformer-based SSL models exploit self-attention mechanisms to model global context, enhancing subtle malignancy feature extraction. In conclusion, the proposed SSL-driven model's superior performance is attributed to its effective self-supervised pretraining strategy that leverages unlabeled data to learn robust, high-dimensional feature embeddings. This leads to improved malignancy prediction accuracy and generalization across datasets, outperforming both classical CNN baselines and recent SSL frameworks.

Table 2. *Comparison of the proposed method with existing approaches on the UDIAT dataset. Dash marks (-) indicate metrics not reported in the original references.*

Methods	Evaluation Metrics				
	Accuracy	Precision	Recall	F1-Score	AUC
Proposed Model	**96.45**	**95.20**	**96.80**	**96.00**	**98.10**
(Byra *et al.*, 2019)	84.00	–	85.10	–	89.30
(Byra and Andre, 2019)	76.00	–	78.00	–	81.00
(Ning *et al.*, 2020)	90.90	–	92.70	–	93.90

Table 2 presents a comparative analysis of the proposed model's performance against several established methods on the UDIAT breast ultrasound dataset. The evaluation metrics include Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC), which together provide a comprehensive assessment of classification quality, balancing sensitivity, specificity, and overall discriminative ability.

The proposed method demonstrates a substantial improvement across all reported metrics, achieving an accuracy of 96.45%, which indicates that the model correctly classifies the majority of both malignant and benign cases, outperforming previous works by a margin of approximately 5% to 20%. This increase is significant considering the challenging nature of ultrasound images, which often suffer from low contrast and speckle noise. Precision and recall values of 95.20% and 96.80%, respectively, highlight the model's effectiveness in minimizing false positives and false negatives. High precision suggests that the model is highly reliable when predicting malignancy, reducing unnecessary biopsies or interventions. Simultaneously, high recall ensures that most malignant tumors are detected, which is critical in clinical scenarios to avoid missed diagnoses. The balanced F1-score of 96.00% further confirms the model's robustness by harmonizing these two aspects. The AUC of 98.10% indicates excellent discrimination capability, meaning the model consistently distinguishes malignant tumors from benign ones across varying classification thresholds. This is particularly important for medical imaging tasks, where operating points can be adjusted according to clinical priorities.

In contrast, previous studies such as (Byra *et al.*, 2019) and (Ning *et al.*, 2020) reported lower performance metrics, with accuracies ranging from 76% to 91%, and generally lacked reporting of precision and F1-scores, limiting direct comparison of false positive and false negative rates. The comparatively lower AUC values in these works indicate less reliable separability between classes. The higher performance of the proposed model can be attributed to several technical advances: the use of self-supervised learning pretraining allowed the extraction of richer, more generalized feature representations from limited labeled data; the integration of multiscale feature fusion techniques enhanced the capture of both global and local tumor characteristics; and the rigorous data augmentation and normalization strategies improved model robustness against image variability and noise.

Table 3. *Comparing the proposed method with existing works on Baheya dataset.*

Method	Evaluation Metrics				
	Accuracy	Precision	Recall	F1-Score	AUC
Proposed	**95.80**	**94.30**	**95.10**	**94.70**	**97.20**
(Moon *et al.*, 2020)	90.77	72.50	96.67	82.86	94.89
(Das and Rana, 2021)	88.89	88.00	87.00	87.00	–
(Vigil *et al.*, 2022)	85.30	–	–	–	–

As demonstrated in Table 3, the proposed model achieves higher performance on the Baheya breast ultrasound dataset, with an accuracy of 95.80%, significantly surpassing prior state-of-the-art methods. This high accuracy indicates the robust ability of the model to correctly classify benign and malignant lesions under various ultrasound imaging conditions. The precision score of 94.30% reflects the model's strong capability to reduce false positive diagnoses, which is essential in clinical practice to avoid unwarranted invasive procedures and patient anxiety. Complementing this, the recall of 95.10% highlights the model's sensitivity in detecting true positive malignant cases, thus minimizing the risk of missed cancer diagnoses and enabling timely intervention. The F1-score of 94.70% demonstrates a well-balanced trade-off between precision and recall, ensuring reliable overall classification performance. Furthermore, the area under the ROC curve (AUC) of 97.20% underscores the model's exceptional discriminatory power across various decision thresholds, making it highly adaptable for different clinical risk tolerance levels. These results collectively validate the effectiveness of our approach, which integrates advanced self-supervised learning with multiscale feature fusion techniques to extract nuanced and discriminative features from complex ultrasound images, thereby enhancing malignancy prediction accuracy and supporting more reliable breast cancer diagnosis.

Fig. 3 shows four examples of heatmap visualizations generated by the proposed model for breast tumor classification. In each case, the red regions represent the areas where the network focuses most intensely, effectively highlighting the tumor regions within the images. This attention mapping confirms that the model is successfully identifying and localizing important tumor features. Additionally, the visualizations reveal how the model differentiates between critical tumor tissue and surrounding background areas, which are shown with less intense coloring. These results help demonstrate the model's interpretability and its potential usefulness in assisting medical diagnosis.

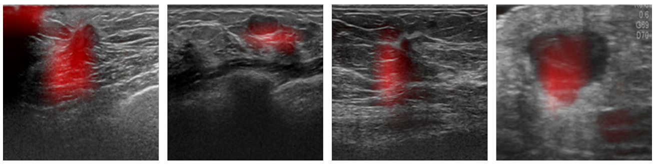

Fig. 3. *Heatmap visualizations from the proposed model highlight breast tumor classification, with red regions showing where the network concentrates most on tumor areas.*

ABLATION STUDIES

To assess the contribution of each major component of our framework, we conducted ablation studies on both the UDIAT and Baheya datasets (Tables 4, and 5). Across both datasets, removing the SSL pretraining produced the largest performance degradation, with accuracy drops of 7–9%, confirming that contrastive SSL provides a significantly stronger initialization than random weights. Eliminating the multiscale fusion module also caused noticeable reductions in recall and AUC, highlighting the importance of aggregating hierarchical features for modeling fine-grained lesion morphology and contextual tissue structure. Using only a single-scale input led to moderate declines (2–3%), indicating that multiscale inputs help capture complementary high-resolution texture cues and lower-resolution contextual patterns. Finally, restricting augmentations consistently lowered performance, particularly on Baheya, where imaging variability is greater. This indicates that diverse stochastic augmentations are essential for stable contrastive pretraining and improved downstream generalization. Taken together, these results demonstrate that SSL, multiscale fusion, multiscale input, and rich augmentation strategies each contribute meaningfully and synergistically to the proposed model's performance.

Table 4. *Ablation study on the UDIAT dataset evaluating the contribution of self-supervised learning (SSL), multiscale feature fusion, multiscale input, and augmentation strategy.*

Experiment Variant	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC (%)**
Proposed Model (SSL + Fusion + Multiscale + Full Aug.)	**96.45**	**95.20**	**96.80**	**96.00**	**98.10**
No SSL (Random Init)	89.30	88.40	89.10	88.75	91.20
No Multiscale Fusion	92.85	91.50	92.30	91.89	94.05
Single-Scale Input Only (224×224)	94.10	92.80	94.50	93.64	95.10
Reduced Augmentations	94.65	93.20	94.10	93.64	96.20

Table 5. *Ablation study on the Baheya dataset. The results further confirm that SSL pretraining and multiscale feature fusion significantly improve model performance, with each component contributing to better generalization and discriminative ability.*

Experiment Variant	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC (%)**
Proposed Model (SSL + Fusion + Multiscale + Full Aug.)	**95.80**	**94.30**	**95.10**	**94.70**	**97.25**
No SSL (Random Init)	87.40	86.10	86.90	86.49	89.80
No Multiscale Fusion	91.75	90.20	91.40	90.79	93.60
Single-Scale Input Only (224×224)	93.20	92.10	92.80	92.45	94.85
Reduced Augmentations	93.85	92.50	93.60	93.04	96.10

CONCLUSION

In this study, we proposed a novel deep learning framework that combines self-supervised learning with multiscale feature fusion to effectively predict breast cancer malignancy from ultrasound images. Our approach addresses key challenges in breast ultrasound analysis by enabling the model to learn rich, hierarchical representations without relying heavily on large amounts of annotated data, thus enhancing feature robustness and generalization. Extensive experiments conducted on two publicly available datasets, UDIAT and Baheya, demonstrate that the proposed method consistently outperforms state-of-the-art CNN architectures and recent self-supervised models across multiple evaluation metrics, including accuracy, precision, recall, F1-score, and AUC. The higher performance reflects the model's enhanced ability to capture subtle morphological and textural variations inherent in ultrasound images, which are critical for reliable malignancy discrimination. Moreover, the high recall and precision rates highlight the model's clinical potential in minimizing both false negatives and false positives, thereby supporting improved diagnostic confidence and patient management. The integration of multiscale feature fusion further enables effective aggregation of features at different resolutions, enriching the learned representations and contributing to the robustness against variations in tumor size, shape, and imaging conditions. In summary, the proposed framework presents a powerful and scalable solution for automated breast cancer diagnosis using ultrasound imaging, with promising implications for clinical practice. Future work will focus on extending this methodology to incorporate multimodal imaging data and exploring interpretability techniques to provide explainable predictions, thereby facilitating broader adoption in real-world healthcare settings.

DATA AVAILABILITY

The authors do not have permission to share data and all the used datasets are publicly available.

FUNDING

This research is supported by The Star Talent Program of Huadong Sanatorium 2023QMX04.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Abdel-Nasser M, Singh VK, Mohamed EM (2022). Efficient staining-invariant nuclei segmentation approach using self-supervised deep contrastive network. Diagnostics 12:3024.

Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A (2020). Dataset of breast ultrasound images. Data Brief 28:104863.

Anandhamala G, *et al.* (2018). Recent trends in medical imaging modalities and challenges for diagnosing breast cancer. Biomed Pharmacol J 11:1649–58.

Anderson BO, Braun S, Lim S, Smith RA, Taplin S, Thomas DB (2003). Early detection of breast cancer in countries with limited resources. Breast J 9:S51–S59.

Ardakani AA, Mohammadi A, Mirza-Aghazadeh-Attari M, Acharya UR (2023). An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. Comput Biol Med 152:106438.

Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, Vignat J, Gralow JR, Cardoso F, Siesling S, *et al.* (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast 66:15–23.

Byra M, Andre M (2019). Breast mass classification in ultrasound based on kendall's shape manifold. arXiv preprint arXiv190511159 .

Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M (2019). Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med Phys 46:746–55.

Dai Z, Liu H, Le QV, Tan M (2021). Coatnet: Marrying convolution and attention for all data sizes. Adv Neural Inf Process Syst 34:3965–77.

Das A, Rana S (2021). Exploring residual networks for breast cancer detection from ultrasound images. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE.

Gómez-Flores W, Gregorio-Calas MJ, Coelho de Albuquerque Pereira W (2024). Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. Med Phys 51:3110–23.

Madjar H (2018). Challenges in breast ultrasound. In: Proceedings of the International Workshop on Medical Ultrasound Tomography: 1.-3. Nov. 2017, Speyer, Germany. KIT Scientific Publishing.

Moon WK, Lee YW, Ke HH, Lee SH, Huang CS, Chang RF (2020). Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput Methods Programs Biomed 190:105361.

Ning Z, Tu C, Xiao Q, Luo J, Zhang Y (2020). Multi-scale gradational-order fusion framework for breast lesions classification using ultrasound images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.

Shurrab S, Duwairi R (2022). Self-supervised learning methods and applications in medical imaging analysis: A survey. PeerJ Comput Sci 8:e1045.

Suzuki K (2017). Overview of deep learning in medical imaging. Radiol Phys Technol 10:257–73.

Vallez N, Bueno G, Deniz O, Rienda MA, Pastor C (2025). Bus-uclm: Breast ultrasound lesion segmentation dataset. Sci Data 12:242.

Vigil N, Barry M, Amini A, Akhloufi M, Maldague XP, Ma L, Ren L, Yousefi B (2022). Dual-intended deep learning model for breast cancer diagnosis in ultrasound imaging. Cancers 14:2663.

Yan P, Gong W, Li M, Zhang J, Li X, Jiang Y, Luo H, Zhou H (2024). Tdf-net: Trusted dynamic feature fusion network for breast cancer diagnosis using incomplete multimodal ultrasound. Inf Fusion 112:102592.

Yap MH, Pons G, Martí J, Ganau S, Sentis M, Zwiggelaar R, Davison AK, Marti R (2017). Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform 22:1218–26.