Informatics in Medicine Unlocked 59 (2025) 101698

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Recognition of breast cancer from heterogeneous ultrasound images: A multi-level deep learning approach

Linh T. Duong [a], Thu T.H. Doan [b], Anh M.T. Bui [c], Phuong T. Nguyen [d,*]

[a] Department of Applied Physics, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden
[b] Gran Sasso Science Institute, Italy
[c] Hanoi University of Science and Technology, Viet Nam
[d] Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, 67100 L'Aquila, Italy

ARTICLE INFO

Keywords:
Deep Learning
Object detection
Breast Cancer Detection

ABSTRACT

Breast ultrasound is a medical imaging technique that employs sound waves to produce breast images, and it has been primarily used to diagnose breast cancer and other related issues. With various machine learning algorithms being applied, many applications have shown promising results and demonstrated outstanding efficiency in giving doctors early accurate diagnoses. By investigating existing state-of-the-art approaches to breast lesion detection, given ConvNeXt-Small architecture as an example, we observe that although they bring a satisfactory performance in classification, their ability to detect small lesions is limited. Therefore, there is still room for improving the performance of DL-based approaches. In this paper, we present a practical Deep Learning-based solution for breast lesion detection, using DetectoRS with Gaussian Receptive Field-based Label Assignment (RFLA) and SegFormer-B4 for recognizing and segmenting small breast lesions, including malignant tumors. Our proposed solution involves three distinct models: ConvNeXt-Small for classification, Swin-Base combined with DetectoRS and RFLA for object detection, and SegFormer-B4 for segmentation. Each model is tailored specifically to address its respective task in breast cancer detection and analysis. The proposed approach has been evaluated on diverse ultrasound datasets. Our deep learning model achieves an Average Precision of 0.270 for small objects (AP_S), and records the highest mean Intersection over Union at 81.55%. The results show that the proposed model outperforms various well-established baselines. We suppose that our method can be integrated into computer-aided diagnosis systems to assist physicians in their clinical activities.

1. Introduction

Breast cancer is the most common cancer among women [1]. Each year, there are about 2.89 million new cases of breast cancer diagnosed worldwide, accounting for about 24.2% of all new cancer cases in women [2]. In fact, breast cancer incidence is higher in developed countries than in less developed countries [3]. However, the mortality rate from breast cancer is higher in less developed countries. This is attributed to the following three reasons: *(i)* Lack of access to early detection and treatment; *(ii)* Socioeconomic factors, such as poverty and lack of education; and *(iii)* Cultural factors, such as taboos about discussing breast cancer. Clinical trials have indicated that early detection and treatment of breast cancer can significantly improve the survival rate [4].

Mammography, digital breast tomosynthesis (DBT), and ultrasound imaging are primary methods to screen for breast cancer in medical facilities. Although they are both capable of detecting abnormalities in the breast, each of them has certain limitations. The mammography method employs radioactivity which poses many health risks to patients. Moreover, it is considered to have low specificity (65%–85%) and high cost, which can lead to the financial burden and unnecessary biopsy operations [5]. Similarly to mammography, DBT also executes radioactivity to render breast images, which makes it become significantly expensive. Compared to the first two methods, ultrasound has a different mechanism of action. It is a non-invasive imaging method, which employs sound waves to generate visual representations of the breast. It is capable of identifying anomalies such as masses, cysts, and other irregularities within the breast. It could bring more benefits, such as real-time imaging, no ionizing radiation, and low cost. These characteristics make it become a common screening or diagnostic tool for breast cancer. However, ultrasound diagnosis requires a high level of technician skill, which makes the diagnostic results could vary between doctors with different training and clinical experiences. This

* Corresponding author.
E-mail addresses: linh.duong@scilifelab.se, lduong@kth.se (L.T. Duong), thihoaithu.doan@gssi.it (T.T.H. Doan), anhbtm@soict.hust.edu.vn (A.M.T. Bui), phuong.nguyen@univaq.it (P.T. Nguyen).

https://doi.org/10.1016/j.imu.2025.101698
Received 27 May 2025; Received in revised form 4 October 2025; Accepted 8 October 2025
Available online 15 October 2025

becomes particularly challenging in developing nations such as Vietnam where there is a shortage of trained physicians and limited access to sonography education [6,7]. Additionally, ultrasound images are often noisy, contain significant artifacts, and have low contrast between tissue structures [8]. Therefore, the development of a computer-assisted breast cancer diagnosis system to aid doctors in their diagnosis is highly desirable.

Many medical imaging techniques have been developed and integrated into various computer-aided detection (CAD) tools to automatically generate diagnoses with acceptable precision [9]. The strong development of deep learning has made significant advancements in the diagnosis and prognosis of various diseases, including breast cancer [6]. This approach has resulted in improved accuracy and understanding of the condition. However, the current CAD systems for breast cancer still stand at a relatively low level of accuracy [10–12]. Previous studies have highlighted that despite being used in clinical facilities for three decades, these existing systems have only led to a small improvement in the accuracy of screening mammograms [13,14]. Therefore, further research is necessary to enhance their accuracy before they can be effectively implemented in real-world scenarios [15,16].

In this paper, we conceptualize a workable solution to the classification of UltraSound images using advanced image processing techniques and deep neural networks as the recognition engines. With the proposed solution, we aim to improve the performance of the DL-based model, working more effectively on small breast lesions existing in UltraSound images. The evaluation shows promising results both in terms of efficiency and effectiveness. In this respect, our work makes the following contributions.

- We developed a unified approach built upon the ConvNeXt model family [17] to address multiple tasks—namely, *classification*, *object detection* and *segmentation* on recently released breast ultrasound image datasets. Our method integrates DetectoRS with Gaussian Receptive Field-based Label Assignment (RFLA) and employs SegFormer-B4 to enhance the recognition and segmentation of small lesions, a combination that, to the best of our knowledge, has not been previously explored.
- We conducted an empirical evaluation to study the proposed approach's performance on the benchmark datasets, comparing to state-of-the-art models. The experimental results show that our proposed approach outperforms various baselines on small-object segmentation and detection.
- We published online a replication package including the data curated and tool developed through this work to foster future research. Our approach can be integrated into computer-aided diagnosis systems to assist physicians in their clinical activities.

The paper is organized as follows. Section 2 provides some background knowledge related to our work. The detail of the proposed approach is described in Section 3. Afterward, Section 4 explains the dataset and metrics used for evaluation. Section 5 reports and analyzes the experimental results, discusses the findings. Section 6 reviews several related works in this field. Finally, we sketch future work and conclude our research in Section 7.

2. Background

This section offers a concise introduction to the ConvNeXt backbones extensively employed in deep learning architectures for the recognition of breast cancer utilizing ultrasound modalities.

Swin Transformer [18] was introduced as a novel architecture that combines the strengths of convolutional neural networks (CNNs) and transformers in computer vision tasks. It employs a multi-stage design, where each stage operates at a distinct feature map resolution. Additionally, a "patchify" layer is used to convert the input image into non-overlapping patches. Swin Transformer has demonstrated impressive performance across various computer vision tasks like image classification, object detection, and instance segmentation.

Liu et al. [17] sought to enhance the performance of the Swin Transformer through modifications in both macro and micro designs. Macro design encompasses the overall network architecture, including the number of stages, blocks in each stage, and the type of blocks used. Micro design focuses on specific details of network layers, such as kernel sizes, activation functions, and normalization layers. They discovered that the following modifications to the Swin Transformer's macro and micro design yielded performance improvements:

- **Altering the stage compute ratio**: The original Swin Transformer had a stage compute ratio of 1:1:3:1. Adjusting the stage compute ratio to 1:1:9:1 resulted in improved network performance is discovered;
- **Utilizing a "patchify" stem**: The original Swin Transformer employed a 7×7 convolution layer with stride 2 as its stem. However, the effectiveness of the "patchify" layer has been demonstrated across diverse tasks such as image classification, object detection, and segmentation. It represents a potent tool capable of enhancing the performance of deep learning models;
- **Employing depthwise convolution**: Depthwise convolution is a type of convolution that operates on each channel of an input tensor independently. This can be useful for reducing the computational cost of convolutions, while still maintaining accuracy. Swin Transformer is a hierarchical vision transformer that uses a combination of depthwise convolution and self-attention to learn long-range dependencies in images. Hence, integrating depthwise convolution into Swin Transformer improved its performance on a variety of image classification and object detection tasks is found;
- **Introducing inverted bottlenecks**: Inverted bottlenecks are a type of design that is used to reduce the number of parameters in a neural network without sacrificing accuracy. They work by first reducing the size of the input data, then expanding it back to its original size. This allows the network to learn more complex features without becoming too large or computationally expensive;
- **Using larger kernel sizes**: Swin Transformer was originally designed with 3×3 convolution kernels. However, researchers later found that using larger kernel sizes, such as 7×7, improved the network's performance. This is because larger kernel sizes allow the network to learn more complex relationships between features;
- **Substituting the activation function ReLU with GELU**: ReLU and GELU are distinct activation functions. ReLU [19] is a simple function that outputs 0 for negative inputs and the input value for positive inputs. Meanwhile, GELU [20] is a more complex function that is a smooth approximation of the error function. The study found that replacing the ReLU activation function with the GELU activation function improved the performance of the Swin Transformer model on a variety of tasks, including image classification;
- **Reducing the number of activation functions**: The original Swin Transformer incorporated an activation function after each convolutional layer. However, eliminating certain activation functions can encourage the network to acquire more intricate data representations, ultimately resulting in improved performance, as demonstrated by the findings of this research;
- **Decreasing the number of normalization layers**: The original Swin Transformer architecture implemented two normalization layers per block to promote training stability and mitigate overfitting. However, reducing the number of normalization layers to one per block did not substantially increase the risk of overfitting. Instead, it yielded performance improvements across various tasks for the network;

- **Substituting BN with LN0**: Batch normalization (BN) [21] is commonly used in CNNs, while layer normalization (LN) [22] is commonly used in transformers. A significant finding of the study is that replacing BN with LN resulted in improved performance for the Swin Transformer.

3. Proposed approach

We propose a practical solution to the classification, detection, and segmentation of breast cancer from heterogeneous ultrasound images. The framework is built exploiting the potential of advanced deep learning models including ConvNeXt for image classification (Section 3.2), Swin-Base architecture for object detection (Section 3.3), and SegFormer for segmentation (Section 3.4).

3.1. The workflow of our approach

Fig. 1 illustrates our proposed paradigm, which is divided into four main stages, i.e., data collection, classification, object detection and segmentation, and performance comparison, corresponding to steps *1*, *2*, *3* and *4*.

The process begins with gathering data from a variety of public repositories to ensure a diverse and comprehensive dataset (*Step 1*). This dataset, essential for training and evaluating the proposed deep learning approach, consists of ultrasound images labeled as *normal*, *benign*, or *malignant* tissue. In the initial analysis phase, we train a ConvNeXt model on this heterogeneous dataset for image classification (*Step 2*), enabling them to distinguish between the different ultrasound image categories. At this stage, we aim to evaluate the efficacy of different ConvNeXt variants to identify the most accurate model in terms of classification. Building on the strong classification results, we advance to *Step 3*, which targets object detection. At this stage, we first assess the ConvNeXt family's ability to localize breast lesions. Our experiments show that, despite ConvNeXt's effectiveness for classification, its performance is comparatively weak on both object detection and segmentation. Accordingly, we evaluated two transformer-based architectures purpose-built for object detection, including Swin-Base [18] and DETR-ResNet50 [23] and benchmarked their performance against the ConvNeXt family. We conduct a focused analysis of small-lesion detection performance. Our results show that integrating Swin-Base with DetectoRS and RFLA achieves the strongest performance in this setting. Concurrently, *Step 4* explores the capability of ConvNeXt family and other Transformer-based architecture for segmentation task. In this stage, segmentation extends beyond detection to the precise delineation of lesion boundaries. Accurately identifying small lesions within the breast cancer category remains particularly challenging, presenting complexities comparable to those in object detection. We adopt the SegFormer family [24] as the segmentation backbone and benchmark its performance against the ConvNeXt models. The results show that SegFormer-B4 outperforms both the other SegFormer variants and all ConvNeXt variants, achieving superior segmentation across small, medium, and large lesions.

3.2. Classification

An efficient and streamlined framework designed for the classification of breast lesions. In our classification task, we assess the performance of the entire ConvNeXt model range (Atto, Femto, Pico, Nano, Tiny, Small, Base, and Large) [17,25]. These variants are distinguished by their convolutional layer configurations, such as kernel size, number of channels per stage, and network depth, among other architectural parameters. The Atto, Femto, Pico, and Nano models are optimized for resource-limited environments, whereas the Tiny, Small, Base, and Large models are designed to deliver high performance in more capable computing settings. Our research benefits from the scaled-down versions of ConvNeXt models provided by HuggingFace Inc., featuring reduced-size backbones. We further evaluate the ConvNeXt models' effectiveness and compare them with other scalable backbone frameworks for image classification.

The ConvNeXt framework represents a novel approach to image classification, drawing on the principles of the Swin Transformer architecture [18] and inheriting its key advantages, as detailed in Section 2. Notable for its simplicity, efficiency, and high performance, ConvNeXt distinguishes itself in this domain. Our extensive experimental results provide compelling evidence that ConvNeXt's streamlined and efficient design plays a critical role in achieving superior classification performance, particularly for the identification of breast cancer lesions.

3.3. Object detection

We benchmark mid-sized ConvNeXt backbones for object detection, with a focused analysis of ConvNeXt-Tiny and ConvNeXt-Small. While these models have shown strong potential in classification, we assess their effectiveness on detection by comparing them against well-established baselines—Swin-Base [18] and DETR-ResNet50 [23].

3.4. Semantic segmentation

When it comes to segmentation tasks, we encounter challenges in accurately distinguishing among small lesions within the breast cancer category, which resemble the difficulties faced in object detection. To tackle this problem, we assess the performance of the Segformer family [24], a fusion of Transformers and lightweight multilayer perceptron decoders designed for identifying breast lesions. SegFormer is a new semantic segmentation framework that combines Transformers with lightweight multilayer perceptron (MLP) decoders. The network is simple, efficient, and it has the following two key features:

- SegFormer incorporates a hierarchically structured Transformer encoder with two components: a coarse encoder and a fine encoder. The coarse encoder extracts coarse features from the input image, while the fine encoder captures fine-grained features. Together, these two encoders generate multiscale features that are highly valuable for effective semantic segmentation;
- The MLP decoder in SegFormer consists of multiple MLP layers that leverage both local and global attention mechanisms to integrate information from different layers. By employing these attention mechanisms, the decoder is capable of generating robust representations, which are then utilized to predict the semantic segmentation mask.

4. Evaluation

This section explains in detail the evaluation process conducted to study the performance of our proposed approach. First, Section 4.1 briefly introduces the evaluation metrics to measure the prediction performance in this work. Second, in Section 4.2 we present the dataset used in the evaluation. Finally, we describe the plan for evaluation in Section 4.3, with more detail about the datasets and our setting environment.

4.1. Evaluation metrics

To evaluate the performance of our approach in classifying UltraSound images, we use four metrics including *Accuracy*, *Precision*, *Recall*, and F_1 *score* widely applied in classification tasks. Given a set of images, associated with a set of ground-truth labels, i.e., $G = (G_1, G_2, \ldots, G_N)$, our model generates predictions for the given images, resulting in a set of predicted labels, i.e., $C = (C_1, C_2, \ldots, C_N)$. Considered $match_i$, $|C_i|$, and $|G_i|$ as the number of correct predictions of class i, the number of samples of class i, and the number of samples predicted to class i, respectively, the four metrics are defined as follows.

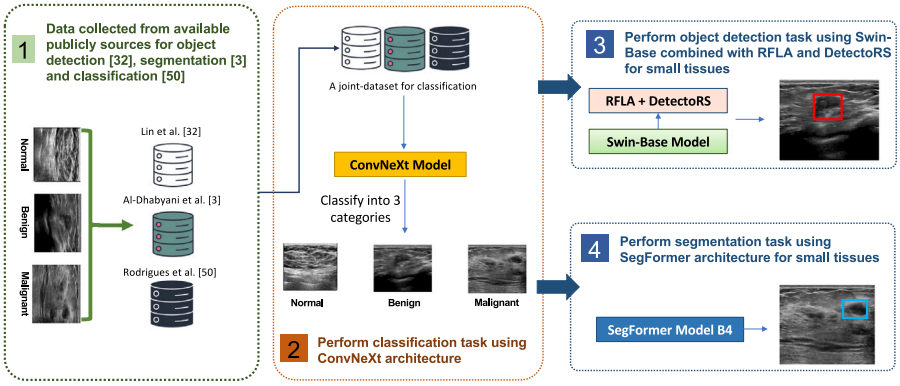


Fig. 1. The proposed workflow of our study. For the sake of clarity, we emphasize on the models with the highest performance for object detection and segmentation tasks (Step 3 and Step 4).

Accuracy. It is calculated as the ratio of the number of correct instances to the total number of items.

$$accuracy = \frac{\sum_i^N match_i}{\sum_i^N |G_i|} \times 100\% \quad (1)$$

Precision and Recall. While *Precision* measures the proportion of correct positive predictions, *recall* measures the coverage of actual positive labels. These two metrics are computed using the following formulas.

$$precision_i = \frac{match_i}{|C_i|} \quad (2)$$

$$recall_i = \frac{match_i}{|G_i|} \quad (3)$$

F_1 score (F-Measure). F_1-score is computed as the harmonic mean of Precision and Recall:

$$F_1 = \frac{2 \cdot precision_i \cdot recall_i}{precision_i + recall_i} \quad (4)$$

Evaluation metrics for the object detection and segmentation task. We use widely-used metrics for quantitatively comparing different ultrasound video breast lesion detection methods: *(i)* The Average Precision (AP) metric; *(ii)* the Area Under the Precision-Recall (PR) Curve (AUC-PR), being calculated for each class separately. Meanwhile, the Mean Average Precision metric is the overall average of the AP values, AP_{mean} and AP_α metric is the Average Precision at the IoU threshold of an α value, for example, AP_{50} and AP_{75}. Furthermore, we also calculate AP for small objects (AP_S) if the area of an object sizes between 10 × 10 and 32 × 32 pixels, AP for medium objects (AP_M) if the area of objects ranges from 32 × 32 to 96 × 96 pixels, and AP for large objects (AP_L) if the area is larger than 96 × 96 pixels.

4.2. Dataset

In our data curation task, we conducted a literature review and searched for publicly available sources such as Zenodo, Kaggle, and others. We then built combined datasets for image recognition tasks using data from three sources [26–28]. Further details of the datasets are provided in Table 1. The dataset from Lin et al. [26], containing 25,407 images, is utilized for the object detection task. The dataset by Al-Dhabyani et al. [27] with 780 images is employed for the segmentation task. Additionally, the dataset from Rodrigues et al. [28], which includes 250 images, is integrated with the previous datasets for the classification task, resulting in a combined total of 26,437 images. These benchmark datasets consist of distinct image samples that are classified into three categories: *benign*, *malignant*, and *normal*. For the classification and object detection tasks, the data is divided into training, validation, and testing sets using an 8:1:1 ratio. For the segmentation task, the split follows a 64:16:20 ratio in alignment with the configuration of the original study [27]. The configuration details for the three tasks – classification, object detection, and segmentation – are presented in Tables 2, 3, and 4, respectively.

Table 1
Summary of datasets collected from various sources.

Source	Original task	Category			Total
		Benign	Malignant	Normal	
[26]	Object detection	9932	15,475	0	25,407
[27]	Segmentation	437	210	133	780
[28]	Not applicable	100	150	0	250
Total	–	10,469	15,835	133	26,437

Table 2
A merged dataset used for a classification task.

Subset	Category			Total
	Benign	Malignant	Normal	
Training	8375	12,668	106	21,149
Validation	1046	1583	13	2642
Testing	1048	1584	14	2646
Total	10,469	15,835	133	26,437

Table 3
The dataset used for object detection.

	Benign	Malignant	Total
Training	7061	10,707	20,408
Validation	1500	1500	3500
Testing	1249	3292	3541
Total	9810	15,499	25,272

Table 4
The dataset used for the segmentation task.

Type	Category			Total
	Benign	Malignant	Normal	
Training	280	134	85	499
Validation	70	34	21	125
Test	87	42	27	156
Total	437	210	133	780

4.3. Experimental settings

The experiments were conducted on a server equipped with an Intel® Xeon® CPU E5-2680v4 @ 2.50 GHz (14 cores), 96 GiB RAM, and an NVIDIA GeForce RTX 3090 GPU, running Ubuntu 20.04.4 LTS.

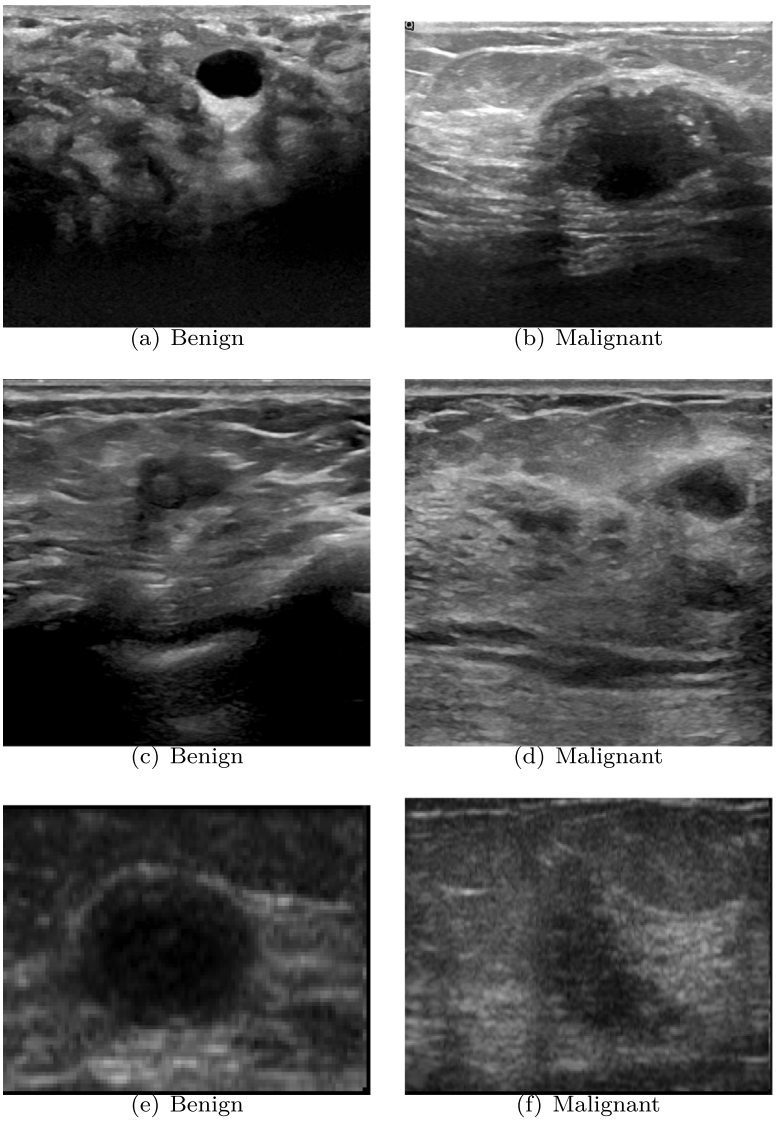


Fig. 2. Breast ultrasound examples of datasets from Al-Dhabyani et al. [27], Lin et al. [26], and Rodrigues et al. [28].

The batch size was configured to fit within the 24 GB VRAM capacity of the RTX 3090. Our deep learning framework was implemented using PyTorch[1] version 1.12 with CUDA 11.6 [29,30], in conjunction with Python 3.10 and Timm version 0.6.11 [25].

Classification task. For the classification task experiments, the dataset was divided into three independent subsets: 80% for training, 10% for validation, and 10% for testing. This corresponds to 21,149 images for training, 2642 images for validation, and 2646 images for testing (further details are provided in Tables 1 and 2). The configuration details and hyper-parameters of all ConvNeXt variants are introduced in Table 5.

Object detection task. For the object detection task, the dataset obtained from Lin et al. [26] was partitioned into three independent subsets: 95% of the original training set for training, 5% of the original training set for validation, and the original validation set serving as the testing set in this study (as shown in Table 3). We employ two ConvNeXt variants including ConvNeXt-Tiny and ConvNeXt-Small, configured as specified in Table 5. Additionally, we compare their performance against Swin-Base [18] and DETR [23] to assess their effectiveness in recognizing small objects. All the implementations were carried out using MMDet version 2.25.3 [31].

Segmentation task. For the segmentation task, the dataset curated from Al-Dhabyani et al. [27] was divided into independent training, validation, and testing sets in a 64%:16%:20% ratio, respectively. Further details on the dataset are provided in Table 4. We also utilize ConvNeXt-Tiny and ConvNeXt-Small with the same configuration specified in Table 5. Their performance is compared against five variants of the SegFormer family [24], which has been specifically developed for segmentation tasks. All the experiments were implemented using MMSeg version 0.29.0 [32].

Several instances of breast issues are given in Fig. 2. For each of the three selected datasets, two images are presented, corresponding to the case of benign and malignant tumors. In detail, three pairs of images, regarded as Figs. 2(a), 2(b), 2(c), 2(d), and 2(e), 2(f), are selected from the datasets curated by Al-Dhabyani et al. [27], Lin et al. [26], and Rodrigues et al. [28], respectively. For ordinary people without in-depth medical knowledge, data engineers in particular, it could be a little hard for them to recognize the breast issue shown in these images. Moreover, from a computer vision perspective, the tumors that represent the issue are usually small in size. From this observation, our work investigates the best deep-learning-based solution to effectively handle small objects for breast tumor recognition.

[1] https://pytorch.org.

Table 5
Experimental configurations for the classification task.

Conf.	Network	Batch size	# of Params (Milions)	MAC	Size (MB)
C_1	ConvNeXt-Atto	1100	3.7	551.16M	13.5
C_2	ConvNeXt-Femto	950	5.22	784.22M	25.5
C_3	ConvNeXt-Pico	650	9.05	1.37G	34.2
C_4	ConvNeXt-Nano	450	15.59	2.45G	59.8
C_5	ConvNeXt-Tiny	350	28.9	4.46G	111.3
C_6	ConvNeXt-Small	210	50.22	8.69G	205.3
C_7	ConvNeXt-Base	150	88.59	15.36G	350.3
C_8	ConvNeXt-Large	90	197.77	34.37G	785.0

All configurations and hyper-parameters for training processes of these recognition tasks can be found in our GitHub repository.[2]

5. Experimental results

Through an extensive set of experiments, we gathered evidence to support the claim that the streamlined and efficient architecture of SegFormer plays a pivotal role in achieving better segmentation results specifically for tiny lesions in the malignant breast class.

5.1. Classification

We are interested in understanding the efficiency and effectiveness of our proposed approach. To this end, we performed prediction with ConvNext family [17,25], aiming to identify the model that brings the best prediction accuracy as well as timing efficiency on the curated dataset, which is merged from various open sources [26–28]. As shown in Fig. 3, ConvNeXt-Small brings the best performance with an accuracy of 93.84%, and ConvNeXt-Base and ConvNeXt-Tiny are the runner-up with accuracies of 93.76% and 93.24%, respectively. The results are a hint to conduct further experiments for object detection and segmentation tasks.

Fig. 3 depicts eight confusion matrices achieved by eight models of the ConvNeXt family in the classification task. We perform this task on the merged dataset which has been introduced in Table 2. The samples in this dataset are categorized into three groups, regarded as *Benign, Malignant*, and *Normal*. Each of the confusion matrices indicates the Precision, Recall, and F1-score corresponding to these categories. It can be seen that there is a significant disparity in the metrics of these eight models. Figs. 3(a), 3(b), 3(c), and 3(h) illustrate four configurations with markedly poor classification performance: the *Normal* class is entirely misclassified, resulting in 0% Precision and 0% Recall. This outcome is attributable to the imbalance of the *Normal* class in the dataset and, moreover, underscores the performance limitations of certain ConvNeXt variants. Besides, the remaining four configurations achieve clearly better results, regarded as ConvNeXt_Nano (Fig. 3(d)), ConvNeXt_Tiny (Fig. 3(e)), ConvNeXt_Small (Fig. 3(f)), and ConvNeXt_Base (Fig. 3(g)). As shown in Fig. 3(d), ConvNeXt_Nano configuration achieves higher Accuracy of 93.31%, compared to three configurations in the same line including ConvNeXt_Atto, ConvNeXt_Femto, and ConvNeXt_Pico. Although it achieves high Precision for category Benign (96.10%) and Malignant (91.86%), category Normal only accounts for 60% of Precision, and 21.43% of Recall, resulting in 72.59% of F1-score in total.

Considered the last three configurations, they achieve better classification performance for both three categories with higher Precision and Recall. In which, ConvNeXt-Small brings the best performance with an accuracy of 93.84%, and ConvNeXt-Base and ConvNeXt-Tiny are the runner-up with accuracies of 93.76% and 93.24%, respectively. F1-score of these three models is in the different order, with the best of 89.07% being also achieved by ConvNeXt-Small, and 86.98% by ConvNeXt-Tiny, and 81.45% by ConvNeXt-Base. The results are a hint to conduct further experiments for object detection and segmentation tasks.

5.2. Object detection

Following the overall methodological framework, once the classification phase is completed, the subsequent step involves object detection. This stage emphasizes the use of two ConvNeXt variants – ConvNeXt-Tiny and ConvNeXt-Small – selected on the basis of their demonstrated superiority over other models in accurately classifying malignant tissues.

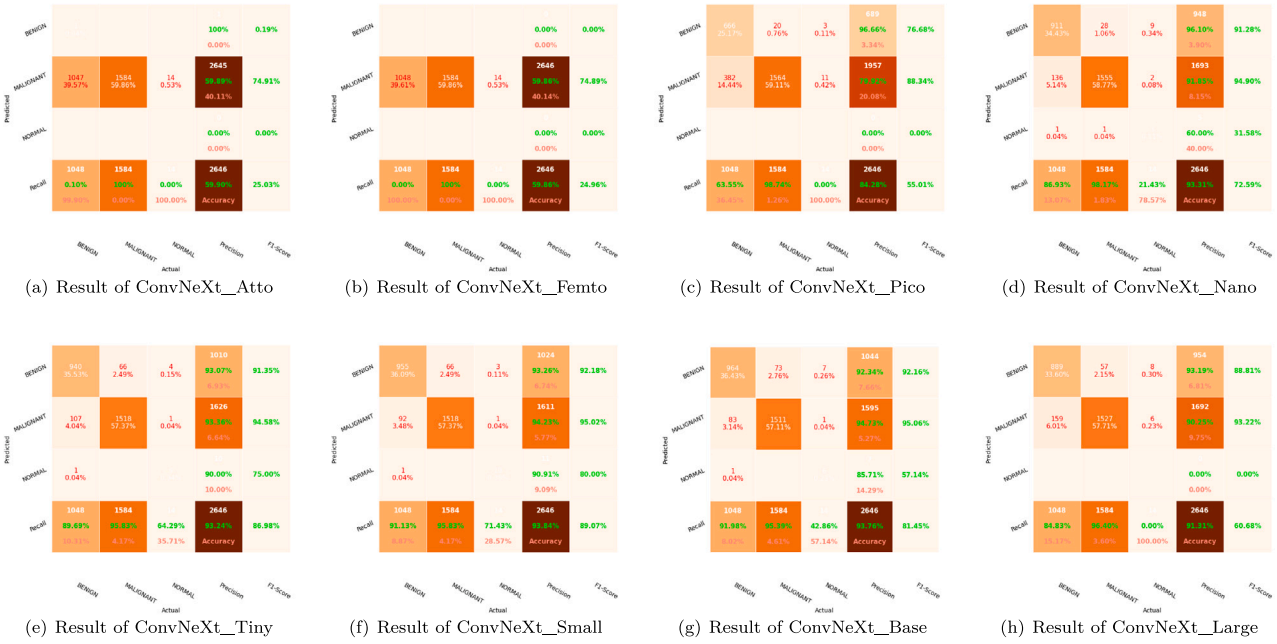
During the object detection experiments conducted on the dataset from Lin et al. [26], we observed that ConvNeXt [17] struggled to accurately detect small objects, with performance particularly limited in the case of small malignant lesions. Recognizing this shortcoming, we expanded our analysis by benchmarking against leading Transformer-based architectures. In particular, we evaluated Swin-Base [18] and DETR-ResNet50 [23], both of which have been established in the literature as state-of-the-art baselines for object detection. Table 6 presents the recognition results of four models in terms of six criteria including $AP_{mean}, AP_{50}, AP_{75}, AP_S, AP_M$ and AP_L where AP_S, AP_M and AP_L are employed to evaluate the average precision of small, medium and large lesion detection. The details of all the metrics have been detailed in Section 4.1.

As shown in Table 6, ConvNeXt-Tiny achieves values of 0.207, 0.384, and 0.217 for AP_{mean}, AP_{50}, and AP_{75}, respectively, while ConvNeXt-Small attains slightly higher scores of 0.244, 0.435, and 0.266 for the same metrics. Notably, both models clearly surpass most of the other approaches in detecting medium objects, as indicated by their leading AP_M scores of 0.011 for ConvNeXt-Tiny and 0.008 for ConvNeXt-Small. In the case of large objects, their performance is second only to the Swin-Base model (0.265), with ConvNeXt-Tiny and ConvNeXt-Small achieving 0.211 and 0.248, respectively, substantially exceeding the results obtained by DETR-ResNet50. Regarding Swin-Base, it is evident that this model performs strongly, achieving the highest scores for AP_{mean}, AP_{50}, AP_{75} and AP_L. Despite these strengths, all four models exhibit extremely poor capability in detecting small tissues, as reflected by an AP_S value of 0.000.

To mitigate the challenges associated with small object detection, we incorporate Gaussian Receptive Field-based Label Assignment (RFLA) [33], a technique that has demonstrated strong effectiveness in satellite imaging tasks, where the target objects are typically very small. Given that this technique is explicitly tailored for small object detection, we present only the AP_{mean} and AP_S values obtained by integrating RFLA with ConvNeXt-Small and Swin-Base, as these models demonstrated promising results in the preceding analysis. As shown in Table 6, Swin-Base outperforms ConvNeXt-Small in small tissue detection when combined with RFLA, achieving an AP_S of 0.265 compared to 0.252 for ConvNeXt-Small. Finally, we evaluate a hybrid configuration that combines Swin-Base with DetectoRS [34] and RFLA, leveraging Swin's strong hierarchical representations [18] together with the Recursive Feature Pyramid (RFP) in DetectoRS for recursive multi-scale refinement [35]. As reported in Table 6, the combination of Swin-Base with DetectoRS and RFLA attains the best performance, achieving the highest AP_{mean} (0.243) and AP_S (0.270).

We select two examples of benign and malignant tumors and perform object detection by using the seven architectures mentioned in Table 6. We visualize their detecting results of bounding boxes in Fig. 5(a–d). Concerning the results in Figs. 4(a), 4(b), 4(c), and 4(d), it can be seen that ConvNeXt models bring outstanding performance for the benign tumor with the probability of 1.00, very close to the ground-truth. However, for the malignant tumor, they detect two objects in which the larger tumor is correctly identified with the probability of 1.00 while the smaller one is over-recognized with relatively low

[2] https://github.com/linhduongtuan/Breast_Ultrasound_recognition.



(a) Result of ConvNeXt_Atto (b) Result of ConvNeXt_Femto (c) Result of ConvNeXt_Pico (d) Result of ConvNeXt_Nano

(e) Result of ConvNeXt_Tiny (f) Result of ConvNeXt_Small (g) Result of ConvNeXt_Base (h) Result of ConvNeXt_Large

Fig. 3. Confusion matrices on the independent test sets using our proposed models.

Table 6
Quantitative comparisons of ConvNeXt-Tiny and ConvNeXt-Small [17], Swin-Base [18], and DETR [23] on the annotated test set [26].

Model name	AP_{mean}	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ConvNeXt-Tiny	0.207	0.384	0.217	0.000	**0.011**	0.211
ConvNeXt-Small	0.244	0.435	0.266	0.000	0.008	0.248
Swin-Base [18]	**0.262**	**0.445**	**0.302**	0.000	0.003	**0.265**
DETR-ResNet50 [23]	0.080	0.140	0.109	0.000	0.001	0.082
ConvNeXt-Small w/ RFLA	0.238	–	–	0.252	–	–
Swin-Base w/ RFLA	0.240	–	–	0.265	–	–
Swin-Base w/ DetectoRS & RFLA	0.243	–	–	**0.270**	–	–

confidence of 0.58 and 0.31 for ConvNeXt-Tiny and ConvNeXt-Small, respectively.

Turning to Swin-Base, this transformer-based model also achieves strong results for both benign and malignant cases, with confidence scores of 1.00. In contrast, DETR frequently mislocalizes both benign and malignant tumors, yielding bounding boxes that fail to correctly cover the target. As illustrated in Fig. 5(e–j), the combination of Swin-Base with RFLA and DetectoRS correctly identifies the targets with a confidence of 1.00, highlighting the effectiveness of RFLA for small-object detection in breast ultrasound images.

Finally, we assess the total prediction time across the seven models (in seconds). Fig. 6 depicts the cumulative time each model requires on the same test set. As expected, augmenting the Swin-Base backbone with RFLA and DetectoRS increases computational overhead, resulting in longer inference times for malignant-lesion detection compared with the other configurations.

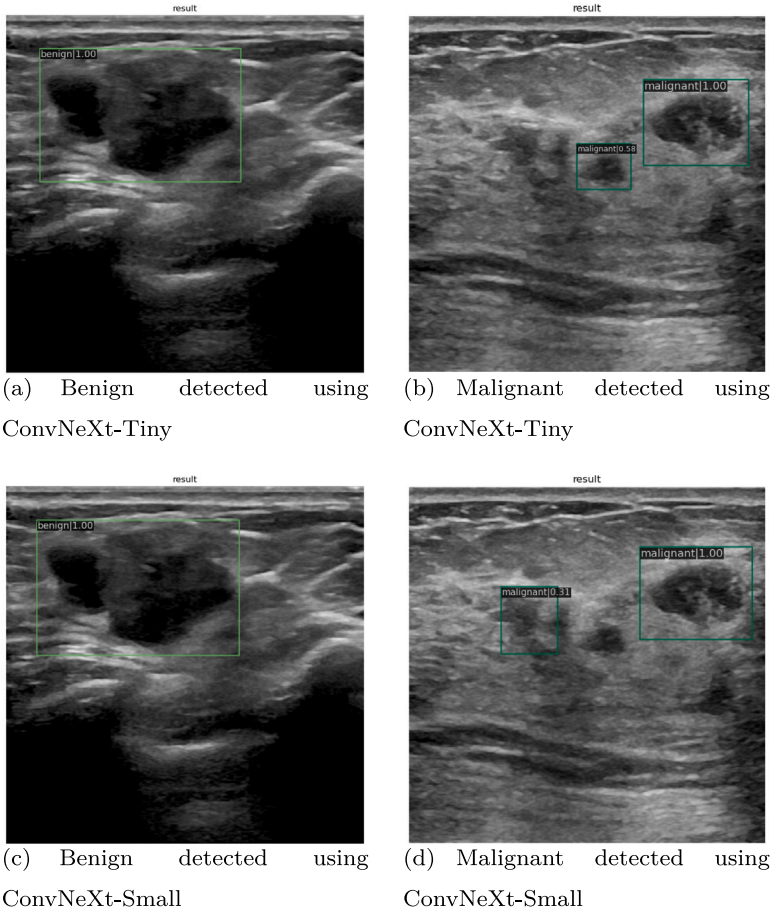
5.3. Segmentation tasks

We investigate the segmentation task on the dataset curated from Al-Dhabyany et al. [27]. In this phase, we examine two ConvNeXt variants – ConvNeXt-Tiny and ConvNeXt-Small – chosen for their strong performance in malignant-tissue classification. We further benchmark them against a Transformer-based segmentation architecture, SegFormer, considering five variants: SegFormer-B0, SegFormer-B1, SegFormer-B2, SegFormer-B3, and SegFormer-B4 [24].

As summarized in Table 7, we focus on *IoU* and *Accuracy* as the primary metrics for comparing model performance on the segmentation task. For each model, the table reports IoU and Accuracy separately for the *Benign* and *Malignant* classes. In addition, we provide the mean IoU (mIoU), mean Accuracy (mAcc), and average Accuracy (aAcc). Focusing on the *Malignant* class, the ConvNeXt family underperforms relative to the SegFormer variants: ConvNeXt-Tiny and ConvNeXt-Small achieve IoUs of 22.05 and 0.48, respectively, whereas all SegFormer variants exceed 50 on the value of IoU. Although ConvNeXt-Small achieves the highest accuracy for the *Benign* class, its accuracy for the *Malignant* class is only 0.500, yielding a mean accuracy (mAcc) of

Table 7
Quantitative comparisons of ConvNeXt-Tiny, ConvNeXt-Small [17], SegFormer-B0, SegFormer-B1, SegFormer-B2, SegFormer-B3, and SegFromer-B4 [24] on the annotated test set [27].

Model name	Category	IoU	Acc.	mIoU	aAcc.	mAcc.
ConvNeXt-Tiny	Benign	66.08	68.41	44.06	69.05	71.10
	Malignant	22.05	73.80			
ConvNeXt-Small	Benign	87.86	**99.64**	44.17	87.87	50.07
	Malignant	0.480	0.500			
SegFormer-B0	Benign	**97.01**	97.34	73.25	93.61	80.92
	Malignant	54.97	64.51			
SegFormer-B1	Benign	93.19	97.58	73.06	93.68	80.43
	Malignant	53.22	63.27			
SegFormer-B2	Benign	93.84	97.66	75.77	94.32	82.94
	Malignant	57.70	68.22			
SegFormer-B3	Benign	94.23	62.65	78.44	94.74	78.44
	Malignant	62.65	77.67			
SegFormer-B4	Benign	95.08	97.10	**81.55**	**95.54**	**90.26**
	Malignant	**68.02**	**83.43**			



(a) Benign detected using ConvNeXt-Tiny (b) Malignant detected using ConvNeXt-Tiny

(c) Benign detected using ConvNeXt-Small (d) Malignant detected using ConvNeXt-Small

Fig. 4. Examples of breast ultrasound detection using architectures of ConvNeXt-Tiny and ConNeXt-Small.

50.07% across the dataset. By contrast, the average accuracy (aAcc) remains high at 87.87%, reflecting class imbalance. Consistent with the object-detection results, these findings indicate that the ConvNeXt family struggles to recognize small malignant lesions while performing well on larger objects.

In comparison with the ConvNeXt family, the SegFormer variants achieve higher IoU for both the *Benign* and *Malignant* classes, leading to correspondingly higher accuracies. Notably, SegFormer-B0 attains the best IoU for *Benign* (97.01%). For *Malignant*, SegFormer-B4 achieves the highest IoU (68.02%), which corresponds to the top Accuracy (83.43%). Overall, SegFormer-B4 delivers the strongest performance on the dataset, with mIoU of 81.55%, aAcc of 95.54%, and mAcc of 90.26%.

We present example segmentation masks produced by SegFormer-B0 and SegFormer-B4 in Fig. 7. Consistent with the quantitative results in Table 7, Fig. 7(a) shows that SegFormer-B4 yields masks that closely match the ground truth, particularly for small malignant lesions.

5.4. Threats to validity

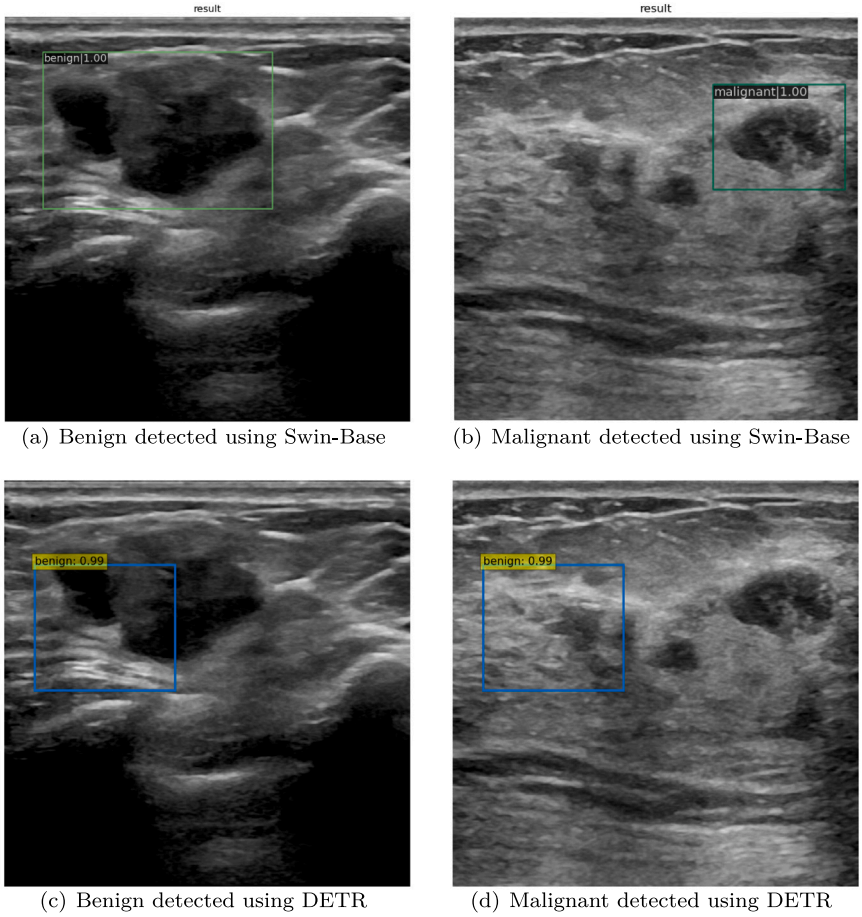
This section discusses the potential limitations of our study, which could affect the internal, external, construct, and conclusion validity of our findings.

Internal validity. These internal factors could influence the evaluation process. One potential concern is the comparison with the baseline methods. To address this concern, we take steps to minimize the impact by conducting experiments using the original implementations. Additionally, we test all the relevant tools on our collected dataset and compared their performance using the same set of metrics.

External validity. The primary concern regarding *external validity* is the generalizability of our findings beyond the scope of this study. To address this issue, we mitigated the threat by conducting evaluations of our proposed models on large and diverse datasets that encompass various categories. We conducted additional rounds of experiments to further validate and reinforce our final results. By taking these measures, we aim to minimize the potential threat to *external validity* and enhance the reliability and applicability of our findings.

Construct validity. This pertains to the experimental setup employed in our study, specifically regarding the simulated configurations used to evaluate our approach. We conduct the evaluation using separate training and test sets, which may not fully represent real-world usage scenarios. To ensure a fair comparison, we employed identical configurations when comparing the different systems. By employing consistent settings, we aim to minimize any potential biases and facilitate a more meaningful and equitable evaluation of the approaches.

Conclusion validity. This pertains to the question of whether the experimental methodology employed in our study has a direct connection to the observed results, or if there are additional factors that could impact the outcomes. The evaluation metrics utilizes, including accuracy, precision, recall, F1 score, mAP, IoU, and execution time, present a potential risk to the validity of our conclusions. To tackle this concern, we take steps to address the issue by utilizing the same set of metrics when comparing the various systems. Our intention in employing consistent evaluation criteria is to reduce any potential biases and ensure an equitable and meaningful comparison between the different approaches.



(a) Benign detected using Swin-Base (b) Malignant detected using Swin-Base

(c) Benign detected using DETR (d) Malignant detected using DETR

Fig. 5(a–d). Examples of breast ultrasound detection using architectures of Swin-Base, and DETR-ResNet50 (continued).

6. Related work

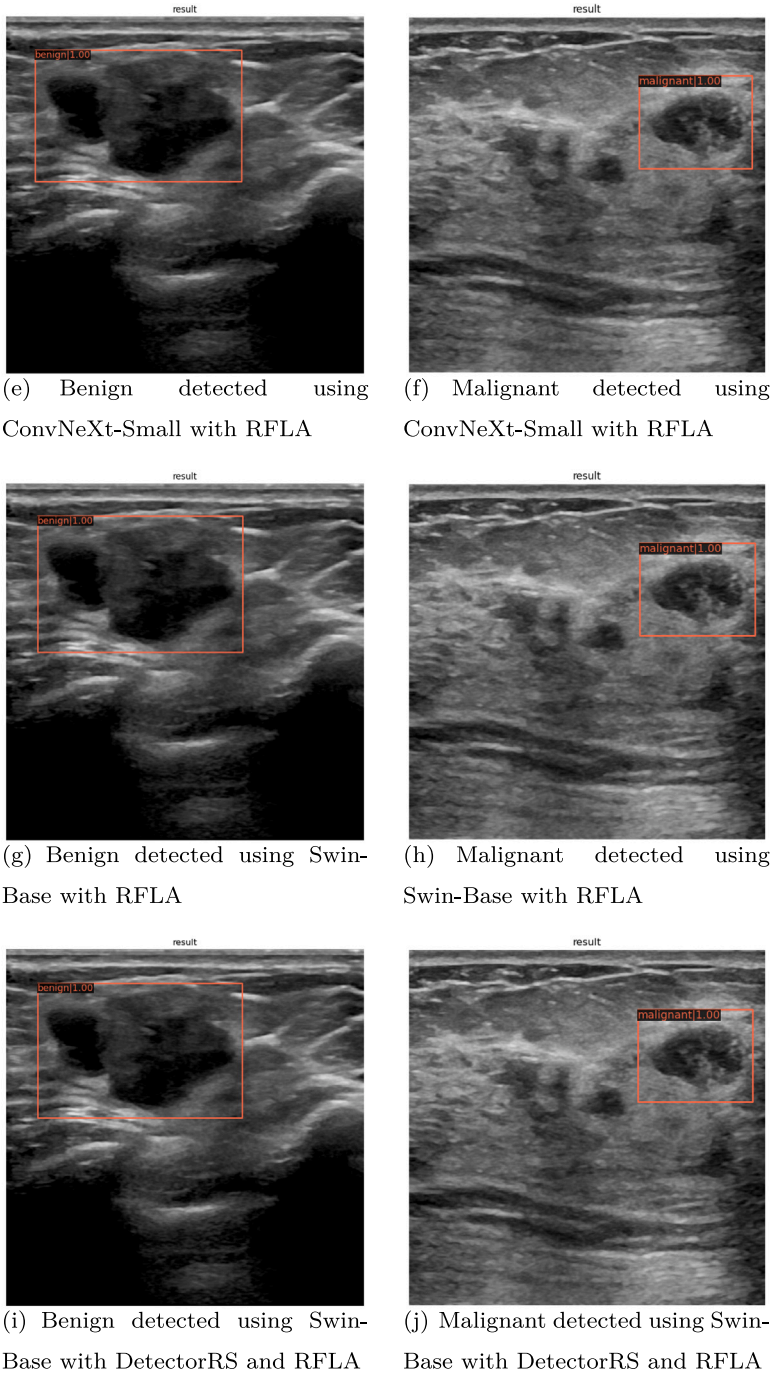
In our research, we specifically concentrate on utilizing ultrasound imaging to identify abnormalities in breast tissue. Therefore, in this section, we delve into the significant research conducted in this field, giving special attention to the datasets utilized and the approaches employed for the breast cancer image identification tasks including classification, detection, and segmentation.

6.1. Classification approaches

Many breast ultrasound (BUS) image detection approaches have been investigated in the last two decades, in which deep learning algorithms have recently drawn the interest of researchers. Notably, Convolutional Neural Networks (CNNs) have emerged as successful tools for classifying and recognizing patterns in medical images, particularly for breast ultrasound images [36–38]. Khanna et al. [36] have proposed a hybrid approach that combined a pre-trained CNN model (i.e., ResNet-50 CNN [39]) with a binary gray wolf optimization algorithm (BGWO) for feature selection. They employed Support Vector Machine (SVM) model for final classification. In this study, two approaches have been investigated, including (i) a transfer learning approach and (ii) machine learning approach. The proposed method achieved an accuracy rate of 84.9% for classification and an AUROC score of 0.97 when evaluated on the BUSI dataset [27]. The concept of employing feature selection to enhance the classification model's performance is similarly applied by Jabeen et al. [40]. In their study, the authors introduced a combination of a pre-trained DarkNet-53 model [41] with two optimization algorithms: reformed differential evaluation (RDE) and reformed gray wolf (RGW) in order to identify the most essential features to classifying breast tissues. Upon transferring the model on the BUSI dataset, the achieved results have showcased an exceptionally high accuracy rate of 99.1%. Transfer learning approach with convolutional neural networks (CNNs) has been also adopted in the work of Byra [37]. The author proposed to adjust the pre-trained CNN to enable better flow of information in the network by using deep representation scaling layers (DRS). This study showed that the application of DRS allows to reduce considerably the number of fine-turned parameters. The proposed approach demonstrated impressive results on the BUSI dataset [27] with an AUC score of 0.955 and an accuracy rate of 0.915.

Different to prior studies which primarily focused on binary classification of benign versus malignant tumors, Behboodi et al. [42] proposed a multi-class classification deep learning-based approach in which the background tissue is added as an additional class. The effectiveness of their method was appraised through the utilization of two distinct pre-trained architectures: ResNet-34 [39] and MobileNet-v2 [43]. This evaluation revealed notable enhancements in the AUC score, with an approximately 31% increase for ResNet-34 and a 9% increase for MobileNet-v2.

Recently, Kalafi et al. [38] have introduced a novel framework for the classification of breast cancer lesions. They proposed to modify the VGG16 architecture by incorporating an attention module to enhance the discernment of features in order to distinguish between background and targeted lesions in ultrasound images. Furthermore, they put forward an ensemble loss function that combines binary cross-entropy with the logarithm of the hyperbolic cosine loss. This combined loss function aimed to improve the alignment between classified lesions and



(e) Benign detected using ConvNeXt-Small with RFLA

(f) Malignant detected using ConvNeXt-Small with RFLA

(g) Benign detected using Swin-Base with RFLA

(h) Malignant detected using Swin-Base with RFLA

(i) Benign detected using Swin-Base with DetectorRS and RFLA

(j) Malignant detected using Swin-Base with DetectorRS and RFLA

Fig. 5(e–j). Examples of breast ultrasound detection using architectures of Swin-Base combined with RFLA and DetectoRS.

their corresponding labels, resulting in accelerated network optimization. The proposed model exhibits remarkable performance in contrast to alternative modified VGG16 architectures, achieving an impressive accuracy rate of 93% on both the dataset curated by Yap et al. [44] and the dataset proposed by the authors themselves.

6.2. Object detection approaches

Object detection models typically consist of two primary components: a backbone pretrained on ImageNet [45] and a detector which is responsible for predicting object classes and bounding boxes. These models can be categorized into two kinds based on the characteristics of the detector: one-stage detector and two-stage detector. Notable examples of one-stage detectors include the YOLO family models [41,46–48], SSD [49] and RetinaNet [50]. These models directly predict object locations and confidences based on the entire feature map [51]. In contrast, two-stage detectors utilize a CNN model to classify proposed boxes and refine their coordinates using a sliding window approach [52]. Popular two-stage detectors include RCNN [53], Fast R-CNN [52] and Faster R-CNN [54]. Recently, anchor-free one-stage detectors like CenterNet [55] and CornerNet [56] have emerged, detecting target objects by matching key points instead of using anchor frames, resulting in improved accuracy and speed.

In order to test the effectiveness of our proposed approach, we conducted a comparison with a variety of breast cancer ultrasound detection methods. Lin et al. [26] present a publicly available benchmark dataset for breast lesion detection in ultrasound videos, along

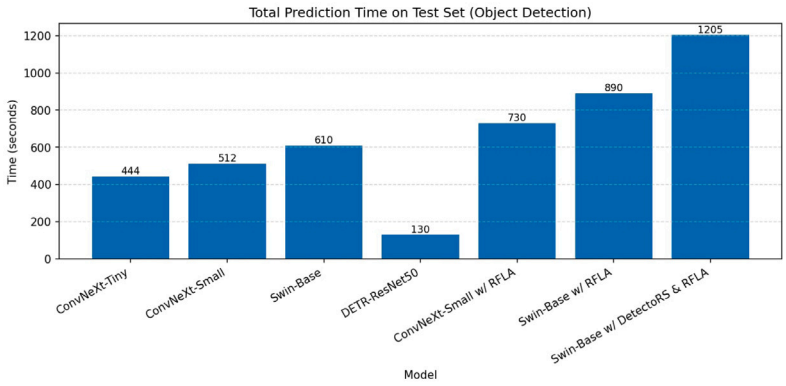
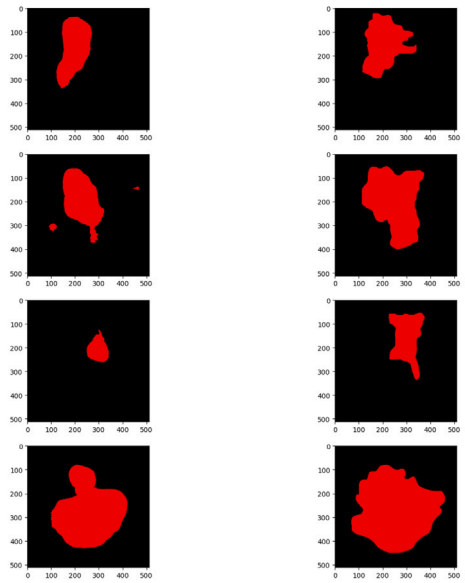
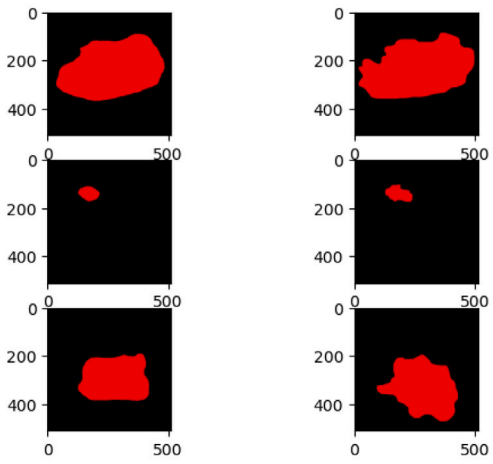


Fig. 6. Prediction time across seven models for object detection task.



(a) Predicted masks (left) compared with ground truths (right) using SegFormer-B0



(b) Predicted masks (left) compared with ground truths (right) using SegFormer-B4

Fig. 7. Examples of breast ultrasound semantic segmentation using SegFormer family.

with annotations. In order to assess the performance of our developed network, we have gathered a dataset comprising 188 ultrasound videos depicting breast lesions. Out of these videos, 113 are malignant and 75 are benign. Overall, the dataset contains a total of 25,272 ultrasound images, with the number of images per video ranging from 28 to 413. In addition, the authors also present CVA-Net, which utilizes a ResNet-50 backbone [39]. The algorithm achieves the highest performance scores in terms of mAP, AP50, and AP75, with values of 36.1, 65.1, and 38.5, respectively.

Mo et al. [57] used YOLO V3 as the detection network and made two modifications to the original YOLO V3. First, they optimized the anchor size by using the K-Means++ algorithm and K-Mediods algorithm. Second, they replaced the residual structure in the original YOLO V3 with a new residual network based on ResNet [39] and DenseNet [58]. Their best mAP using YOLO V3-anchor and V3-res with datasets merged from [27] and [44] are 0.749 and 0.772, respectively.

Wang and Yao [59] proposed a two-stage method: *(i)* Enhancing the contrast of breast ultrasound images using segmentation-based methods, and *(ii)* Using an anchor-free network to detect and classify breast lesions. The method achieved a mean average precision (mAP) score of 0.902 on the datasets similar to Mo et al.'s datasets [57].

6.3. Segmentation methods

Huang et al. [60] introduced a novel convolutional operator called the shape-adaptive convolutional operator, which incorporates a pixel selection mechanism for convolution calculations. By integrating this operator with the original convolutional operator, they are able to extract higher-order convolutional features. Their image segmentation approach showcases exceptional performance, surpassing existing methods and achieving state-of-the-art results. It attains an Intersection over Union (IoU) of 77.90% on the dataset from [44] and 72.12% on the dataset from [27].

Using the BUSI dataset, Al-Dhabyani et al. [27] developed BUSnet, a deep learning model specifically designed to detect breast tumor lesions accurately in ultrasound images [61]. Their approach consists of a two-stage method, incorporating unsupervised region proposal and bounding-box regression algorithms. Furthermore, they introduce a post-processing technique to enhance the accuracy of the detections. The proposed method is evaluated on the benchmark dataset, demonstrating favorable results in terms of effectiveness and accuracy. Specifically, the intersection over union (IoU) scores for benign and malignant lesions are 0.566 and 0.521, respectively.

Lyu et al. [62] developed an improved Pyramid Attention Network called AMS-PAN specifically for breast ultrasound image segmentation. The model incorporates both an Attention mechanism and Multi-Scale features to enhance its performance. To capture diverse information, they have employed depthwise separable convolution on the encoding

side of the model. This allows for the extraction of features at multiple scales and the creation of a feature pyramid. On the decoding side, they have utilized Global Attention Upsample (GAU) feature fusion. Furthermore, they have introduced a Spatial and Channel Attention (SCA) module to emphasize edge texture information and refine the focus of the segmentation process. By conducting thorough experiments on the BUSI dataset curated by [27] and the OASBUD dataset curated by [63], the authors have successfully demonstrated the effectiveness of the proposed method. The results of these experiments have shown impressive segmentation performance, achieving IoU scores of 68.53% and 67.52% on the BUSI and OASBUD datasets, respectively.

To enhance the accuracy of tumor segmentation for different tumor sizes, Shareef et al. [64] introduce a novel deep learning architecture known as Small Tumor-Aware Network (STAN). This architecture effectively combines rich context information and high-resolution image features. The proposed approach is thoroughly evaluated using two public breast ultrasound datasets [27,65], employing seven quantitative metrics. In the task of segmenting small breast tumors, their approach surpasses the performance of existing state-of-the-art methods. Specifically, it achieves Jaccard index scores of 0.847 and 0.695 on the BUSI [27] and Dataset B [65] datasets, respectively.

Huang et al. [66] introduced a fully automatic segmentation algorithm for breast ultrasound (BUS) images. The algorithm is composed of two components: a fuzzy fully convolutional network (FCN) and a conditional random fields (CRF) post-processing step. The FCN is used to extract features from the BUS image. The features are then processed using fuzzy logic, which takes into account the uncertainty of the image data. The CRF post-processing step refines the segmentation result by taking into account the spatial relationships between the different breast anatomy layers. The algorithm is tested on a dataset of 325 breast ultrasound images that the researchers create themselves. The results show that the algorithm achieved state-of-the-art performance compared to existing methods. For the overall segmentation performance across five categories (fat layer, mammary layer, muscle layer, background, and tumor), the algorithm achieves a mean intersection over union (mIoU) of 80.47%.

Recently, Xie et al. [67] introduced PIF-Net, a feature-fusion network for knee magnetic resonance image (MRI) segmentation. The method strengthens feature representations and filters out misleading, look-alike structures with different labels by leveraging multi-level (hierarchical) cues from the input.

7. Conclusions and future work

In this paper, we proposed a workable solution for the recognition of breast lesions using various ultrasound datasets. Our proposed models have been tackled in three domains of computer vision fields, such as classification, object detection, and segmentation. The experimental results demonstrate that the application of deep learning on the considered datasets substantially improves the overall evaluation metrics in comparison to relevant well-established baselines. We see that even the basic network architectures, such as ConvNeXt-Nano, ConvNeXt-Tiny, or ConvNeXt-Small, obtain a decent classification performance. Our approach to object detection presents a greater challenge due to the small size of the objects. While we attained the highest mean Average Precision (mAP = 0.262) and Average Precision for large objects ($AP\_L = 0.265$) using Swin-BASE, its performance in detecting small lesions was not satisfactory. To address this limitation, we experimented with DetectoRS incorporating Gaussian Receptive Field based Label Assignment (RFLA) and achieved an Average Precision for small objects (AP_S) of 0.270. When it came to segmenting small lesions, particularly malignant tumors, we encountered a similar challenge. To address this issue, we managed to achieve the highest mean Intersection over Union (mIoU) of 81.55% by employing SegFormer-B4.

Taken together, recognizing breast cancer in ultrasound images using computer vision poses several challenges, with the detection of small malignant lesions being one of the most significant hurdles. To tackle this particular challenge, we propose the utilization of three distinct models: ConvNeXt-Small for classification, Swin-Base combined with DetectoRS with RFLA for object detection, and SegFormer-B4 for segmentation. Each model is specifically tailored to address its respective task in the detection and analysis of breast cancer.

CRediT authorship contribution statement

Linh T. Duong: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Thu T.H. Doan:** Writing – review & editing, Writing – original draft. **Anh M.T. Bui:** Writing – review & editing, Writing – original draft. **Phuong T. Nguyen:** Writing – review & editing, Writing – original draft, Supervision.

Ethical statement

The research study adheres to the highest standards of ethics. It does not involve the use of human participants and animals. All the analysis are made by the publicly available sources and scientific literature.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We have no funding sources to be acknowledged.

References

[1] Nath MK, Sundararajan K, Mathivanan S, Thandapani B. Analysis of breast cancer classification and segmentation techniques: A comprehensive review. Inform Med Unlocked 2025;56:101642. http://dx.doi.org/10.1016/j.imu.2025.101642, URL https://www.sciencedirect.com/science/article/pii/S2352914825000309.

[2] Khazaei Z, Sohrabivafa M, Momenabadi V, Moayed L, Goodarzi E, et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide prostate cancers and their relationship with the human development index. Adv Hum Biol 2019;9(3):245. http://dx.doi.org/10.4103/2321-8568.262891.

[3] Ghoncheh M, Pournamdar Z, Salehiniya H. Incidence and mortality and epidemiology of breast cancer in the world. Asian Pac J Cancer Prev 2016;17(3):43–6. http://dx.doi.org/10.7314/apjcp.2016.17.s3.43.

[4] Shin HJ, Kim HH, Cha JH. Current status of automated breast ultrasonography. Ultrasonography 2015;34(3):165. http://dx.doi.org/10.14366/usg.15002.

[5] Cheng H, Shan J, Ju W, Guo Y, Zhang L. Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recognit 2010;43(1):299–317. http://dx.doi.org/10.1016/j.patcog.2009.05.012, URL https://www.sciencedirect.com/science/article/pii/S0031320309002027.

[6] Duong LT, Chu CQ, Nguyen PT, Nguyen ST, Tran BQ. Edge detection and graph neural networks to classify mammograms: A case study with a dataset from vietnamese patients. Appl Soft Comput 2023;134:109974. http://dx.doi.org/10.1016/j.asoc.2022.109974, URL https://www.sciencedirect.com/science/article/pii/S1568494622010237.

[7] Trieu PDY, Mello-Thoms C, Brennan PC. Female breast cancer in Vietnam: a comparison across Asian specific regions, vol. 12, Chinese Anti-Cancer Association; p. 238–45. http://dx.doi.org/10.7497/j.issn.2095-3941.2015.0034, arXiv:26487968 URL https://pubmed.ncbi.nlm.nih.gov/26487968, (3).

[8] Qian X, Pei J, Zheng H, Xie X, Yan L, Zhang H, Han C, Gao X, Zhang H, Zheng W, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat Biomed Eng 2021;5(6):522–32. http://dx.doi.org/10.1038/s41551-021-00711-2.

[9] Hasan M, Shawon ARM, Aeyas A, Uddin MA. Cyclic peptides as an inhibitor of metastasis in breast cancer targeting mmp-1: Computational approach. Inform Med Unlocked 2022;35:101128. http://dx.doi.org/10.1016/j.imu.2022.101128, URL https://www.sciencedirect.com/science/article/pii/S2352914822002659.

[10] Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, den Heeten A, Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–12. http://dx.doi.org/10.1016/j.media.2016.07.007, URL https://www.sciencedirect.com/science/article/pii/S1361841516301244.

[11] Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Sci Rep 2018;8(1):4165. http://dx.doi.org/10.1038/s41598-018-22437-z.

[12] Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. Sci Rep 2019;9(1):12495. http://dx.doi.org/10.1038/s41598-019-48995-4.

[13] Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep 2016;6(1):27327. http://dx.doi.org/10.1038/srep27327.

[14] Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning to distinguish recalled but benign mammography images in breast cancer screeningdeep learning in mammography. Clin Cancer Res 2018;24(23):5902–9. http://dx.doi.org/10.1158/1078-0432.ccr-18-1115.

[15] Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology 2019;292(1):60–6. http://dx.doi.org/10.1148/radiol.2019182716.

[16] Lehman CD, Mercaldo S, Lamb LR, King TA, Ellisen LW, Specht M, Tamimi RM. Deep learning vs traditional breast cancer risk models to support risk-based mammography screening. JNCI: J Natl Cancer Inst 2022. http://dx.doi.org/10.1093/jnci/djac142.

[17] Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2022, doi: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01167.

[18] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 10012–22, doi: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986.

[19] Agarap AF. Deep learning using rectified linear units (relu). 2018, CoRR abs/1803.08375 arXiv:1803.08375 URL http://arxiv.org/abs/1803.08375.

[20] Hendrycks D, Gimpel K. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. 2016, CoRR abs/1606.08415 arXiv:1606.08415 URL http://arxiv.org/abs/1606.08415.

[21] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd international conference on international conference on machine learning - volume 37, iCML'15. JMLR.org; 2015, p. 448–56, URL https://proceedings.mlr.press/v37/ioffe15.html.

[22] Ba JL, Kiros JR, Hinton GE. Layer normalization. 2016, http://dx.doi.org/10.48550/arXiv.1607.06450.

[23] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Computer vision – ECCV 2020. Springer-Verlag: 16th European Conference, Glasgow, UK, August (2020) 23–28, Proceedings, Part I; 2020, p. 213–29. http://dx.doi.org/10.1007/978-3-030-58452-8_13.

[24] Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. Segformer: Simple and efficient design for semantic segmentation with transformers. 2021, http://dx.doi.org/10.48550/ARXIV.2105.15203, URL https://arxiv.org/abs/2105.15203.

[25] Wightman R. Pytorch image models. 2025, http://dx.doi.org/10.5281/zenodo.4414861, https://github.com/rwightman/pytorch-image-models. [Retrieved 30 March 2025].

[26] Lin Z, Lin J, Zhu L, Fu H, Qin J, Wang L. A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. Medical image computing and computer assisted intervention – MICCAI 2022. Cham: Springer Nature Switzerland; 2022, p. 614–23. http://dx.doi.org/10.1007/978-3-031-16437-8_59.

[27] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief 2020;28:104863. http://dx.doi.org/10.1016/j.dib.2019.104863, URL https://www.sciencedirect.com/science/article/pii/S2352340919312181.

[28] Rodrigues PS. Breast ultrasound image. 2017, http://dx.doi.org/10.17632/wmy84gzngw.1, Retrieved on 20 January 2025.

[29] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in pytorch. 2017, URL https://api.semanticscholar.org/CorpusID:40027675. [Retrieved 30 November 2024].

[30] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. Pytorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2019, doi: https://dl.acm.org/doi/10.5555/3454287.3455008.

[31] Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J, Zhang Z, Cheng D, Zhu C, Cheng T, Zhao Q, Li B, Lu X, Zhu R, Wu Y, Li, Dai J, Wang J, Shi J, Ouyang W, Loy CC, Lin D. Mmdetection: Open mmlab detection toolbox and benchmark. 2019, ArXiv abs/1906.07155 URL https://api.semanticscholar.org/CorpusID:189927886.

[32] MMSegmentation. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. 2025, https://github.com/open-mmlab/mmsegmentation. [Retrieved 30 March 2025].

[33] Xu C, Wang J, Yang W, Yu H, Yu L, Xia G-S. Rfla: Gaussian receptive field based label assignment for tiny object detection. 2022, http://dx.doi.org/10.48550/ARXIV.2208.08738, URL https://arxiv.org/abs/2208.08738.

[34] Qiao S, Chen L-C, Yuille A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: 2021 IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2021, p. 10208–19. http://dx.doi.org/10.1109/CVPR46437.2021.01008.

[35] Hendria WF, Phan QT, Adzaka F, Jeong C. Combining transformer and cnn for object detection in uav imagery. ICT Express 2023;9(2):258–63. http://dx.doi.org/10.1016/j.icte.2021.12.006, URL https://www.sciencedirect.com/science/article/pii/S2405959521001715.

[36] Khanna P, Sahu M, Kumar Singh B. Improving the classification performance of breast ultrasound image using deep learning and optimization algorithm. In: 2021 IEEE international conference on technology, research, and innovation for betterment of society. TRIBES, 2021, p. 1–6. http://dx.doi.org/10.1109/TRIBES52498.2021.9751677.

[37] Byra M. Breast mass classification with transfer learning based on scaling of deep representations. Biomed Signal Process Control 2021;69:102828. http://dx.doi.org/10.1016/j.bspc.2021.102828, URL https://www.sciencedirect.com/science/article/pii/S1746809421004250.

[38] Kalafi EY, Jodeiri A, Setarehdan SK, Lin NW, Rahmat K, Taib NA, Ganggayah MD, Dhillon SK. Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks. Diagnostics 2021;11(10). http://dx.doi.org/10.3390/diagnostics11101859, URL https://www.mdpi.com/2075-4418/11/10/1859.

[39] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition. CVPR, 2016, p. 770–8. http://dx.doi.org/10.1109/CVPR.2016.90.

[40] Jabeen K, Khan MA, Alhaisoni M, Tariq U, Zhang Y-D, Hamza A, Mickus A, Damaševičius R. Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. Sensors 2022;22(3). http://dx.doi.org/10.3390/s22030807, URL https://www.mdpi.com/1424-8220/22/3/807.

[41] Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv.org; 2018, p. 1–6, arXiv:1804.02767v1 URL http://arxiv.org/abs/1804.02767v1.

[42] Behboodi B, Rasaee H, Tehrani AK, Rivaz H. Deep classification of breast cancer in ultrasound images: more classes, better results with multi-task learning. In: Medical imaging 2021: ultrasonic imaging and tomography. vol. 11602, SPIE; 2021, p. 170–5. http://dx.doi.org/10.1117/12.2581930.

[43] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. CVPR, Los Alamitos, CA, USA: IEEE Computer Society; 2018, p. 4510–20. http://dx.doi.org/10.1109/CVPR.2018.00474, URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00474.

[44] Yap MH, Pons G, Martí J, Ganau S, Sentí s M, Zwiggelaar R, Davison AK, Martí R. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Heal Inform 2018;22(4):1218–26. http://dx.doi.org/10.1109/JBHI.2017.2731873.

[45] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–52. http://dx.doi.org/10.1007/s11263-015-0816-y.

[46] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR. Los Alamitos, CA, USA: IEEE Computer Society; 2016, p. 779–88. http://dx.doi.org/10.1109/CVPR.2016.91, URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.91.

[47] Redmon J, Farhadi A. Yolo9000: Better, faster, stronger. In: CVPR. IEEE Computer Society; 2017, p. 6517–25, URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2017.html#RedmonF17.

[48] Bochkovskiy A, Wang C-Y, Liao H-YM. Yolov4: Optimal speed and accuracy of object detection. 2020, http://dx.doi.org/10.48550/arXiv.2004.10934, arXiv:2004.10934.

[49] Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu C-Y, Berg AC. Ssd: Single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. ECCV (1). Lecture notes in computer science, vol. 9905, Springer; 2016, p. 21–37, URL http://dblp.uni-trier.de/db/conf/eccv/eccv2016-1.html#LiuAESRFB16.

[50] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42(2):318–27. http://dx.doi.org/10.1109/TPAMI.2018.2858826.

[51] Zhang X, Lin X, Zhang Z, Dong L, Sun X, Sun D, Yuan K. Artificial intelligence medical ultrasound equipment: Application of breast lesions detection. Ultrason Imaging 2020;42(4–5):191–202. http://dx.doi.org/10.1177/0161734620928453, arXiv:10.1177/0161734620928453, pMID: 32546066.

[52] Liu S, Huang D, Wang Y. Receptive field block net for accurate and fast object detection. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer vision - ECCV 2018-15th European conference, munich, Germany, September (2018) 8-14, proceedings, part XI. Lecture notes in computer science, vol. 11215, Springer; 2018, p. 404–19. http://dx.doi.org/10.1007/978-3-030-01252-6_24.

[53] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE conference on computer vision and pattern recognition. 2014, p. 580–7. http://dx.doi.org/10.1109/CVPR.2014.81.

[54] Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. In: Proceedings of the 29th international conference on neural information processing systems - volume 1. NIPS'15, Cambridge, MA, USA: MIT Press; 2015, p. 91–9. http://dx.doi.org/10.5555/2969239.2969250.

[55] Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. Centernet: Keypoint triplets for object detection. In: 2019 IEEE/CVF international conference on computer vision. ICCV, 2019, p. 6568–77. http://dx.doi.org/10.1109/ICCV.2019.00667.

[56] Law H, Deng J. Cornernet: Detecting objects as paired keypoints. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer vision - ECCV 2018-15th European conference, munich, Germany, September (2018) 8-14, proceedings, part XIV. Lecture notes in computer science, vol. 11218, Springer; 2018, p. 765–81. http://dx.doi.org/10.1007/978-3-030-01264-9_45.

[57] Mo W, Zhu Y, Wang C. A method for localization and classification of breast ultrasound tumors. In: Tan Y, Shi Y, Tuba M, editors. Advances in swarm intelligence - 11th international conference, ICSI 2020, belgrade, Serbia, July (2020) 14-20, proceedings. Lecture notes in computer science, vol. 12145, Springer; 2020, p. 564–74. http://dx.doi.org/10.1007/978-3-030-53956-6_52.

[58] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society; 2017, p. 2261–9. http://dx.doi.org/10.1109/CVPR.2017.243.

[59] Wang Y, Yao Y. Breast lesion detection using an anchor-free network from ultrasound images with segmentation-based enhancement. Sci Rep 2022;12(1):14720. http://dx.doi.org/10.1038/s41598-022-18747-y.

[60] Huang K, Zhang Y, Cheng HD, Xing P. Shape-adaptive convolutional operator for breast ultrasound image segmentation. In: 2021 IEEE international conference on multimedia and expo. ICME, 2021, p. 1–6. http://dx.doi.org/10.1109/ICME51207.2021.9428287.

[61] Li Y, Gu H, Wang H, Qin P, Wang J. Busnet: A deep learning model of breast tumor lesion detection for ultrasound images. Front Oncol 2022;12. http://dx.doi.org/10.3389/fonc.2022.848271, URL https://www.frontiersin.org/articles/10.3389/fonc.2022.848271.

[62] Lyu Y, Xu Y, Jiang X, Liu J, Zhao X, Zhu X. Ams-pan: Breast ultrasound image segmentation model combining attention mechanism and multi-scale features. Biomed Signal Process Control 2023;81:104425, URL https://www.sciencedirect.com/science/article/pii/S1746809422008795.

[63] Piotrzkowska-Wróblewska H, Dobruch-Sobczak K, Byra M, Nowicki A. Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. Med Phys 2017;44(11):6105–9. http://dx.doi.org/10.5281/zenodo.603138.

[64] Shareef B, Xian M, Vakanski A. Stan: Small tumor-aware network for breast ultrasound image segmentation. In: 2020 IEEE 17th international symposium on biomedical imaging. ISBI, 2020, p. 1–5. http://dx.doi.org/10.1109/ISBI45749.2020.9098691.

[65] Xian M, Zhang Y, Cheng H-D, Xu F, Huang K, Zhang B, Ding J, Ning C, Wang Y. A benchmark for breast ultrasound image segmentation (BUSIS). Infin Study 2018. http://dx.doi.org/10.3390/healthcare10040729.

[66] Huang K, Zhang Y, Cheng H, Xing P, Zhang B. Semantic segmentation of breast ultrasound image with fuzzy deep learning network and breast anatomy constraints. Neurocomputing 2021;450:319–35. http://dx.doi.org/10.1016/j.neucom.2021.04.012, URL https://www.sciencedirect.com/science/article/pii/S0925231221005300.

[67] Xie X, Xie L, Pan X, Shao F, Zhao W, An J. Pif-net: A parallel interweave fusion network for knee joint segmentation. Biomed Signal Process Control 2025;109:107967. http://dx.doi.org/10.1016/j.bspc.2025.107967, URL https://www.sciencedirect.com/science/article/pii/S1746809425004781.