A Benchmark for Breast Ultrasound Image Classification

Bryar Shareef[1], Min Xian[1], Shoukun Sun[1], Aleksandar Vakanski[1], Jianrui Ding[2], Chunping Ning[3], Heng-Da Cheng[4]

[1] Department of Computer Science, University of Idaho, Idaho Falls, ID 83401, U.S.A.

[2]School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China.

[3]Department of Ultrasound, Affiliated Hospital of Medical College Qingdao University, Qingdao, China.

[4] Department of Computer Science, Utah State University, Logan, UT 84322 U.S.A.

Corresponding author: mxian@uidaho.edu

A Benchmark for Breast Ultrasound Image Classification

Bryar Shareef[1], Min Xian[1], Shoukun Sun[1], Aleksandar Vakanski[1], Jianrui Ding[2], Chunping Ning[3], Heng-Da Cheng[4]

[1] Department of Computer Science, University of Idaho, Idaho Falls, ID 83401, U.S.A.

[2]School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China.

[3]Department of Ultrasound, Affiliated Hospital of Medical College Qingdao University, Qingdao, China.

[4] Department of Computer Science, Utah State University, Logan, UT 84322 U.S.A.

Abstract

Classification of breast ultrasound (BUS) images is an essential yet challenging task in computer-aided diagnosis systems. Recently, deep learning-based approaches for BUS image classification have demonstrated state-of-the-art performance; however, it is difficult to reproduce their results and identify the most useful strategies due to the lack of public datasets and method implementations, and inconsistencies in the reported evaluation metrics. Therefore, there is a pressing need to develop a benchmark, to objectively compare current approaches and gain insights on techniques that improve the generalization of BUS image classification. In this work, we build a benchmark for BUS image classification that consists of a large public dataset with 3,641 B-mode BUS images, provide open-source code of state-of-the-art approaches, and identify the best strategies for deep learning-based BUS classification. Moreover, we propose a novel multitask learning approach which incorporates a small-tumor aware network as the backbone network, and consists of one primary task (tumor classification) and a secondary task (tumor segmentation). We evaluate the proposed approach and 10 deep learning-based approaches using seven quantitative metrics on the benchmark dataset. Extensive experiments demonstrate that the proposed approach achieves state-of-the-art performance with high sensitivity and specificity of 90.4% and 89.8%, respectively.

Keywords: breast ultrasound benchmark; breast cancer detection; deep learning; computer-aided diagnosis

1. Introduction

Breast cancer has become one of the most common cancers worldwide, accounting approximately for 12% of all new cancer cases [1]. In the U.S., it is estimated that breast cancer affected 30% of all new female cancer cases in 2021 [1]. Early detection of breast cancer can significantly reduce mortality and expand treatment options. Among the different imaging modalities, mammography and ultrasound are the two most popular imaging tools for detecting breast abnormality. However, mammography is less commonly implemented in most low- and middle-income countries, because of the high costs of the required infrastructure [2]. Furthermore, mammography produces high false-positive rates in women with dense breasts, which leads to anxiety and additional examination steps, such as biopsy [3]. Rebolj et al. [4] reported that ultrasound detected approximately 40% more cancer cases than mammography in women with dense breasts. According to [5-9], women with dense breasts have a four to six times greater risk of breast cancer than those with fatty breast tissue. Asian women of age < 45 have 1.2 more dense breasts than white women of that age, and the ratio increases to 1.6 for age 65 and older. In contrast, black women have 1.7 more dense breasts than white women for age 65 and younger, while black, Hispanic, and white women have a similar breast density for ages>65.

BUS image processing is challenging due to the presence of speckle noise, low contrast, weak boundary, and artifacts [10]. Therefore, analyzing ultrasound images requires extensive experience and training. To alleviate this challenge, computer-aided diagnosis (CAD) systems have been developed to assist radiologists with breast tumor diagnosis. The idea of CAD was first introduced in the 1960s [11]. These systems can reduce operator dependency and identify breast tumors/cancers more accurately [12]. CADs can be broadly classified into conventional and deep learning-based systems [13]. The conventional BUS CAD systems typically comprise four modules: image preprocessing, tumor segmentation, feature extraction and selection, and tumor classification [12] (see Fig. 1(a)). In deep learning-based CAD systems, the modules of preprocessing [11,12] and segmentation [16,17] become optional (see Fig. 1(b)). Automatic

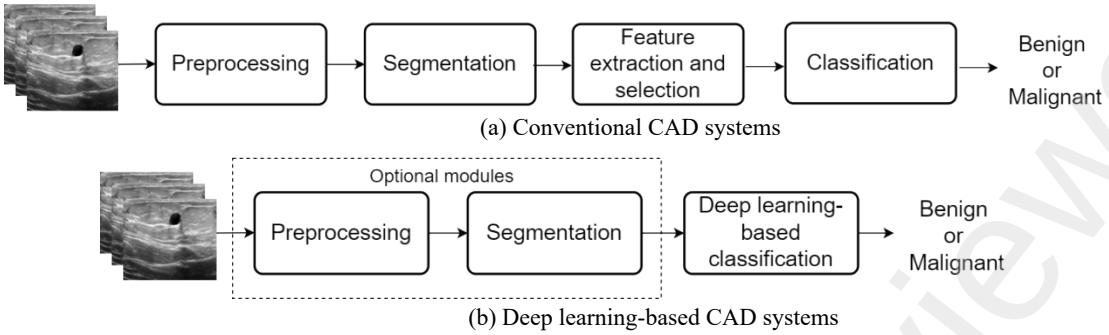

Fig. 1. Key modules in conventional and deep learning-based CAD systems.

feature learning without human intervention is a substantial advantage of deep learning-based approaches over conventional approaches [13]. On the other hand, conventional approaches rely on radiologists' knowledge to extract and select meaningful features [18].

Given recent advancements in deep learning approaches for medical image applications, prior work demonstrated the effectiveness of deep learning to classify breast tumors in ultrasound images (see Table 1). However, due to the lack of large, publicly available, high-quality BUS datasets, and unified quantitative metrics, a fair evaluation of the current approaches and strategies is impossible. Furthermore, most existing deep learning architectures for BUS image classification are simply adopted from general-purpose image classification tasks, and there is limited research on identifying the best architectures and strategies of deep learning for BUS image classification. In this paper, the focus is on benchmarking deep learning-based CAD systems for BUS image classification. Refer to [10] for more details on a BUS benchmark for breast tumor segmentation.

The paper is organized as follows. Section 2 discusses the fundamentals of BUS image classification using deep learning; Section 3 describes the benchmark setup. Section 4 illustrates the proposed approach; Section 5 presents comprehensive experimental results. Finally, Sections 6 and 7 provide a discussion and conclusion, respectively.

Table 1. Deep learning approaches for BUS image classification.

References	Approaches	Year	Dataset/Availability	Performance	Pretrained dataset
Huynh, et al. [20]	Feature extractor (AlexNet) + SVM	2016	1,125 cases/private	AUC: 88%	ImageNet
Shia, et al. [25]	Fine-tuned (ResNet101) + SVM	2021	2,099/private	Sen: 94%, Spec: 93%, AUC: 94%	ImageNet
Liang, et al. [21]	Feature extractor (Mask-R-CNN)	2019	150 cases/private 163 cases/public	Acc : 80%, TPR: 63%, TNR: 87%	Coco datasets
Liao et al. [31]	Fine-tuned (VGG19, ResNet50, DenseNet121, Inceptions V3) + Elastography images + B-mode images	2020	256/private	AUC:98%, Acc:93%, Sen:91%, Spec: 95%, F_1: 93%	ImageNet
Fei et al. [32]	Designed DL network (SVM + Elastography) + Transfer learning	2020	265/private	Acc:87%, Sen: 86%, Spec: 87%, Youden index (YI): 73%	--
Yap et al. [56]	Fine-tuned (FCN-AlexNet)	2018	306/private 163/public	Sen (Benign:83%, Malignant: 57%)	ImageNet
Zhang, et al. [26]	Fine-tuned (VGG16, ResNet50, InceptionV3, VGG19)	2020	6,007/private	Sen: 85%, AUC: 91%, PPV:64%, Acc:83%, NPV: 93.7%, Spec: 81.5%	ImageNet
Hijab et al. [23]	Fine-tuned (VGG16) + ROIs	2019	1,300/private	Acc: 97%, AUC: 98%	ImageNet
Cao et al. [24]	Fine-tuned (4 ROIs on five networks)	2019	1,041/private	APR: 97%, ARR:67%, F_1:79%, Acc: 87.5%	ImageNet
Xie et al. [27]	Network design (Dual-sampling (2 Encoders) network)	2020	1,272/private 163/public	Acc: 92%, Sen: 95%, Spec: 89%, PPV: 88%, NPV: 95%, AUC: 94%	ImageNet
Xing et al. [29]	Prior knowledge (BI-RADS + CNN)	2020	Training: 9,373/private Tested: 810/public	AUC: 91%, Acc: 87%, Sen: 82%, Spec: 89%, Precision: 80%	ImageNet
Zhuang et al. [30]	Prior knowledge (hand crafted features +SVM +DL)	2021	1,682/public	Acc: 93%, Precision: 91%, Sen: 95%, F_1: 93%, Spec: 91%	ImageNet
Han et al. [28]	Adopting modified network (GoogleNet) +ROIs	2017	7,408/private	AUC: 96%, Acc: 91%, Sen:84%, Spec: 96%	ImageNet
Al-Dhabyani et al. [14]	Data augmentation (GAN to produce data)	2019	780/public	Acc: 99%	ImageNet
Tanaka et al. [57]	Ensemble Learning (VGG19 +ResNet152)	2019	1,543/private	Acc: 86%, Precision: 85%, Sen: 89%, F_1 : 87%, Spec: 83%, AUC:94%	--
Byra et al. [15]	Preprocessing (Input Channel)	2019	Training:882/ private Tested: 163/public	AUC: 94%, Acc: 89%, Sen: 85%, Spec: 90%	ImageNet
Zhuang et al. [33]	Preprocessing (Decomposition of BUS images)	2020	2,280/public	AUC: 98%, Acc: 92%, Sen: 98%, Spec: 86%, F_1: 93%	ImageNet
Zhang et al. [35]	Multitask learning + attention mechanism	2021	647/public	Acc:94%, Sen: 89%, Spec: 96%, F_1:93%	--
Moon et al. [58]	Ensemble learning (BUS + tumor masks + segmented tumor + fused images)	2020	647/public 1,687/private	AUC: 95%, Acc:91%, Sen: 97%, Spec: 95%, F_1: 83%, Precision: 73%	--

Acc: Accuracy, AUC: area under curve, Sen: sensitivity, Spec: Specificity, TNR: true negative rate, TPR: true positive rate, PPV: positive predictive value, NPV: negative predictive value, APR: average precision rate, ARR: average recall rate,

2. Fundamentals of BUS image classification using deep learning

2.1 Transfer learning

Deep learning typically requires large and high-quality labeled data. However, many medical applications have scarce data due to expensive data collection, high labeling costs, and privacy issues. To address these issues, many approaches have adopted the transfer learning strategy. In transfer learning approaches, a deep learning network, which is previously pretrained for another task on a large-scale dataset is employed for BUS classification. For example, the ImageNet [19] dataset is widely used by deep learning approaches for learning feature representations. The pretrained model can be used as 1) a fixed feature extractor or 2) an initial model for fine-tuning.

For the fixed feature extractor, the pretrained layers are kept unchanged, and the prediction layers are trained based on the target task. Huynh et al. [20] employed a pretrained model (AlexNet) as a feature extractor and combined it with a support vector machine (SVM) algorithm to classify BUS images by using 1,125 whole images and 2,393 regions of interest (ROIs). Liang et al. [21] proposed using Mask R-CNN to segment and classify breast tumors simultaneously, where a ResNet50 pretrained on the COCO [22] dataset was used as a backbone to extract features.

In models for fine-tuning, the whole network including the pretrained layers and the prediction layers is retrained using new data. The fine-tuning approach uses the pretrained weights to initialize the network, and tune it to a target task. Hijab et al. [23] adopted transfer learning to train VGG16 for classifying BUS images. The authors studied three different training techniques, and the results demonstrated that the fine-tuned network outperformed both training from scratch and transfer learning without fine-tuning. Cao et al. [24] studied breast tumor detection and classification using five models with and without transfer learning. Moreover, [25] used a pretrained deep residual network as a feature extractor and a support vector machine (SVM) algorithm to classify BUS images, and their classification performance on 2,099 BUS images outperformed physicians. Zhang et al. [26] used a balanced training set and compared four

pretrained classifiers (InceptionV3, VGG16, ResNet50, and VGG19), and pretrained InceptionV3 with fine-tuning outperformed all other three models.

2.2 Network architectures

Developing network architectures based on domain knowledge can enhance the generalizability of deep learning-based approaches. Xie et al. [27] proposed the DSCNN to combine convolutional and residual layers for BUS image classification. DSCNN outperformed pretrained and fine-tuned AlexNet, ResNet18, VGG16, GoogleNet, and EfficientNet, and the three experienced radiologists. Han et al. [28] modified GoogleNet with different regions of interest (ROIs) which accepted single-channel images and removed two auxiliary classification branches. The proposed approach achieved a sensitivity of 86% and an AUC of 90% on a private dataset.

2.3 Incorporating prior knowledge

Xing et al. [29] integrated BI-RADS information into a three-layer residual network. The proposed approach showed promising results and outperformed all other transfer learning and non-transfer learning approaches on two public datasets and one private dataset. Zhuang et al. [30] extracted four characteristic semantic features (i.e., orientation, characteristics of posterior shadowing region, shape complexity, and edge indistinctness) and combined them with computational features learned from VGG16. The proposed approach outperformed the general-purpose-designed deep learning approaches. Liao et al. [31] extracted computational features using two VGG19 models from B-mode BUS images and strain elastography images, respectively; and all features are concatenated and input into a 3-layer network to conduct classification. The results showed that the proposed approach can achieve better sensitivity and specificity compared to deep learning approaches trained solely on B-mode images. Similarly, [32] transferred knowledge from elastography ultrasound through transfer learning to improve the diagnostic accuracy of breast cancer.

2.4 Preprocessing

The image quality and size of a BUS dataset have a significant impact on deep learning models. Researchers have employed a variety of preprocessing techniques to enlarge, standardize, and enhance datasets. Al-Dhabyani et al. [14] implemented a new augmentation approach by combing generative adversarial networks (GANs) with traditional augmentation methods; and the classification accuracy of VGG16, Inception, ResNet, and NasNet was improved by 16%, 17%, 16%, and 15%, respectively. Byra et al. [15] introduced a matching layer to rescale grayscale BUS images to RGB images. The results showed that this technique improved the performance of a pretrained VGG19 network. Zhuang et al. [33] used fuzzy enhancement, bilateral filtering, and image morphology operation to produce a set of decomposed images which were combined to feature maps using three deep learning models. The approach showed promising results, with the specificity and sensitivity reaching 98% and 94%, respectively.

2.5 Multitask learning

Multitask learning has been proved to be an effective approach to improve the generalizability of deep learning approaches by learning shared representations from multiple tasks. Vakanski et al. [34] implemented a deep multitask network that comprised both tumor segmentation and classification subnetworks, and the performance of tumor classification was significantly improved by learning representations focused on tumor regions. Zhang et al. [35] employed soft and hard attention mechanisms to perform tumor classification and segmentation simultaneously; and the classification accuracy increased by 2.45% compared with the single task model. Shi et al. [36] proposed the EMT-NET, a light-weighted multitask learning approach for both breast tumor classification and segmentation to replace the single task MobileNet; and its sensitivity increased by 18.81%.

2.6 Challenges

Conclusively, despite the potential of deep learning approaches for accurately classifying BUS images, considerable challenges still need to be addressed: 1) most deep learning approaches require large and high-

quality labeled datasets, but most publicly available BUS datasets are small. It is time-consuming and expensive to collect a large BUS dataset. 2) The end-to-end learning scheme of deep learning approaches makes BUS image classification a black box, which leads to poor explainability. 3) Existing deep learning approaches have poor robustness and are vulnerable to adversarial attacks. 4) Most deep learning approaches are computationally intensive, which makes it impossible to deploy them to devices with limited resources. To the best of our knowledge, there is an absence of benchmarking studies focusing on deep learning approaches in classifying breast ultrasound images. Therefore, we are introducing a BUS benchmark to identify the most useful strategies for classifying breast tumors using a combined dataset of 3,641 BUS images.

3. Benchmark setup

This section provides a detailed description of the BUS image datasets, deep learning approaches, experimental setup, and evaluation metrics.

3.1 BUS image dataset

Existing public BUS datasets are small. We prepared a large and diverse BUS dataset from five sources, HMSS [37], BUSI [38], BUSIS [10], Thammasat [39], and Dataset B [40]. It contains a total of 3,641 B-model BUS images, of which 1,854 contain benign tumors and 1,763 have malignant tumors. Detailed information on the five datasets is shown in Table 2. We develop a set of scripts to prepare the images which are publicly available at http://busbench.midalab.net. Note that we do not own the images, and researchers need to obtain permissions to use the datasets from the original authors.

A total of 2,006 BUS images are from the HMSS [37] dataset, of which 882 images have benign tumors and 1,100 have malignant tumors. HMSS was collected by Dr. Geertsma, an experienced radiologist at Gelederse Vallei hospital in Netherland, in a collaboration with Hitachi Medical Systems Europe. BUSI [38] dataset was collected from Baheya Hospital for Early Detection & Treatment of Women's Cancer (Cairo, Egypt) using LOGIQ E9 ultrasound system and LOGIQ E9 Agile ultrasound system with the ML6-

Table 2. Five public BUS datasets.

BUS dataset	BUS images	Class distribution	Ground truth availability	Country
HMSS [37]	2,006	B: 846, M: 1,160	Classification: Yes Segmentation: No	Netherlands
BUSI [38]	647	B: 437, M: 210	Classification: Yes Segmentation: Yes	Egypt
BUSIS [10]	562	B: 306, M: 256	Classification: Yes Segmentation: Yes	China
Thammasat [39]	263	B:120, M: 143	Classification: Yes Segmentation: No	Thailand
Dataset B [40]	163	B: 109, M: 54	Classification: Yes Segmentation: Yes	Spain
Total # of images	3,641	Total # of Benign (B): 1,823 (50.06%) Total # of Malignant (M): 1,818 (49.94%)		

15-D Matrix linear probe transducers. The dataset has a total of 780 images, of which 133 are normal, 437 are benign, and 210 are malignant. It was collected from 600 women patients aged between 25 and 75 years old. We excluded the normal cases, resulting in a total of 647 BUS images. BUSIS [10] dataset was collected from the Second Affiliated Hospital of Harbin Medical University, the Affiliated Hospital of Qingdao University, and the Second Hospital of Hebei Medical University using the GE VIVID 7, LOGIQ E9, Hitachi EUB-6500, Philips iU22, and Siemens ACUSON S2000 systems. It contains 562 images, of which there are 306 benign and 256 malignant images. Thammasat dataset [39] was collected by the Biomedical Engineering Unit at the Thammasat University Hospital, and Philips iU22 ultrasound workstation was used. We get a total number of 263 (120 benign and 143 malignant) BUS images from the Thammasat dataset. Dataset B [40] consists of 163 breast ultrasound images (53 malignant and 110 benign), provided by the UDIAT Diagnostic Centre of the Parc Taul´ı Corporation, Sabadell (Spain). The images were collected using the Siemens ACUSON Sequoia C512 system with a 17L5 linear array transducer (8.5 MHz). Refer to the original publications of the datasets for more details.

Because most deep learning approaches require square images as input, all BUS images in the benchmark dataset are zero-padded and reshaped to form square images without distortions. Note that directly reshaping an original BUS image to a square shape will result in morphologic changes in breast tumors and their surrounding tissues. Refer to our scripts for preparing the benchmark dataset.

Table 3. The sizes of the selected classifiers.

	List of generic deep learning classifiers		
	Classifiers	Number of parameters (million)	Size of trained models (megabytes)
1	MobileNet	4.2	29 MB
2	EfficientNetB0	5.3	37 MB
3	DenseNet121	8	59 MB
4	Xception	22.9	168 MB
5	InceptionV3	23	176 MB
6	ResNet50	25	189 MB
7	VGG16	138.3	172 MB
	List of BUS-specific deep learning classifiers		
1	Shi, et al. [36]	5.1	60 MB
2	Zhang, et al. [35]	8.2	130 MB
3	Vakanski, et al. [34]	27.3	312.6 MB

3.2 Deep learning approaches and setup

In this study, we evaluate seven generic widely used deep learning-based classifiers [41-47] and three recently published state-of-the-art approaches [34-36] for BUS image classification (see Table 3). The generic approaches include MobileNet V1 [41], EfficientNet [42], DenseNet121 [43], ResNet50 [44], VGG16 [45], Xception [46], and InceptionV3 [47]. These classifiers are among the most commonly used architectures in medical image applications, thus, providing new insights into their performance will benefit the development of CAD systems and the research community. In addition, the approaches range from lightweight to heavyweight models, and evaluating them could help build applications with hardware limitations. The 5-fold cross-validation is utilized to assess the performance of all approaches. The maximum number of training epochs is set to 50, and the batch size is 32. In addition, a validation set that comprises 20% of the training set is used, and all BUS images of the benchmark dataset are resized to the original classifier's input size. In the benchmark dataset, multiple images may come from one patient/case. To prevent data leakage and bias, we split the train and test set based on the cases, i.e., all images from one case are assigned to only one of the training, validation, and test sets.

The approaches are implemented in Keras and TensorFlow using Python (version 3.7) programming language. All experiments were performed on a GPU server with seven NVIDIA Quadro RTX 8000 GPUs, two Intel Xeon Silver 4210R CPUs (2.40GHz), and 512 GB of RAM.

3.3 Evaluation metrics

To evaluate the performance of the classifiers, we use the following quantitative metrics: accuracy (Acc), sensitivity (Sens), specificity (Spec), F_1 score, false positive rate (FPR), false negative rate (FNR), and Area Under the Receiver Operating Characteristic Curve (AUC).

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{1}$$

$$\mathrm{Sens} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{Spec} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{3}$$

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + (\mathrm{FP} + \mathrm{FN})} \tag{4}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{5}$$

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} \tag{6}$$

In Eqs. (1-6), TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

3.4 Loss functions

We explore three different loss functions to improve the overall performance and identify the best strategy that can better balance the sensitivity and specificity for breast cancer detection. The adopted loss functions include binary cross-entropy loss, focal loss [48], and weighted cross-entropy loss. The binary cross-entropy is widely employed in binary classification, and it is defined by

$$L_{BCE} = -\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\left[(t_i \cdot log\,(p_i) + (1 - t_i) \cdot log\,(1 - p_i)\right] \quad (7)$$

where N denotes the number of image samples; t_i is the target label of the *i*th training sample; p_i denotes the prediction. Cross-entropy loss calculates the difference between two probability distributions and all classes are treated equally. To reduce the risk of false negatives, we employed the weighted cross-entropy function. The normal weighted cross-entropy is given by

$$L_{WBCE} = -\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\left[(w_z \cdot t_i \cdot log\,(p_i) + (1 - t_i) \cdot log\,(1 - p_i)\right] \quad (8)$$

where w_z is the weight parameter that penalizes the false-negative predictions and could also mitigate the issue of imbalanced classes. To avoid overflow issues and produce stable results, we utilized a numerically stable weighted cross-entropy which was implemented in [36] and is defined by

$$L_{NS-WBCE} = -\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\left((1 - t_i) \cdot l_i + s_i \cdot \log\left(1 + e^{-l_i}\right)\right) \quad (9)$$

where l_i is the logits of the predicted probability p_i, and s_i is from the positive weight coefficient. They defined as $l_i = \log\left(\frac{p_i}{1-p_i}\right)$ and $s_i = 1 + t_i \cdot (w_z - 1)$.

Furthermore, to focus more on difficult predictions, we utilized the focal loss function [50]. In the focal loss, a factor $(1 - p_i)^{\gamma}$ is added to the cross-entropy loss, where γ is a focusing parameter that makes the model focus on hard samples. The focal loss is defined by

$$L_{Focal} = -\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\left[(\alpha \cdot t_i \cdot (1 - p_i)^{\gamma} \cdot \log\,(p_i) + (1 - t_i) \cdot (1 - \alpha) \cdot p_i \cdot \log\,(1 - p_i)\right] \quad (10)$$

where α is a weighting factor, and takes values from [0, 1]. We use nine combination of focal loss weights (γ = {2, 3, 4}, and α = {0.25, 0.50, 0.8}) and five weights for L_{WBC} (1, 2, 3, 4, and 5).

4. The proposed method

Multitask learning (joint BUS segmentation and classification) can significantly improve the generalization ability of deep learning approaches trained using datasets with limited sizes. The performance of the primary task could be improved using better representations regularized by a secondary task. In BUS images, tumor categories are determined by features inside or around a tumor; if we could regularize a deep neural network to learn representations of tumor regions, a more accurate and robust model could be trained. Inspired by this, we propose a new deep multitask network, namely MT-ESTAN, which consists of both tumor segmentation and classification tasks. The network architecture is shown in Fig. 2.

In our previous work [16, 17], small-tumor aware networks were proposed to accurately segment tumors with different sizes. [16] used row-column-wise kernels to extract and fuse BUS context information at different scales. It consists of two parallel encoder branches: the enhanced small-tumor aware network (ESTAN) and basic encoders. In this work, we use the network in [16] as the backbone of MT-ESTAN to ensure sensitivity to tumors with different sizes; and ResNet50 is used as the building blocks of the basic encoder. Refer to [16] for the implementation details of ESTAN. There are several major differences between the proposed MT-ESTAN and our ESTAN in [16]: 1) MT-ESTAN performs tumor classification and segmentation simultaneously, and tumor classification is the primary task. ESTAN [16] only has a tumor segmentation task; 2) the loss function of MT-ESTAN is a balanced combination between $L_{NS-WBCE}$ and Dice loss, while ESTAN only has the Dice loss; and 3) the basic encoder was pretrained on ImageNet in MT-ESTAN, but trained from scratch in ESTAN.

Segmentation Task. The segmentation task is supplementary to the classification task. The segmentation branch comprises four blocks, and each has an upsampling layer and three consecutive convolution kernels (see Figs. 2(a) and (c)). Each block receives two skip connections from blocks in the two encoders, i.e. a skip connection from the basic encoder and another from the ESTAN encoder.

Classification Task. The primary task of the proposed MT-ESTAN is to classify BUS tumors into benign and malignant. The classification branch receives input from the combined basic and ESTAN encoders. It consists of a Global Average Pooling (GAP) layer followed by two dense layers using ReLU activation with 512, and 128 nodes, respectively. A dropout layer with a rate of (50%) is added after the first dense layer. The final prediction consists of a single node employing a sigmoid activation function.

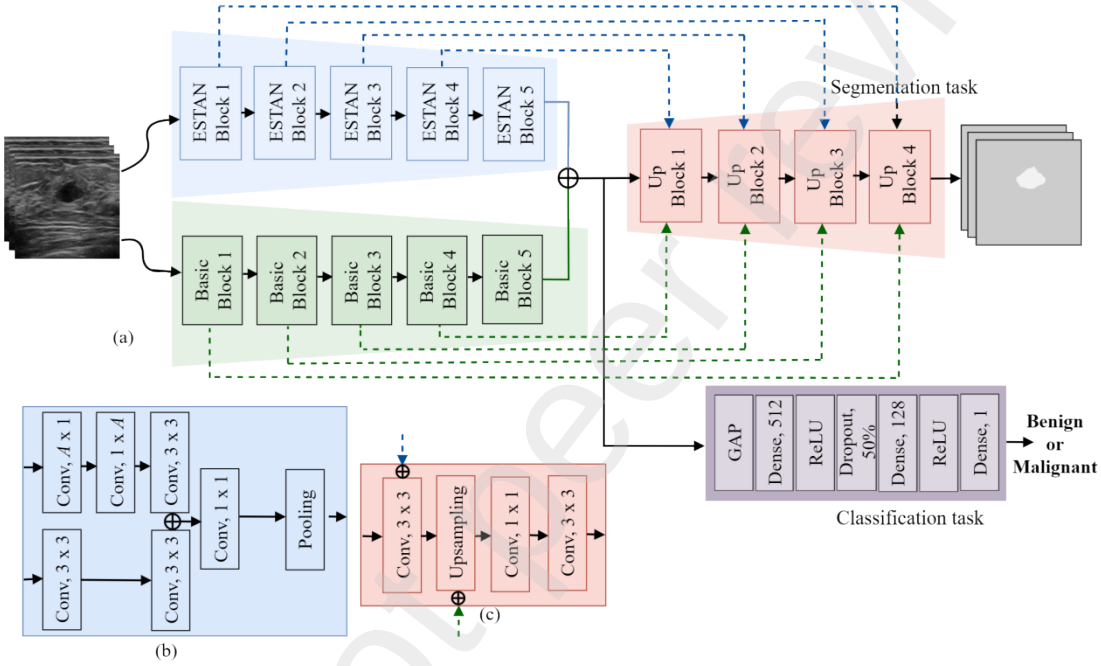

Fig. 2. MT-ESTAN architecture. (a) Overall architecture; (b) the ESTAN block; and (c) the upsampling (Up) block. ⊕ denotes the concatenation operator, and A denotes kernel size.

Loss function. In disease diagnosis, the models that produce higher sensitivity are more vital than that vice versa. We utilize the weighted cross-entropy loss function for the classification task to perform a trade-off between sensitivity and specificity with minimum sacrifice of overall accuracy. A numerically stable weighted cross-entropy from [36] is adopted and is defined in Eq. (9). The final multitask loss (L_{mtl}) function is defined by

$$L_{mtl} = w \cdot L_{NS-WBCE} + L_{Dice} \tag{11}$$

where the weight (w) of the classification task is set to 3, and the positive weight of $L_{NS-WBCE}$ is set to 3. In addition, the best model with the minimum validation loss will be saved during training.

The proposed approach and [34-36] share the same two tasks. However, two major differences exist. 1) [34], [35], and [36] used U-Net, DenseNet, and MobileNet, respectively, as the backbone network. The proposed multitask network applies the ESTAN as the backbone, and is more robust to tumors of different sizes. 2) [34] and [35] used the cross-entropy function as the loss of the classification loss, and have no control on the balance of sensitivity and specificity. For example, [35] obtained high specificity but relatively low sensitivity. The proposed network utilizes the numerically-stable weighted cross-entropy loss that enables the flexibility to balance sensitivity and specificity.

5. Experimental results

In this section, we evaluate the proposed approach and 10 deep learning-based approaches for BUS image classification using the proposed benchmark dataset. The five most useful strategies in deep learning are validated by experiments in Sections 5.1 and 5.2.1; and the effectiveness of the proposed approach is validated and discussed in Section 5.2.2.

5.1 Evaluate useful strategies in deep neural networks for BUS image classification

Training from scratch versus transfer learning. In the transfer learning setup, all classifiers are pretrained on ImageNet, and the last prediction layer is replaced with two dense layers with 512 and 64 units, respectively. ReLU is used as the activation. All model parameters are trainable in the fine-tuning stage. For training from scratch, all seven models are trained from scratch using BUS images. Additionally, all experiments were conducted without using regularization, augmentation, and postprocessing techniques.

The results presented in Table 4 show that all seven models with transfer learning outperform those with training from scratch. It is worth noting that transfer learning significantly enhances the performance of the less complex classifiers with small model sizes. The reason could be that small models are prone to underfit when trained from scratch on a limited number of images. For example, the EfficientNetB0 model is a lightweight classifier with only 5.3 million parameters, and its accuracy, F_1 score, and AUC improved

by 19.3%, 12.1%, and 19.5%, respectively. On the other hand, VGG16 is a heavyweight classifier with 138 million parameters, and its accuracy, F_1 score, and AUC improved by 10.1%, 7.8%, and 10.2%, respectively. The pretrained VGG16 classifier outperformed all other classifiers by achieving the best F_1 score and AUC. Because transfer learning improves the overall classification performance, it is used in the remaining sections.

Table 4. Training from Scratch (S) vs. Transfer learning (TL).

Classifiers	Accuracy (%) ↑		Sensitivity (%) ↑		Specificity (%) ↑		F_1 ↑		AUC (%) ↑		FPR (%) ↓		FNR (%) ↓	
	S	TL	S	TL	S	TL	S	TL	S	TL	S	TL	S	TL
DenseNet121	64.8	73.3	69.3	70.9	59.8	75.9	0.66	0.72	64.5	73.4	40.2	24.1	30.7	29.1
InceptionV3	64.5	71.6	69.0	62.8	59.6	**80.5**	0.66	0.69	64.3	71.7	40.4	**19.5**	31.0	37.2
MobileNet	61.7	75.3	74.5	76.9	49.1	74.2	0.66	**0.76**	61.8	75.5	50.9	25.8	25.5	23.1
ResNet50	62.0	70.3	74.2	**79.1**	50.6	61.8	0.66	0.73	62.4	70.4	49.4	38.2	25.8	**20.9**
VGG16	68.9	**76.7**	75.7	75.8	62.2	77.8	0.70	**0.76**	68.9	**76.8**	37.8	22.2	24.3	24.2
Xception	63.1	72.7	75.4	73.0	52.1	72.6	0.67	0.73	63.8	72.8	47.9	27.4	24.6	27.0
EfficientNetB0	59.7	74.0	75.6	73.0	43.8	75.4	0.65	0.74	59.7	74.2	56.2	24.6	24.4	27.0

Table 5. Augmentation (Aug.) vs. no augmentation (No Aug.).

Classifiers	Accuracy (%) ↑		Sensitivity (%) ↑		Specificity (%) ↑		F_1 ↑		AUC (%) ↑		FPR (%) ↓		FNR (%) ↓	
	No Aug.	Aug.	No Aug.	Aug.	No Aug.	Aug.	No Aug.	Aug.	No Aug.	Aug.	No Aug.	Aug.	No Aug.	Aug.
DenseNet121	73.3	76.9	70.9	72.2	75.9	**81.9**	0.72	0.76	73.4	77.0	24.1	**18.1**	29.1	27.8
InceptionV3	71.6	75.7	62.8	73.4	80.5	78.4	0.69	0.75	71.7	75.9	19.5	21.6	37.2	26.6
MobileNet	75.3	**77.2**	76.9	75.1	74.2	79.6	0.76	**0.77**	75.5	**77.4**	25.8	20.4	23.1	24.9
ResNet50	70.3	76.2	79.1	74.0	61.8	78.5	0.73	0.75	70.4	76.3	38.2	21.5	20.9	26.0
VGG16	76.7	76.6	75.8	**77.6**	77.8	75.8	0.76	**0.77**	76.8	76.7	22.2	24.2	24.2	**22.4**
Xception	72.7	76.0	73.0	72.1	72.6	79.9	0.73	0.75	72.8	76.0	27.4	20.1	27.0	27.9
EfficientNetB0	74.0	76.7	73.0	74.0	75.4	79.6	0.74	0.76	74.2	76.8	24.6	20.4	27.0	26.0

Image augmentation. Several augmentation techniques are explored to improve models' generalizability. An optimal augmentation technique should not distort the BUS images, because tumor shapes, boundaries, echo patterns, and margins in breast cancer classification are essential in determining the tumor type. The classifiers are trained on six different augmentation techniques individually: horizontal flip, height shift, width shift, zoom, shear, and rotation. A combination of the four best-performed techniques including the horizontal flip, height shift (0.2), width shift (0.2), and rotation (20%), is chosen to augment the training set. The results in Table 5 demonstrate that the augmentation combination improves the overall performance of DenseNet121, InceptionV3, MobileNet, ResNet50, Xception, and

Table 6. Results of different loss functions.

Classifiers	Loss	Accuracy (%) ↑	Sensitivity (%) ↑	Specificity (%) ↑	F_1 ↑	AUC (%) ↑	FPR (%) ↓	FNR (%) ↓
DenseNet121	L_{BCE}	76.9	72.2	**81.9**	0.76	77.0	**18.1**	27.8
	L_{WBCE} (w_z=4)	72.7	90.1	55.7	**0.77**	72.9	44.3	9.90
	$L_{Focal}(\gamma = 3, \alpha = 0.8)$	70.3	88.6	52.6	0.75	70.6	47.4	11.4
InceptionV3	L_{BCE}	75.7	73.4	78.4	0.75	75.9	21.6	26.6
	$L_{WBCE}(w_z = 3)$	71.6	86.8	57.1	0.75	71.9	42.9	13.2
	$L_{Focal}(\gamma = 2, \alpha = 0.8)$	68.3	90.3	47.4	0.74	68.8	52.6	9.70
MobileNet	L_{BCE}	**77.2**	75.1	79.6	**0.77**	**77.4**	20.4	24.9
	$L_{WBCE}(w_z = 3)$	74.0	87.6	60.8	**0.77**	74.2	39.2	12.4
	$L_{Focal}(\gamma = 3, \alpha = 0.8)$	72.3	87.2	57.6	0.76	72.4	42.4	12.8
ResNet50	L_{BCE}	76.2	74.0	78.5	0.75	76.3	21.5	26.0
	$L_{WBCE}(w_z = 3)$	72.6	86.2	59.4	0.76	72.8	40.6	13.8
	$L_{Focal}(\gamma = 3, \alpha = 0.8)$	71.3	88.3	54.4	0.75	71.4	45.6	11.70
VGG16	L_{BCE}	76.6	77.6	75.8	**0.77**	76.7	24.2	22.4
	$L_{WBCE}(w_z = 3)$	74.5	86.7	62.6	**0.77**	74.7	37.4	13.3
	$L_{Focal}(\gamma = 2, \alpha = 0.8)$	70.3	90.2	50.9	0.75	70.5	49.1	9.80
Xception	L_{BCE}	76.0	72.1	79.9	0.75	76.0	20.1	27.9
	$L_{WBCE}(w_z=3)$	72.9	88.7	57.7	**0.77**	73.2	42.3	11.30
	$L_{Focal}(\gamma = 2, \alpha = 0.8)$	68.2	**91.9**	45.1	0.74	68.5	54.9	**8.10**
EfficientNetB0	L_{BCE}	76.7	74.0	79.6	0.76	76.8	20.4	26.0
	$L_{WBCE}(w_z=3)$	73.8	86.8	61.2	**0.77**	74.0	38.8	13.2
	$L_{Focal}(\gamma = 2, \alpha = 0.8)$	69.6	91.3	48.5	0.75	69.9	51.5	8.70

EffienetNetB0 classifiers except for VGG16. This is because the VGG16 without augmentation has less overfitting than other approaches, and extra augmented images do not improve its performance significantly. The proposed combination of augmentation techniques is utilized for all classifiers to expand the dataset size in the remaining experiments.

Loss functions. As described in section 3.4, the binary cross-entropy loss(L_{BCE}), focal loss [48] (L_{Focal}), and weighted cross-entropy loss (L_{WBCE}) are evaluated. Table 6 shows the performance of different models with the loss parameter(s) that leads to the best overall and sensitivity values. By utilizing the L_{WBCE} , the sensitivity of DenseNet121, InceptionV3, MobileNet, ResNet50, VGG16, Xception, and EfficientNet

improved by 19.8%, 15.4%, 14.2%, 14.1%, 10.4%, 18.7%, and 14.7%, respectively. Additionally, with the Focal loss, the sensitivity has further improved, but the overall performance degrades considerably. For example, the sensitivity of InceptionV3 and Xception has increased by 18.7%, and 21.5%, respectively; however, the AUC is reduced by 9.3%, and 9.8%, respectively. The best trade-off between sensitivity and specificity is achieved by MobileNet and VGG16 when L_{WBCE} is used.

Table 7. Results of different optimizers.

Classifier	Optimizer	Accuracy (%) ↑	Sensitivity (%) ↑	Specificity (%) ↑	F_1 ↑	AUC (%) ↑	FPR (%) ↓	FNR (%) ↓
DenseNet121	**ADAM**	72.7	**90.1**	55.7	**0.77**	72.9	44.3	**9.9**
	SGD	71.6	89.0	54.8	0.76	71.9	45.2	11.0
	NADAM	71.1	87.7	55.0	0.75	71.3	45.0	12.3
InceptionV3	ADAM	71.6	86.8	57.1	0.75	71.9	42.9	13.2
	SGD	73.0	88.4	57.6	**0.77**	73.0	42.4	11.6
	NADAM	70.5	86.5	54.6	0.74	70.6	45.4	13.5
MobileNet	ADAM	74.0	87.6	60.8	**0.77**	74.2	39.2	12.4
	SGD	74.0	87.4	61.3	**0.77**	74.4	38.7	12.6
	NADAM	72.4	83.6	61.5	0.75	72.5	38.5	16.4
ResNet50	**ADAM**	72.6	86.2	59.4	0.76	72.8	40.6	13.8
	SGD	70.8	87.6	54.6	0.75	71.1	45.4	12.4
	NADAM	70.8	85.2	56.7	0.74	70.9	43.3	14.8
VGG16	**ADAM**	**74.5**	86.7	**62.6**	**0.77**	**74.7**	**37.4**	13.3
	SGD	70.2	89.7	51.1	0.75	70.4	48.9	10.3
	NADAM	71.5	86.3	57.0	0.75	71.7	43.0	13.7
Xception	ADAM	72.9	88.7	57.7	**0.77**	73.2	42.3	11.3
	SGD	73.7	88.5	59.6	**0.77**	74.0	40.4	11.5
	NADAM	69.1	87.6	50.0	0.74	68.8	50.0	12.4
EfficientNetB0	**ADAM**	73.8	86.8	61.2	**0.77**	74.0	38.8	13.2
	SGD	73.8	86.2	61.7	**0.77**	73.9	38.3	13.8
	NADAM	72.4	85.1	59.7	0.75	72.4	40.3	14.9

Optimizers. We compare three popular optimizers: Adaptive Moment Estimation (ADAM) [49], Stochastic Gradient Descent (SGD) with momentum, and Nesterov-accelerated Adaptive Moment Estimation (NADAM) [50]. In the experiments, ADAM is applied with a learning rate of 0.00001, SGD with a learning rate of 0.002 and momentum of 0.9, and NADAM with a learning rate of 0.00001, beta_1 of 0.9, beta_2 of 0.999, and epsilon of 1e-08. All other parameters take default values in Keras.

Table 8. Results of five deep NNs using multitask learning**.**

Classifiers	Accuracy (%) ↑		Sensitivity (%) ↑		Specificity (%) ↑		F_1 ↑		AUC (%) ↑		FPR (%) ↓		FNR (%) ↓	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
DenseNet121	82.2	85.0	75.3	79.1	87.1	88.9	0.76	0.80	81.2	84.0	12.9	11.1	24.7	20.9
MobileNet	85.1	87.0	78.1	81.1	90.2	91.0	0.81	**0.83**	84.1	**86.1**	9.8	9.0	21.9	18.9
ResNet50	85.1	86.1	78.5	80.1	89.2	89.0	0.80	0.81	83.8	85.0	10.7	10.9	21.5	21.3
VGG16	**86.1**	87.1	81.0	**81.3**	**91.2**	90.9	**0.82**	**0.83**	**86.1**	**86.1**	**8.8**	9.1	19.0	**18.7**
EfficientNetB0	84.2	**87.5**	**81.2**	81.0	86.9	**91.2**	0.80	**0.83**	84.0	**86.1**	13.1	**8.8**	**18.8**	19.0

As shown in Table 7, DenseNet121, ResNet50, VGG16, and EfficientNet classifiers achieved better F_1 scores and AUC values using the ADAM optimizer. On the other hand, InceptionV3, MobileNet, and Xception achieved better results using the SGD optimizers. It is worth mentioning that the optimizers have the slightest impact on the generalization performance among all the strategies that we tested. DenseNet121 achieved the best sensitivity with 90.1% by using ADAM optimizers, and EfficientNetB0 attained the lowest sensitivity with 85.1% by using the NADAM optimizer. In addition, the VGG16 using Adam and MobileNet using SGD achieved the best AUC by 74.7% and 74.4%, respectively.

5.2 Multitask learning

The multitask learning approaches need ground truth labels for both tumor class and tumor boundaries, and a combined dataset (BUSI and BUSIS) with a total of 1,209 BUS images is used. BUSI and BUSIS are chosen because they have accurate annotations for both tumor boundaries and classes. The 5-fold cross-validation is utilized to evaluate the performance of all approaches. The max epoch is set to 70, and the batch size is 32. We optimize all approaches using ADAM [49].

5.2.1 The effectiveness of multitask learning using generic deep learning models

Many previous studies [34-36] have demonstrated the effectiveness of integrating tumor segmentation tasks into tumor classification networks. In BUS images, the shared representations between tumor classification and segmentation tasks include tumor morphology, size, shape, and echo pattern. We evaluate multitask learning networks with five different pretrained (ImageNet) backbone networks, DenseNet121,

MobileNet, ResNet50, VGG16, and EfficientNetB0. A subnetwork [34] is added to perform breast tumor segmentation at the end of the convolutional layers of the backbone network. The subnetwork consists of four blocks, each of which contains one upsampling layer, and two consecutive 3×3 convolution layers with batch normalization and ReLU activation. The loss function is a combination of both the Dice loss and binary cross-entropy loss. The weight for the binary cross-entropy loss is set to 1.5 by experiments.

As shown in Table 8, with the additional segmentation task, the overall performance of the five approaches can be improved. VGG16, MobileNet, and EfficientNetB0 achieve the best AUC of 86.1% among all the approaches. The sensitivity of DenseNet121 is improved by 5%. It is worth noticing that, in all approaches, the specificity values are significantly higher compared to the sensitivity values. We observed the same outcome in [34] and [35]. This issue could be addressed by choosing the weighted binary cross-entropy function.

Table 9. Results of three multitask learning approaches developed for BUS image classification.

Approaches	Accuracy (%) ↑	Sensitivity (%) ↑	Specificity (%) ↑	F_1 ↑	AUC (%) ↑	FPR (%) ↓	FNR (%) ↓
Zhang, et al. [35]	87.4	81.4	**91.4**	0.83	86.4	**8.6**	18.6
Vakanski, et al. [34]	83.6	77.4	87.8	0.78	82.6	12.2	22.5
Shi, et al. [36]	83.9	87.3	81.7	0.80	84.5	18.3	12.6
MT-ESTAN	**90.0**	**90.4**	89.8	**0.88**	**90.1**	10.2	**9.6**

5.2.2 The effectiveness of the proposed MT-ESTAN

In this section, we compare the proposed MT-ESTAN with three multitask learning approaches [34-36]. We obtained the source code from the authors of [34, 36], and implemented the approach in [35], all model parameters were adopted from the papers.

As shown in Table 9, the AUC of the proposed MT-ESTAN is significantly higher than those of [34], [36], and [35], and MT-ESTAN outperforms all approaches reported in Table 8. For example, compare to the best performed multitask network (VGG16) in Table 8, the proposed MT-ESTAN improves the sensitivity, F_1 score, and AUC by 11.2%, 7.3%, and 4.6%, respectively. However, [34-36] are not significantly better than the multitask learning approaches reported in table 8. [34-36] achieves high

specificity values, but at the cost of low sensitivity values, which leads to high false negative rates (FNRs), e.g., the FNR of [36] is 18.6%. In addition, all multitask learning approaches in have low sensitivity values and high FNRs. The proposed MT-ESTAN achieves a better balance between the sensitivity and specificity, and has a low FNR of 9.6%.

6. Discussion

The experiments and similar outcomes in [36, 50- 51] demonstrate that the transfer learning (TL) strategy consistently outperforms training from scratch for deep learning approaches for BUS image classification, which implies that knowledge learned from a different domain (e.g., nature images) could be transferred and used to improve BUS image classification. BUS images share common image elements in natural images, e.g., object boundaries, image contrast, and texture, and deep neural networks learning the representations of those elements from nature images can also contribute to BUS image classification. Inspired by this, medical image datasets sharing common features with BUS images could be applied to further improve the performance of deep learning approaches for BUS image classification. For example, ultrasound images from other organs and breast images from other modalities (e.g., MRI, CT, and Mammogram) can be used to pretrain BUS image classifiers.

Our results and previous studies [14] suggest that image augmentation techniques could improve the generalizability of most deep learning approaches for BUS image classification. Augmentation techniques introduce variations and enlarge the training set size, and could prevent overfitting [52]; and model training using an augmented dataset alleviates the issue of the small size of the medical dataset. To further increase the generalizability of deep learning models, the simplest way is to add more images from different sources to the model training. The additional images could be either new real BUS images or synthetic images generated using algorithms [53].

Many BUS image classification approaches have achieved promising overall performance (e.g., accuracy and F_1 score), but failed to balance the sensitivity and specificity. They used the binary cross-entropy as the loss function and treat cancer and non-cancer cases equally, which makes predictions that

favor the dominant class, e.g., benign class, and produce low sensitivities. Sensitivity is the most important assessment metric in breast cancer detection because missing malignant cases may risk patients' lives; and a well-balanced model should achieve both high overall performance and high sensitivity. One solution is to utilize the numerically-stable weighted cross-entropy function discussed in Section 3.4 to achieve a better balance between the sensitivity and specificity.

Multitask learning (MTL) is a promising future direction to improve the robustness and generalization of deep learning approaches for BUS image classification. Table 8 demonstrates that MTL networks with a primary BUS tumor classification task and a secondary segmentation task outperform single-task networks with only the classification task. The segmentation task incorporates semantic information, i.e., tumor region, during the training, which enables an MTL network to learn meaningful and focused representations in tumor regions rather than random features from a whole BUS image. This secondary task performs as a regularizer that could also improve models' convergence using small or medium datasets. Inspired by this finding, researchers can further advance BUS image classification by incorporating other semantic knowledge, e.g., breast anatomy and BI-RADs descriptors, into MTL networks.

Last but not least, to improve the adoption and trustworthiness of CAD systems for breast cancer detection, the explainability of approaches should be improved. Existing deep learning-based methods still have a black-box nature in which limited information is provided to help understand the BUS image classification process [54-55]. This gap discourages radiologists from using BUS CADs in clinical practice. Therefore, solving this gap by introducing explainability into models [54] is a promising direction for BUS image classification.

7. Conclusion

In this work, we build a public benchmark for the classification of B-mode BUS images which consists of a diverse dataset, useful strategies, and findings for developing deep learning-based approaches, and a novel MTL network, MT-ESTAN, for accurate BUS image classification.

The benchmark dataset comprises 3,641 B-mode BUS images from five countries, and a set of public software tools for data preparing and preprocessing. The BUS images were collected with different ultrasound devices and patient populations, and have a wide variation in image contrast, brightness, level of noise, etc. We highlight three major findings by evaluating 10 deep learning-based approaches using the benchmark dataset: 1) Transfer learning and image augmentation are effective strategies to significantly improve the overall performance of deep learning-based BUS image classifiers; 2) the numerically-stable weighted cross-entropy loss function offers a better balance between the sensitivity and specificity; 3) MTL networks with both the breast tumor segmentation and classification tasks is one of the most useful strategies to improve the generalization of deep learning approaches for BUS image classification.

The newly proposed MT-ESTAN incorporates a small-tumor aware network as the backbone network, and consists of one primary task (tumor classification) and a secondary task (tumor segmentation). The results show that MT-ESTAN achieves state-of-the-art performance, and significantly improved the sensitivity of the model.

In the future, we will be continuously adding more BUS images, new findings, and emerging approaches to the benchmark.

Acknowledgments

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM104420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] American Cancer Society.Cancer Facts & Figures, (2022). http://cancerstatisticscenter.cancer.org.

[2] F.Z. Francies, R. Hull, R. Khanyile, Z. Dlamini, Breast cancer in low-middle income countries: abnormality in splicing and lack of targeted treatment options, Am. J. Cancer Res. 10 (2020) 1568–1591.

[3] S.H. Kim, H.H. Kim, W.K. Moon, Automated Breast Ultrasound Screening for Dense Breasts, Korean J. Radiol. 21 (2020) 15–24. https://doi.org/10.3348/kjr.2019.0176.

[4] M. Rebolj, V. Assi, A. Brentnall, D. Parmar, S.W. Duffy, Addition of ultrasound to mammography in the case of dense breast tissue: Systematic review and meta-analysis, Br. J. Cancer. 118 (2018) 1559–1570. https://doi.org/10.1038/s41416-018-0080-3.

[5] C. Byrne, C. Schairer, J. Wolfe, N. Parekh, M. Salane, L.A. Brinton, R. Hoover, R. Haile, Mammographic features and breast cancer risk: Effects with time, age, and menopause status, J. Natl. Cancer Inst. 87 (1995) 1622–1629. https://doi.org/10.1093/jnci/87.21.1622.

[6] G. Ursin, H. Ma, A.H. Wu, L. Bernstein, M. Salane, Y.R. Parisky, M. Astrahan, C.C. Siozon, M.C. Pike, Mammographic density and breast cancer in three ethnic groups, Cancer Epidemiol. Biomarkers Prev. 12 (2003) 332–338.

[7] R.M. Tamimi, C. Byrne, G.A. Colditz, S.E. Hankinson, Endogenous hormone levels, mammographic density, and subsequent risk of breast cancer in postmenopausal women, J. Natl. Cancer Inst. 99 (2007) 1178–1187. https://doi.org/10.1093/jnci/djm062.

[8] M.N. Linver, 4-19 Mammographic Density and the Risk and Detection of Breast Cancer, Breast Dis. 18 (2008) 364–365. https://doi.org/10.1016/S1043-321X(07)80400-0.

[9] L. Yaghjyan, G.A. Colditz, L.C. Collins, S.J. Schnitt, B. Rosner, C. Vachon, R.M. Tamimi, Mammographic breast density and subsequent risk of breast cancer in postmenopausal women according to tumor characteristics, J. Natl. Cancer Inst. 103 (2011) 1179–1189. https://doi.org/10.1093/jnci/djr225.

[10] Y. Zhang, M. Xian, H. Da Cheng, B. Shareef, J. Ding, F. Xu, K. Huang, B. Zhang, C. Ning, Y. Wang, BUSIS: A Benchmark for Breast Ultrasound Image Segmentation, Healthc. 10 (2022).

[11] G.S. Lodwick, T.E. Keats, J.P. Dorst, The coding of roentgen images for computer analysis as applied to lung cancer., Radiology. 81(2) (1963) 185–200.

[12] H.D. Cheng, J. Shan, W. Ju, Y. Guo, L. Zhang, Automated breast cancer detection and classification using ultrasound images: A survey, Pattern Recognit. 43 (2010) 299–317. https://doi.org/10.1016/j.patcog.2009.05.012.

[13] Q. Huang, F. Zhang, X. Li, Machine Learning in Ultrasound Computer-Aided Diagnostic Systems: A Survey, Biomed Res. Int. 2018 (2018). https://doi.org/10.1155/2018/5137904.

[14] W. Al-Dhabyani, A. Fahmy, M. Gomaa, H. Khaled, Deep learning approaches for data augmentation and classification of breast masses using ultrasound images, Int. J. Adv. Comput. Sci. Appl. 10 (2019) 618–627. https://doi.org/10.14569/ijacsa.2019.0100579.

[15] M. Byra, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, M. Andre, Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion, Med. Phys. 46 (2019) 746–755. https://doi.org/10.1002/mp.13361.

[16] B. Shareef, A. Vakanski, P.E. Freer, M. Xian, ESTAN: Enhanced Small Tumor-Aware Network for Breast Ultrasound Image Segmentation, Healthcare. 10 (2022) 2262. https://doi.org/10.3390/healthcare10112262.

[17] B. Shareef, M. Xian, A. Vakanski, STAN : Small Tumor-Aware Network for Breast Ultrasound Image Segmentation, IEEE 17th Int. Symp. Biomed. Imaging (ISBI 2020). (2020).

[18] M. Xian, Y. Zhang, H.D. Cheng, F. Xu, B. Zhang, J. Ding, Automatic breast ultrasound image segmentation: A survey, Pattern Recognit. 79 (2018) 340–355. https://doi.org/10.1016/j.patcog.2018.02.012.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. 115 (2015) 211–252. https://doi.org/10.1007/s11263-015-0816-y.

[20] B. Huynh, K. Drukker, M. Giger, Computer-Aided Diagnosis of Breast Ultrasound Images Using Transfer

Learning From Deep Convolutional Neural Networks, Med. Phys. 43 (2016) 3705–3705.
[21] Y. Liang, R. He, Y. Li, Z. Wang, Simultaneous segmentation and classification of breast lesions from ultrasound images using Mask R-CNN, IEEE Int. Ultrason. Symp. IUS. 2019-Octob (2019) 1470–1472. https://doi.org/10.1109/ULTSYM.2019.8926185.
[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common Objects in Context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Comput. Vis. -- ECCV 2014, Springer International Publishing, Cham, 2014: pp. 740–755.
[23] A. Hijab, M.A. Rushdi, M.M. Gomaa, A. Eldeib, Breast Cancer Classification in Ultrasound Images using Transfer Learning, Int. Conf. Adv. Biomed. Eng. ICABME. 2019-Octob (2019) 1–4. https://doi.org/10.1109/ICABME47164.2019.8940291.
[24] Z. Cao, L. Duan, G. Yang, T. Yue, Q. Chen, An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures, BMC Med. Imaging. 19 (2019) 1–9. https://doi.org/10.1186/s12880-019-0349-x.
[25] W.C. Shia, D.R. Chen, Classification of malignant tumors in breast ultrasound using a pretrained deep residual network model and support vector machine, Comput. Med. Imaging Graph. 87 (2021) 101829. https://doi.org/10.1016/j.compmedimag.2020.101829.
[26] H. Zhang, L. Han, K. Chen, Y. Peng, J. Lin, Diagnostic Efficiency of the Breast Ultrasound Computer-Aided Prediction Model Based on Convolutional Neural Network in Breast Cancer, J. Digit. Imaging. 33 (2020) 1218–1223. https://doi.org/10.1007/s10278-020-00357-7.
[27] J. Xie, X. Song, W. Zhang, Q. Dong, Y. Wang, F. Li, C. Wan, A novel approach with dual-sampling convolutional neural network for ultrasound image classification of breast tumors, Phys. Med. Biol. 65 (2020). https://doi.org/10.1088/1361-6560/abc5c7.
[28] S. Han, H.K. Kang, J.Y. Jeong, M.H. Park, W. Kim, W.C. Bang, Y.K. Seong, A deep learning framework for supporting the classification of breast lesions in ultrasound images, Phys. Med. Biol. 62 (2017) 7714–7728. https://doi.org/10.1088/1361-6560/aa82ec.
[29] J. Xing, C. Chen, Q. Lu, X. Cai, A. Yu, Y. Xu, X. Xia, Y. Sun, J. Xiao, L. Huang, Using BI-RADS Stratifications as Auxiliary Information for Breast Masses Classification in Ultrasound Images, IEEE J. Biomed. Heal. Informatics. XX (2020) 1–1. https://doi.org/10.1109/jbhi.2020.3034804.
[30] Z. Zhuang, Z. Yang, S. Zhuang, A.N.J. Raj, Y. Yuan, R. Nersisson, Multi-Features-Based Automated Breast Tumor Diagnosis Using Ultrasound Image and Support Vector Machine, Comput. Intell. Neurosci. 2021 (2021). https://doi.org/10.1155/2021/9980326.
[31] W.X. Liao, P. He, J. Hao, X.Y. Wang, R.L. Yang, D. An, L.G. Cui, Automatic Identification of Breast Ultrasound Image Based on Supervised Block-Based Region Segmentation Algorithm and Features Combination Migration Deep Learning Model, IEEE J. Biomed. Heal. Informatics. 24 (2020) 984–993. https://doi.org/10.1109/JBHI.2019.2960821.
[32] X. Fei, S. Zhou, X. Han, J. Wang, S. Ying, C. Chang, W. Zhou, J. Shi, Doubly supervised parameter transfer classifier for diagnosis of breast cancer with imbalanced ultrasound imaging modalities, Pattern Recognit. 120 (2021) 108139. https://doi.org/10.1016/J.PATCOG.2021.108139.
[33] Z. Zhuang, Z. Yang, A.N.J. Raj, C. Wei, P. Jin, S. Zhuang, Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion, Comput. Methods Programs Biomed. (2021) 106221. https://doi.org/10.1016/j.cmpb.2021.106221.
[34] A. Vakanski, M. Xian, Evaluation of Complexity Measures for Deep Learning Generalization in Medical Image Analysis, 2021 IEEE 31st Int. Work. Mach. Learn. Signal Process. (n.d.) 1–15.
[35] G. Zhang, K. Zhao, Y. Hong, X. Qiu, K. Zhang, B. Wei, SHA-MTL : soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification, Int. J. Comput. Assist. Radiol. Surg. (2021). https://doi.org/10.1007/s11548-021-02445-7.
[36] J. Shi, A. Vakanski, M. Xian, J. Ding, C. Ning, EMT-NET: Efficient multitask network for computer-aided diagnosis of breast cancer, in: 2022 IEEE 19th Int. Symp. Biomed. Imaging, IEEE, 2022: pp. 1–5.
[37] T. Geertsma, Ultrasoundcases.info, FujiFilm. (2014). https://www.ultrasoundcases.info/.
[38] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Br. 28 (2020) 104863. https://doi.org/10.1016/j.dib.2019.104863.
[39] A. Rodtook, K. Kirimasthong, W. Lohitvisate, S.S. Makhanov, Automatic initialization of active contours and level set method in ultrasound images of breast abnormalities, Pattern Recognit. 79 (2018) 172–182. https://doi.org/10.1016/j.patcog.2018.01.032.

[40] M.H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, A.K. Davison, R. Martí, Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks, IEEE J. Biomed. Heal. Informatics. 22 (2018) 1218–1226. https://doi.org/10.1109/JBHI.2017.2731873.
[41] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, (2017). http://arxiv.org/abs/1704.04861.
[42] M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, 36th Int. Conf. Mach. Learn. ICML 2019. 2019-June (2019) 10691–10700.
[43] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, 2017: pp. 4700–4708. https://doi.org/10.1109/CVPR.2017.243.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem (2016) 770–778. https://doi.org/10.1109/CVPR.2016.90.
[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. (2015) 1–14.
[46] F. Chollet, Xception: Deep learning with depthwise separable convolutions, Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. 2017-Janua (2017) 1800–1807. https://doi.org/10.1109/CVPR.2017.195.
[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem (2016) 2818–2826. https://doi.org/10.1109/CVPR.2016.308.
[48] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal Loss for Dense Object Detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2018) 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.
[49] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, ArXiv:1412.6980. (2014). http://arxiv.org/abs/1412.6980.
[50] T. Dozat, Incorporating Nesterov Momentum into Adam, ICLR Work. (2016) 2013–2016.
[51] T. Xiao, L. Liu, K. Li, W. Qin, S. Yu, Z. Li, Comparison of Transferred Deep Neural Networks in Ultrasonic Breast Masses Discrimination, Biomed Res. Int. 2018 (2018). https://doi.org/10.1155/2018/4605191.
[52] S.S. Yadav, S.M. Jadhav, Deep convolutional neural network based medical image classification for disease diagnosis, J. Big Data. 6 (2019). https://doi.org/10.1186/s40537-019-0276-2.
[53] H. Wang, M. Xian, A. Vakanski, B. Shareef, SIAN: Style-Guided Instance-Adaptive Normalization for Multi-Organ Histopathology Image Synthesis, (2022). http://arxiv.org/abs/2209.02412.
[54] B. Zhang, A. Vakanski, M. Xian, BI-RADS-Net: An Explainable Multitask Learning Approach for Cancer Diagnosis in Breast Ultrasound Images, in: IEEE 31st Int. Work. Mach. Learn. Signal Process., 2021.
[55] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable deep learning models in medical image analysis, J. Imaging. 6 (2020) 1–19. https://doi.org/10.3390/JIMAGING6060052.
[56] M.H. Yap, M. Goyal, F.M. Osman, R. Martí, E. Denton, A. Juette, R. Zwiggelaar, Breast ultrasound lesions recognition: end-to-end deep learning approaches, J. Med. Imaging. 6 (2018) 1. https://doi.org/10.1117/1.JMI.6.1.011007.
[57] H. Tanaka, S.W. Chiu, T. Watanabe, S. Kaoku, T. Yamaguchi, Computer-aided diagnosis system for breast ultrasound images using deep learning, Phys. Med. Biol. 64 (2019). https://doi.org/10.1088/1361-6560/ab5093.
[58] W.K. Moon, Y.W. Lee, H.H. Ke, S.H. Lee, C.S. Huang, R.F. Chang, Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks, Comput. Methods Programs Biomed. 190 (2020). https://doi.org/10.1016/j.cmpb.2020.105361.