A Novel Deep Learning System for Breast Lesion Risk Stratification in Ultrasound Images

Ting Liu, Xing An, Yanbo Liu, Yuxi Liu, Bin Lin, Runzhou Jiang, Wenlong Xu, Longfei Cong, and Lei Zhu(✉)

Shenzhen Mindray BioMedical Electronics, Co., Ltd., Shenzhen, China
zhulei@mindray.com

Abstract. This paper presents a novel deep learning system to classify breast lesions in ultrasound images into benign and malignant and into Breast Imaging Reporting and Data System (BI-RADS) six categories simultaneously. A multi-task soft label generating architecture is proposed to improve the classification performance, in which task-correlated labels are obtained from a dual-task teacher network and utilized to guide the training of a student model. In student model, a consistency supervision mechanism is embedded to constrain that a prediction of BI-RADS is consistent with the predicted pathology result. Moreover, a cross-class loss function that penalizes different degrees of misclassified items with different weights is introduced to make the prediction of BI-RADS closer to the annotation. Experiments on our private and two public datasets show that the proposed system outperforms current state-of-the-art methods, demonstrating the great potential of our method in clinical diagnosis.

Keywords: Classification, Breast ultrasound image, Multitask soft label, Consistency supervision, Cross-class loss

1 Introduction

Breast cancer has become the most commonly diagnosed cancer in women with a critical mortality [1]. In clinic, ultrasound has been widely used for breast screening and lesion diagnosis. Experienced radiologists can recognize the malignant risk of lesions from their appearance in ultrasound, and issue reports referring to the Breast Imaging Reporting and Data System (BI-RADS) [2] guideline to advise treatments. The BI-RADS guideline includes several essential terms for describing breast lesions and criteria for classifying the risk. Specifically, BI-RADS divides the malignancy likelihood of a lesion into 6 categories, i.e. BI-RADS 2, 3, 4a, 4b, 4c and 5. A higher BI-RADS grade means a larger malignant risk [2] as shown in Fig. 1. Patients with any suspicious mass that is assessed as category 4a, 4b, 4c or 5 will normally be suggested to undergo a preliminary biopsy or surgical excision [2].

Over the last decade, various computer-aided diagnosis (CAD) systems have been developed to assess the malignant risk of breast lesions in ultrasound, which can effectively relieve the workload of physicians and improve their diagnostic performance [3]. Han et al. [4] adopted a pretrained convolutional neural network (CNN) model to

categorize lesions in ultrasound as benign and malignant. Qian et al. [5] developed a combined ultrasonic B-mode and color Doppler system to classify the malignant risk of a lesion into four BI-RADS categories, i.e. BI-RADS 2, 3, 4 (including 4a, 4b and 4c) and 5, and the sum of the probabilities of BI-RADS 2 and 3 was considered as the benign probability while that of BI-RADS 4 and 5 was regard as malignant likelihood. Liu et al. [6] and Xing et al. [7] developed deep learning models with two branches, one for classifying lesions into BI-RADS six categories and the other for pathologic binary classification (benign and malignant).

Although many researches have achieved good performances, two main issues remain. First, several studies demonstrated that combining BI-RADS stratifications and pathology classification could improve the classification performance, but they achieved the combination by simply adding two branches following the CNN, ignoring the relevance between them [7-8]. Second, no clear distinction between adjacent BI-RADS categories exists in BI-RADS guideline, and the general rule is that a higher BI-RADS grade represents a larger malignant likelihood [2]. Therefore, minor differences between BI-RADS assessments of a lesion given by different physicians or CADs exist and are acceptable, while huge distinctions are unreasonable. However, to the best of our knowledge, few studies focused on minimizing the gap between their predicted BI-RADS categories and the annotations.

To solve the problems above, an automatic breast lesion risk stratification system that classifies lesions into benign and malignant and into BI-RADS categories simultaneously is presented. we propose a multitask soft label generating architecture, in which task-correlated soft labels are generated from a dual-task teacher network and utilized to train a student model. Moreover, a consistency supervision mechanism is proposed to constrain that the prediction of BI-RADS category is consistent with the predicted pathology result, and a cross-class loss function that penalizes the different degrees of misclassification items with different weights is introduced to make the prediction of BI-RADS categories as closer as the annotations.

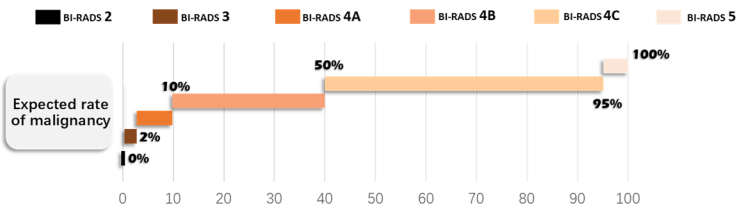

Fig. 1. Expected malignant rate of BI-RADS categories [2].

2 Methodology

The overall architecture is illustrated in Fig. 2. The teacher model is a dual-task network, one task is to classify lesions into BI-RADS six categories, the other is for benign and malignant classification. Task-correlated soft labels are obtained from the teacher network and utilized to train the student model. In student model, consistency

supervision mechanism (CSM) constrains that a lesion predicted as BI-RADS 2 or 3 (BI-RADS 4c or 5) is categorized as benign (malignant), thus making the predictions of two branches consistent. The cross-class loss function (CCLF) penalizes different degrees of mis-classified items of BI-RADS categories with different weights to make the prediction closer to the annotation. Details are described as follows.

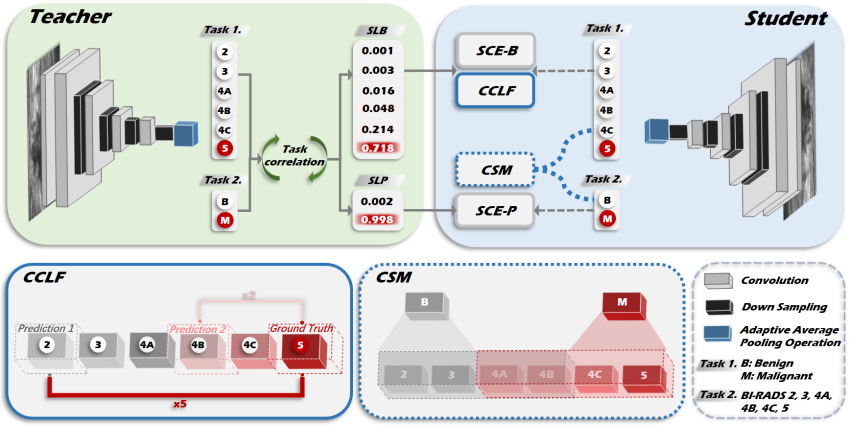

Fig. 2. Illustration of our method. *SLB:* Soft label of BI-RADS. *SLP:* Soft label of pathology. *SCE-B:* Soft label cross entropy of BI-RADS. *SCE-P:* Soft label cross entropy of pathology. *CCLF:* Cross-class loss function. *CSM:* Consistency supervision mechanism.

2.1 Multitask Label Generating Architecture

Soft labels possess information between different classes and are usually used to improve classification performance [8]. Multitask learning can improve prediction results by message sharing and correlating [9]. In this paper, we combine the multitask learning with soft label strategy to obtain task-correlated soft labels which not only contain the relation between different classes but also include relevance between BI-RADS categories and pathology classes. Specifically, we propose a multitask teacher model to learn the task-correlated labels and employ them to train the student network. The teacher model consists of a CNN module, an adaptive average pooling operation, and two classification branches at the end. RepVGG-A2 [10] is adopted as the backbone network to extract features. Hard labels of BI-RADS categories and pathologic information are utilized to train the teacher model using two normal cross entropy loss functions.

The soft labels of BI-RADS (SLB_i) and pathology (SLP_j) are expressed as:

$$\begin{cases} SLB_i = \frac{1}{N_i}\sum tb'(x_{ij}), tbc'(x_{ij}) = i \;\&\; tpc'(x_{ij}) = j \\ SLP_j = \frac{1}{N_j}\sum tp'(x_{ij}), tpc'(x_{ij}) = j \;\&\; tbc'(x_{ij}) = i \end{cases} \tag{1}$$

where x_{ij} denotes an input image belonging to the i^{th} BI-RADS and the j^{th} pathology. $i \in \{0, 1, 2, 3, 4, 5\}$ and j is 0 or 1. We ran the trained teacher model on the training set, and obtained predicted results. $tb'(x)$ and $tbc'(x)$ represent the output probability vector and predicted category of BI-RADS, and $tp'(x)$ and $tpc'(x)$ are that of pathology respectively. To compute soft label of the i^{th} BI-RADS (SLB_i), a predicted probability vector of BI-RADS is summed up if the predicted BI-RADS result is i and the pathology result equals to the annotation ($tbc'(x_{ij}) = i$ & $tpc'(x_{ij}) = j$). N_i is the number of all qualified cases. The calculation of SLP_j is similar to SLB_i.

The task-correlated labels are then used to train the student model that has the same structure as the teacher network. The loss function is defined as:

$$L_{sl} = -\sum b \cdot log b'(x) - \sum p \cdot log p'(x) \tag{2}$$

where x denotes the input image. b and p represent the labels from SLB and SLP respectively. $b'(x)$ and $p'(x)$ are the prediction probabilities of BI-RADS and pathology from the student model.

2.2 Consistency Supervision Mechanism

According to [2], a lesion annotated as BI-RADS 2 or 3 is more likely to be benign while a lesion with BI-RADS 4c or 5 is more likely to be malignant. We propose a consistency supervision mechanism (CSM) to constrain the above relevance to make the predictions of two branches consistent. That is, the CSM restricts that a lesion with prediction of BI-RADS 2 or 3 is predicted as benign while a lesion predicted as BI-RADS 4c or 5 is classified as malignant. The consistency loss function is defined as:

$$L_c = -\sum \begin{cases} p_B \cdot log(1 - \sum_{4c}^{5} b'(x)), & p_B \geq 0.5 \\ p_M \cdot log(1 - \sum_{2}^{3} b'(x)), & else \end{cases} \tag{3}$$

where $p_B + p_M = 1$. p_B and p_M represent the benign and malignant value in the soft label from SLP. $p_B \geq 0.5$ means the input lesion is benign while $p_B < 0.5$ represents that the lesion is malignant. $b'(x)$ is the predicted probability of BI-RADS. $\sum_{4c}^{5} b'(x)$ represents the sum of predicted probabilities of BI-RADS 4c and 5. $\sum_{2}^{3} b'(x)$ is the sum of probabilities of BI-RADS 2 and 3. By optimizing the L_c, the predicted probability of BI-RADS 2 or 3 (BI-RADS 4c or 5) is positively associated with that of benign (malignant). Blurred operation was made for BI-RADS 4a and 4b, since the malignant probability of them is not significant (varies from 2% to 50%) [2].

2.3 Cross-Class Loss Function

As mentioned above, minor differences between BI-RADS assessments given by different physicians or CADs are acceptable, while huge distinctions are unreasonable. Hence, we propose a cross-class loss function that penalizes the different degrees of misclassification items with different weights to make the prediction of BI-RADS closer to the annotation. The cross-class loss function is defined as:

$$L_{cc} = e^{\|n-m\|} \cdot (b_m{}'(x) - b_m)^2 \quad (4)$$

where n and m represent the n^{th} category annotated by radiologists and m^{th} category predicted by the model respectively. The prediction is wrong when n is not equal to m. $b_m{}'(x)$ and b_m are the m^{th} value in the predicted probability vector and the soft label from SLB respectively. $e^{\|n-m\|}$ is the penalty coefficient, and a larger misclassified degree results in a greater punishment coefficient.

The total loss in the student network is summarized as:

$$\mathcal{L} = \alpha \cdot L_{sl} + \beta \cdot L_c + \gamma \cdot L_{cc} \quad (5)$$

where α, β and γ are balance parameters, and are set to 1, 0.5 and 0.5 respectively.

3 Experiments

Dataset. Pathologically confirmed breast lesions in women were collected from 32 hospitals and one lesion per patient was included. BI-RADS assessment categories were annotated by several high-experienced radiologists according to the BI-RADS lexicon. A total of 5012 patients with 3220 benign and 1792 malignant lesions were collected, including 14042 images (7793 benign and 6249 malignant images). The dataset was divided at patient-level. The patients were divided into training set (4010 lesions with 11258 images), validation set (501 lesions with 1382 images) and testing set (501 lesions with 1402 images) according to 8:1:1. The average patient age was 43.8±13.2 years for the training set, 44.0±12.5 years for the validation set and 42.8±13.1 for the testing set. The details of each set are described in Table 1 (B: benign, M: malignant) and Fig. 3.

Table 1. Benign and malignant distribution at patient-level.

	Training		Validation		Testing	
	B	M	B	M	B	M
Patients	2564	1446	329	172	327	174
Images	6207	5051	798	584	788	614
Average age (years)	43.8±13.2		44.0±12.5		42.8±13.1	
Average size (cm)	1.86±1.00		1.82±1.03		1.80±0.94	

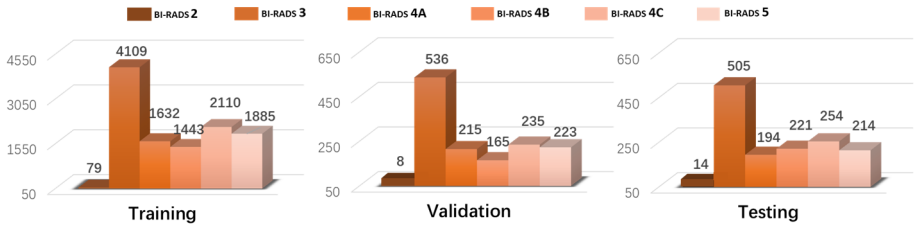

Fig. 3. BI-RADS distribution at image-level.

Implementation Details. The region of a lesion with its 60-pixel border was cropped to reduce the irrelevant background. The cropped images were converted to square by adding 0 to shorter edges, and then resized to 224×224. Both teacher and student models used the same dataset distributions and parameter settings. The pretrained RepVGG-A2 was adopted and the optimizer was set as SGD with a learning rate of 0.01 in the training stage. Random horizontal flipping, scaling, and brightness and contract transformation were utilized as data augmentation. We trained the model for 100 epochs with a batch size of 128. The model that performed with the highest AUC on validation set was chosen as the final model for testing.

Results Analysis. We measured the AUC (AUC^P), Accuracy (ACC), Sensitivity (SENS), Specificity (SPEC), positive predictive value (PPV), and negative predictive value (NPV) of pathology classification as well as AUC (AUC^B) and Kappa of BI-RADS categorization to evaluate the classification performance of the proposed method. The novel VGG19 [11], ResNet18 [12], EfficientNet [13], DenseNet [14] and RepVGG-A2 were re-implemented with two classification branches and evaluated on the same dataset for comparison. Results are shown in Table 2. Our method outperforms all these algorithms in all above metrics.

Table 2. Performance comparison on our test set (%).

Method	AUC^P	ACC	SENS	SPEC	PPV	NPV	AUC^B	Kappa
VGG19	0.897	0.809	0.750	0.867	0.847	0.780	0.880	0.544
ResNet18	0.893	0.803	0.736	0.873	0.858	0.760	0.861	0.577
EfficientNet	0.876	0.799	0.738	0.859	0.839	0.768	0.842	0.548
DenseNet	0.905	0.833	0.819	0.843	0.795	0.863	0.863	0.579
RepVGG-A2	0.908	0.835	0.814	0.850	0.809	0.855	0.868	0.622
Ours	**0.958**	**0.897**	**0.865**	**0.921**	**0.895**	**0.897**	**0.931**	**0.677**

We also test our method on BUSI [15] and UDIAT [16] to compare with others. The results are shown in Table 3 (Italics: test in a subset of the dataset). Our method outperforms [3], [4], [7], and [17] in most metrics on both datasets.

Table 3. Performance comparison on BUSI and UDIAT.

Dataset	Method	AUC^P	ACC	SENS	SPEC	PPV	NPV	AUC^B
BUSI	Shen et al [3]	0.927	-	0.905	0.842	0.672	0.949	-
	Xing et al [7]	0.889	0.843	0.758	0.883	0.751	-	0.832
	Ours	0.900	**0.859**	0.735	**0.916**	**0.803**	0.881	**0.884**
UDIAT	*Zhang et al [17]*	*0.889*	*0.92*	-	-	-	-	-
	Byra et al [4]	*0.893*	*0.840*	*0.851*	*0.834*	-	-	-
	Xing et al [7]	0.870	0.859	0.685	0.945	0.860	-	0.872
	Ours	**0.905**	0.877	0.685	**0.972**	**0.925**	**0.862**	**0.916**

Ablation Study. A series of ablation experiments were conducted to validate the effectiveness of the proposed methods. The results are shown in Table 4 and Fig. 4. Comparing with teacher model (RepVGG-A2 in Table 2), the task-correlated soft labels make the prediction results improved about 2% over all metrics on both classification tasks (S). Based on S, the AUC^P and AUC^B increase by 1.2% ($p = 0.0160$) and 3.7% ($p < 0.0001$) respectively with the help of CSM. After using the CCLF, not only AUC^P and AUC^B but also accuracy (increased by around 4%) and sensitivity (grown by over 2.5%) are risen comparing to S+CSM. The final student network (S+CSM+CCLF) outperforms the teacher model in a huge margin over most metrics demonstrating the advantage of our method.

Table 4. Results of ablation studies.

Method	AUC^P	ACC	SENS	SPEC	PPV	NPV	AUC^B	Kappa
Teacher	0.908	0.835	0.814	0.850	0.809	0.855	0.868	0.622
Student (S)	0.920	0.852	0.832	0.867	0.830	0.869	0.882	0.635
P value	0.0318	-	-	-	-	-	0.1186	-
S+CSM	0.932	0.859	0.839	0.876	0.840	0.875	0.919	0.642
P value	0.0160	-	-	-	-	-	<0.0001	-
S+CSM+CCLF	**0.958**	**0.897**	**0.865**	**0.921**	**0.895**	**0.897**	**0.931**	**0.677**
P value	<0.0001	-	-	-	-	-	0.0991	-

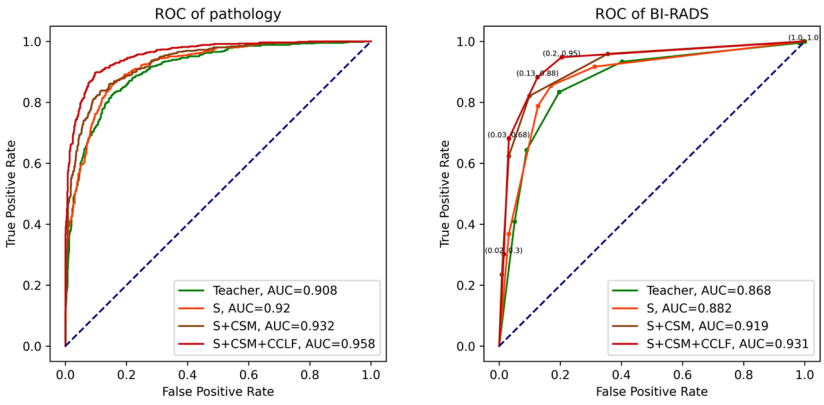

Fig. 4. ROC comparison among ablation studies.

Fig. 5 compares the class activation maps [18] (CAMs) between teacher model and final student model. The CAMs of two tasks in the final model (the last two columns) are more consistent than in the teacher model (the second and third column), displaying that the proposed method successfully utilized the relevance between the two tasks. The first two rows show that the teacher model misclassifies a benign lesion with annotation of BI-RADS 3 and a malignant lesion labeled as BI-RADS 4c in term

of pathologic classes but the final student model does not, which demonstrates that our method is able to correct misclassification by constraining the corelation between the pathologic classes and BI-RADS categories. The third and fourth rows present that the BI-RADS category predicted by student model is closer to the annotation than that by teacher model indicating the effectiveness of cross-class function.

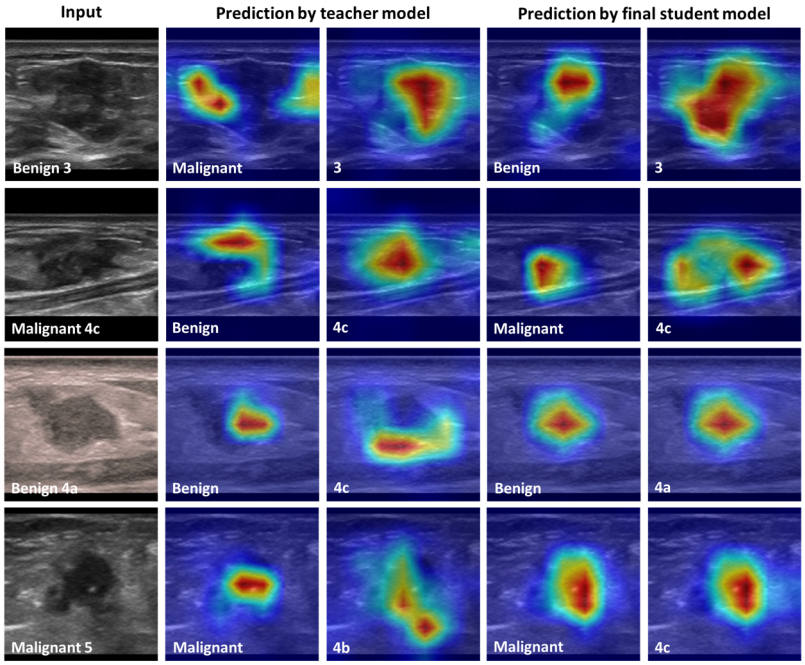

Fig. 5. Comparison of class activation maps.

4 Conclusion

In conclusion, we proposed a novel deep learning system for benign and malignant classification and for BI-RADS categorization simultaneously to assist clinical diagnosis. The task-correlated soft labels successfully improved the classification performance, demonstrating the effectiveness of the multitask label generating architecture. Moreover, the consistency supervision mechanism guaranteed that the prediction of BI-RADS category was consistent with the predicted pathology result, meanwhile the cross-class loss function improved the classification accuracies by utilizing different weights to penalize different degrees of misclassified items. Furthermore, the experiment results on two public datasets indicated the great potential of our method in clinical diagnosis. In the future, we will apply our method for thyroid nodule risk stratification in ultrasound images.

References

1. Sung H., Ferlay J., Siegel RL., Laversanne M., Soerjomataram I., Jemal A., Bray F.: Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin, 71(3), 209-249 (2021)
2. Medelson EB., Böhm-Vélez M., Berg W.A., et al.: ACR BI-RADS® Ultrasound. In ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA, American College of Radiology. (2013)
3. Shen Y., Shamout FE., Oliver JR., et al.: Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nat Commun*, 12(1), 5645 (2021)
4. Byra M., Galperin M., Ojeda-Fournier H., Olson L., O'Boyle M., Comstock C., Andre M.: Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med Phys, 46(2), 746-755 (2019)
5. Qian X., Zhang B., Liu S., Wang Y., Chen X., Liu J., Yang Y., Chen X., Wei Y., Xiao Q., Ma J., Shung KK., Zhou Q., Liu L., Chen Z.: A combined ultrasonic B-mode and color Doppler system for the classification of breast masses using neural network. Eur Radiol, 30(5), 3023-3033 (2020)
6. Liu J., et al.: Integrate Domain Knowledge in Training CNN for Ultrasonography Breast Cancer Diagnosis. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, vol 11071 (2018)
7. Xing J., Chen C., Lu Q., Cai X., Yu A., Xu Y., Xia X., Sun Y., Xiao J., Huang L.: Using BI-RADS Stratifications as Auxiliary Information for Breast Masses Classification in Ultrasound Images. IEEE J Biomed Health Inform. 25(6), 2058-2070 (2021)
8. Hinton., Geoffrey., Oriol Vinyals., Jeff Dean.: "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* 2.7 (2015)
9. Li Y., Kazameini A., Mehta Y., et al.: Multitask learning for emotion and personality detection[J]. arXiv preprint arXiv:2101.02346 (2021)
10. X. Ding., X. Zhang., N. Ma., J. Han., G. Ding. and J. Sun.: RepVGG: Making VGG-style ConvNets Great Again. *CVPR*, pp. 13728-13737 (2021)
11. Simonyan, Karen., Andrew Zisserman.: Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR,* abs/1409.1556 (2015)
12. K. He., X. Zhang., S. Ren., J. Sun.: Deep Residual Learning for Image Recognition. *CVPR*, pp. 770-778 (2016)
13. Tan., Mingxing., Quoc V. Le.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ArXiv* abs/1905.11946 (2019)
14. G. Huang., Z. Liu., L. Van Der Maaten., K. Q. Weinberger.: Densely Connected Convolutional Networks. *CVPR*, pp. 2261-2269 (2017)
15. W. Al-Dhabyani., M. Gomaa., H. Khaled., A. Fahmy.: Dataset of breast ultrasound images. Data in brief, vol. 28, p. 104863 (2020)
16. M. H. Yap., G. Pons., et al.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics, vol. 22, no. 4, pp. 1218–1226 (2017)
17. Zhang E., Seiler S., Chen M., Lu W., Gu X.: Boundary-aware Semi-supervised Deep Learning for Breast Ultrasound Computer-Aided Diagnosis. Annu Int Conf IEEE Eng Med Biol Soc, 2019:947-950 (2019)
18. Selvaraju., Ramprasaath R., Michael Cogswell., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. IEEE international conference on computer vision, pp. 618-626. (2017).