Impact of Ultrasound Image Reconstruction Method on Breast Lesion Classification with Deep Learning

Michal Byra[1](✉), Tomasz Sznajder[2], Danijel Korzinek[2], Hanna Piotrzkowska-Wroblewska[1], Katarzyna Dobruch-Sobczak[1], Andrzej Nowicki[1], and Krzysztof Marasek[2]

[1] Department of Ultrasound, Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland
mbyra@ippt.pan.pl

[2] Department of Multimedia, Polish-Japanese Academy of Information Technology, Warsaw, Poland

Abstract. In this work we investigate the usefulness and robustness of transfer learning with deep convolutional neural networks (CNNs) for breast lesion classification in ultrasound (US). Deep learning models can be vulnerable to adversarial examples, engineered input image pixel intensities perturbations that force models to make classification errors. In US imaging, distribution of US image pixel intensities relies on applied US image reconstruction algorithm. We explore the possibility of fooling deep learning models for breast mass classification by modifying US image reconstruction method. Raw radio-frequency US signals acquired from malignant and benign breast masses were used to reconstruct US images, and develop classifiers using transfer learning with the VGG19, InceptionV3 and InceptionResNetV2 CNNs. The areas under the receiver operating characteristic curve (AUCs) obtained for each deep learning model developed and evaluated using US images reconstructed in the same way were equal to approximately 0.85, and there were no associated differences in AUC values between the models (DeLong test p-values > 0.15). However, due to small modifications of the US image reconstruction method the AUC values for the models utilizing the VGG19, InceptionV3 and InceptionResNetV2 CNNs significantly decreased to 0.592, 0.584 and 0.687, respectively. Our study shows that the modification of US image reconstruction algorithm can have significant negative impact on classification performance of deep models. Taking into account medical image reconstruction algorithms may help develop more robust deep learning computer aided diagnosis systems.

Keywords: Adversarial attacks · Breast lesion classification · Computer aided diagnosis · Deep learning · Robustness · Ultrasound imaging · Transfer learning

Supported by the National Science Center (Poland), Grant Number UMO-2014/13/B/ST7/01271.

A. Morales et al. (Eds.): IbPRIA 2019, LNCS 11867, pp. 41–52, 2019.
https://doi.org/10.1007/978-3-030-31332-6_4

1 Introduction

Ultrasound (US) imaging is widely used for breast mass detection and differentiation in clinics. However, US data acquisition needs to be carried out by an experienced radiologist or physician who knows how to efficiently operate the ultrasound scanner. The operator has to locate the mass within the examined breast and properly record US images. Moreover, interpretation of the US images is not straightforward, but requires deep knowledge of characteristic image features related to breast mass malignancy.

Various computer-aided diagnosis (CADx) systems have been proposed to support the radiologists and improve differentiation of malignant and benign breast masses [6,10,11]. Currently, with the rise of deep learning methods, CADx systems based on convolutional neural networks (CNNs) are gaining momentum for breast mass classification [2–4,13,18,26]. These networks process input images using convolutional filters to learn useful data representations and provide the desired output, such as a single binary decision related to the presence of particular object in the input image. However, better performing deep CNNs were developed using large sets of natural images [8]. Since medical image datasets are usually too small to train efficient CNNs from scratch, transfer learning methods are applied to develop deep learning models [20]. The aim of the transfer learning techniques is to employ a CNN model pre-trained on a large dataset of images from a different domain to address the medical image analysis problem of interest. In the case of the breast mass classification, deep models pre-trained on natural images were used to extract high level image features and utilize those to train binary classifiers, such as logistic regression or support vector machine algorithm [2–4].

In this paper we assess the usefulness of several deep learning models for transfer learning based breast mass classification. In comparison to the previous studies we investigate the impact of US B-mode image reconstruction algorithm on the classification performance [2–4,13,18,26]. Our work is motivated by several studies reporting that deep learning systems can be vulnerable to adversarial examples, input images engineered to cause misclassification due to complex nonlinear behaviors of deep models [9]. Adversarial attacks can be performed by, for example, adding small artificially crafted perturbations to input image pixel intensities, which slightly modifies appearance of objects' edges and texture, and force deep model to perform wrong classification [12,15,16]. In medical image analysis, the vulnerability of deep learning models to adversarial attacks was demonstrated in the case of chest X-rays and dermoscopy images [9], raising concerns about the robustness of CADx systems based on CNNs [24]. Appearance of tissues in US imaging is related to applied image reconstruction algorithm. US scanners record raw radio-frequency (RF) backscattered signals and process them to reconstruct B-mode images. During routine US scanning the operator can modify scanner settings to differently reconstruct B-mode images to enhance specific B-mode image features. Due to high dynamic range RF US signals are commonly non-linearly compressed before B-mode image reconstruction. Modifications of the compression level result in different brightness levels of tissue interfaces and different speckle patterns. Here, we investigate the impact of US image reconstruction algorithm on breast mass classification with deep learning. We study whether small modifications of compression threshold levels related to applied B-mode image reconstruction may cause CNN based models to make classification errors.

2 Materials and Methods

2.1 Dataset

To develop deep learning models for breast mass classification we used an extension of the freely available breast mass dataset, the OASBUD (Open Access Series of Breast Ultrasonic Data) [5,17], which includes RF US data (before B-mode image reconstruction) recorded from breast focal masses during routine scanning performed in the Maria Skłodowska-Curie Memorial Cancer Centre and Institute of Oncology in Warsaw. The study was approved by the Institutional Review Board. The data were collected using the Ultrasonix Sonix-Touch Research ultrasound scanner with an L14-5/38 linear array transducer. The dataset includes RF signals recorded from 231 breast masses, 82 masses were malignant and 149 masses were benign. All malignant masses were histologically assessed by core needle biopsy. Benign masses were assessed either by the biopsy or a two year observation (every six months). For each scan a region of interest was determined by an experienced radiologist to correctly indicate breast mass area in B-mode image. More details regarding the dataset can be found in the original paper [17].

2.2 Ultrasound Image Reconstruction

Reconstruction scheme of a single B-mode image line is presented in Fig. 1. First, the RF signal acquired by the transducer is used to detect the envelope with the Hilbert transform. Second, since the dynamic range of US signal amplitudes is too high to fit on the screen directly, the amplitude samples are logarithmically compressed. In this work we used the following formula to compress amplitude samples:

$$A_{log} = 20log_{10}(A/A_{max}) \tag{1}$$

where A and A_{log} are the amplitude and the log-compressed amplitude of the ultrasonic signal, respectively. A_{max} indicates the highest value of the amplitude in the data. Next, the compressed amplitude samples are mapped to B-mode image pixel intensities based on a specified threshold level. Figure 2 shows three B-mode images of benign and malignant breast masses reconstructed using threshold levels of 45 dB, 50 dB and 55 dB, which are typically used in practice. Moreover, Fig. 3 shows the RF signal amplitude to pixel intensity mapping functions for these three different threshold levels. Physicians commonly select the threshold level to obtain desired image quality e.g. good speckle pattern

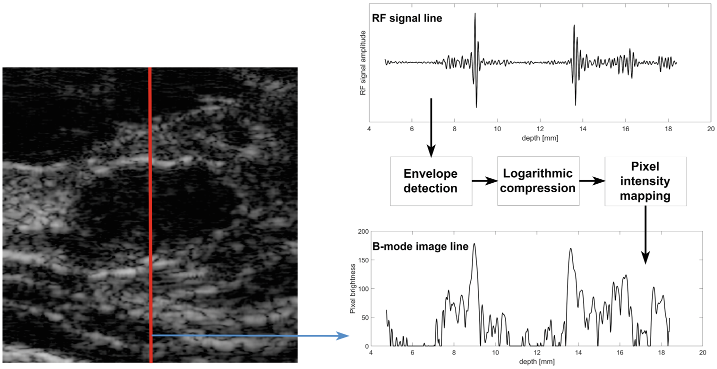

Fig. 1. Pipeline illustrating reconstruction of a single B-mode image line based on a radio-frequency ultrasound signal acquired by the transducer. The scheme includes envelope detection, logarithmic compression and mapping of compressed amplitude samples to B-mode image pixel intensities.

visibility or edge enhancement. For example, setting low threshold level results in removal of speckles that originates from US echoes of low intensities. Setting high threshold level may result in removal of important edge details.

2.3 Transfer Learning with Convolutional Neural Networks

We used three deep CNNs to perform transfer learning and classify breast masses, namely the VGG19, InceptionV3 and InceptionResNetV2 [14,22,23], all pre-trained on the ImageNet dataset [8] and implemented in TensorFlow [1]. These models achieved good performance on the ImageNet dataset and were used for breast mass classification with transfer learning in the previous studies [2,4,13]. In this work, we employed one of the most widely used transfer learning approaches, which aims to extract high level neural features from the last layers of the pre-trained model and use those to develop a classifier. In the case of the VGG19 CNN, we extracted features from the first fully connected layer. Moreover, average pooling layers of the InceptionV3 and InceptionResNetV2 CNNs were used to extract neural features.

2.4 Experiments and Evaluation

We performed several experiments to evaluate the usefulness of each CNN for the breast mass classification, and to explore the possibility of fooling the models by the compression threshold level modification. The experimental setup is presented in Fig. 4. We selected average compression threshold level of 50 dB and investigated how small perturbations (in range from 45 dB to 55 dB) can affect the classification. To assess the classification performance we applied leave-one-out cross validation. For each cross validation round, B-mode images in the training set were reconstructed using compression threshold level of 50 dB. In

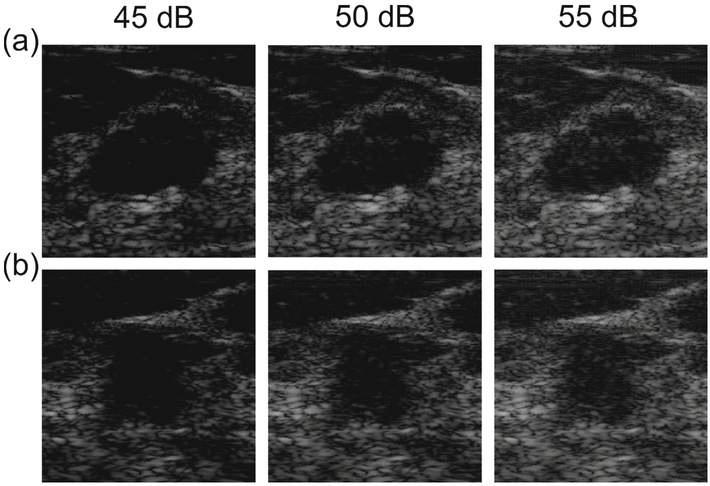

Fig. 2. B-mode images of (a) benign and (b) malignant masses reconstructed using compression threshold levels of 45 dB, 50 dB and 55 dB, respectively.

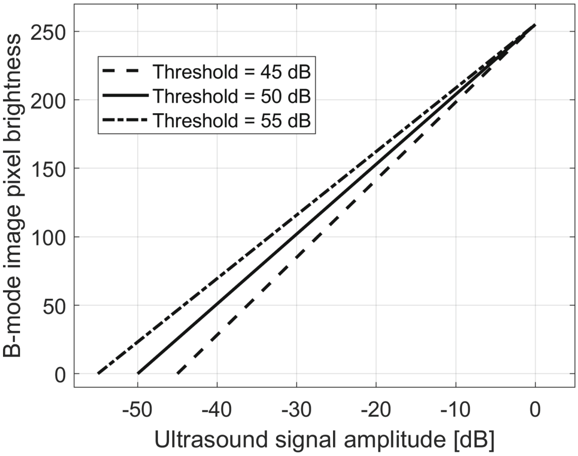

Fig. 3. B-mode image pixel brightness mapping function for logarithmic compression using compression threshold levels of 45 dB, 50 dB and 55 dB, respectively. Small modifications of the threshold level result in small change of B-mode image pixel intensities.

the case of the first experiment, each test B-mode image was reconstructed in the same way as those in the training set, using the threshold level of 50 dB. Therefore the perturbations were not applied for the first experiment. Next, to explore the possibility of fooling the models, we performed the second experiment. Again, all training B-mode images were reconstructed using the threshold level of 50 dB. But this time we reconstructed each test B-mode image using different threshold levels, ranging from 45 dB to 55 dB. Each classification model

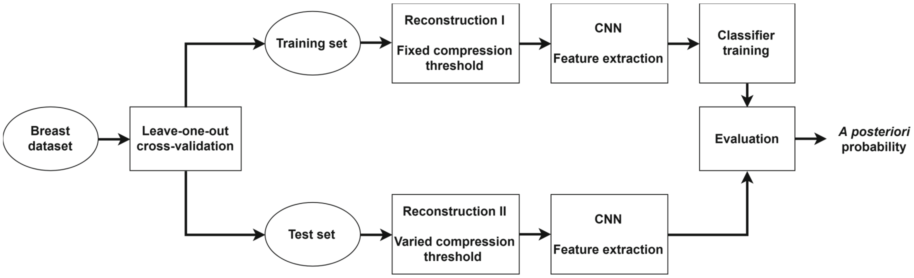

Fig. 4. Pipeline illustrating the experiments performed in our study. B-mode images for training were reconstructed using fixed compression threshold level of 50 dB. In the case of the test set, for the first experiment B-mode images were reconstructed using threshold level 50 dB, for the second (third) experiment the threshold was selected to maximally decrease (increase) the classification performance of each deep learning model. CNN - convolutional neural network.

developed on the training set was evaluated using all differently reconstructed test B-mode images, and we selected the B-mode image corresponding to the worst possible classification performance. If the test breast mass was malignant (benign), then we selected the B-mode image corresponding to the lowest (highest) obtained *a posteriori* probability of malignancy determined by the model. Studies on adversarial attacks in deep learning usually focus on efficient engineering of adversarial examples that would result in classification errors. In comparison to those studies, we also explored the possibility of using B-mode image perturbations to increase deep learning model classification performance. While the second experiment corresponded to the worst possible scenario, the third experiment corresponded to the best possible scenario. This time the test B-mode images were perturbed with the aim to increase classification performance.

To extract features for classification from deep CNNs we applied the following approach. Each B-mode image was cropped using the region on interest provided by the radiologist to contain the mass and a 5 mm band of surrounding tissues, see Fig. 2. Next, the US images were resized using bi-cubic interpolation to match the resolution originally designed for each neural network, 224×224 for the VGG19 CNN and 299×299 for the two other CNNs. Intensities of each image were copied along RGB channels and preprocessed in the same way as in the original papers [21–23]. The same approach utilizing the VGG19 CNN was employed in the previous studies on breast mass classification with transfer learning [2,3,13]. To perform binary classification we used the logistic regression algorithm. To address the problem of class imbalance, we used class weights inversely proportional to class frequencies in the training set. We used a linear classifier to omit possible issues related to the properties of non-linear classifiers, which could introduce additional non-linearity behaviors to the models in addition to those already related to deep CNNs.

To asses the classification performance we calculated the receiver operating characteristic (ROC) curves using model outputs obtained in each experiment. Next, we determined the areas under the ROC curves (AUC) for different models, the sensitivity, specificity and accuracy of the classifiers were calculated based on the ROC curve for the point on the curve that was the closest to (0, 1). The AUC value of 0.5 in the case of binary classification indicates random guessing, while the AUC value of 1 correspond to perfect classification. The AUC values of different models were compared with the DeLong test [7,19]. All calculations were performed in a programming environment including Python, R and Matlab (Mathworks, USA).

3 Results

Table 1 summarizes the classification performances obtained in all three experiments. In the case of the first experiment, for the test B-mode images reconstructed in the same way as the training images, the classification models achieved AUC values of 0.858, 0.829 and 0.860 for the VGG19, InceptionV3 and InceptionResNetV2 CNNs, respectively. There were no associated statistical differences between the AUC values obtained for the models developed using different deep CNNs (DeLong test p-values > 0.15).

In the case of the second experiment, based on the B-mode image reconstruction method modification we were able to decrease classification performance of each deep learning model. Results presented in Table 1 show that the AUC values significantly decreased (DeLong test p-values < 0.05). For the VGG19, InceptionV3 and InceptionResNet CNNs the AUC values were equal to 0.592, 0.584 and 0.687, respectively. The model trained based on features extracted from the InceptionResNetV2 CNN was less vulnerable to B-mode image modification than the other models. Figure 5 shows four adversarial examples engineered with our approach corresponding to two malignant and two benign breast masses. For example, benign breast mass present in Fig. 5a) was correctly classified as benign by all models, the corresponding *a posteriori* probabilities of malignancy were equal to 0.31, 0.38 and 0.23 for the models developed using VGG19, InceptionV3 and InceptionResNet CNNs, respectively. Due to the reconstruction threshold value modification, the corresponding probabilities increased to 0.62, 0.68 and 0.36, what caused classification errors in the case of the models developed using the VGG19 and InceptionV3 CNNs. The adversarial examples in Fig. 5 are very similar to the original B-mode images, with only slightly modified edge visibility and speckle patterns.

Additionally, Table 1 shows the results obtained in the case of the third experiment, which aimed to maximally increase classification performance by perturbing B-mode image pixel intensities. The AUC values for the VGG19, InceptionV3 and InceptionResNet CNNs significantly increased (DeLong test p-values < 0.05) to 0.970, 0.961 and 0.963, respectively.

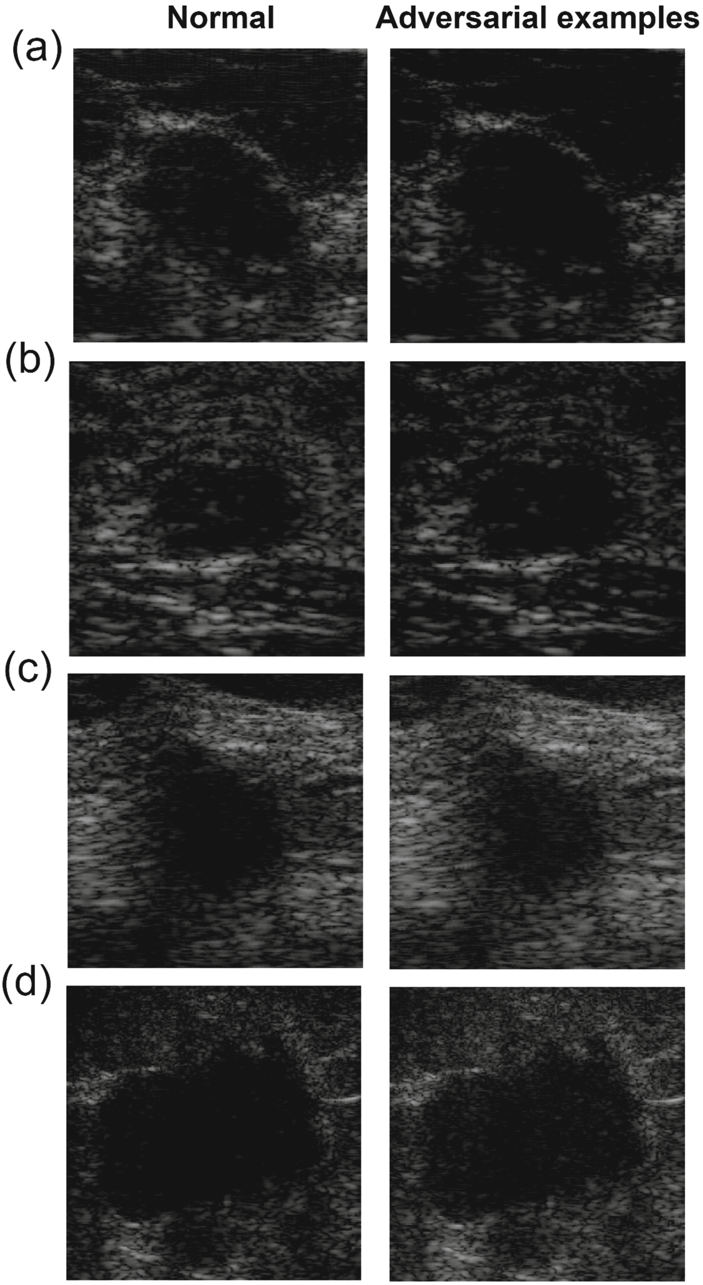

Fig. 5. Correctly classified breast mass B-mode images reconstructed using compression threshold level of 50 dB and corresponding B-mode images reconstructed to cause misclassification, (a), (b) benign masses and (c), (d) malignant masses.

Table 1. Classification performance of each deep learning model developed using transfer learning. The regular results were obtained for the models developed and evaluated using train and test B-mode images reconstructed in the same way. The worst (best) results were determined for the test B-mode images perturbed with the aim to maximally decrease (increase) classification performance. AUC - area under the receiver operating characteristic curve, standard deviations were calculated using bootstrap.

Network	Type	AUC	Accuracy	Sensitivity	Specificity
VGG19	Regular	0.858 ± 0.027	0.822 ± 0.038	0.768 ± 0.038	0.852 ± 0.034
	Worst	0.592 ± 0.034	0.649 ± 0.023	0.548 ± 0.040	0.7018 ± 0.030
	Best	0.970 ± 0.007	0.926 ± 0.015	0.890 ± 0.022	0.946 ± 0.025
InceptionV3	Regular	0.829 ± 0.028	0.757 ± 0.023	0.7682 ± 0.038	0.752 ± 0.038
	Worst	0.584 ± 0.030	0.584 ± 0.027	0.573 ± 0.062	0.590 ± 0.061
	Best	0.961 ± 0.008	0.896 ± 0.017	0.878 ± 0.026	0.901 ± 0.027
InceptionResNetV2	Regular	0.860 ± 0.026	0.792 ± 0.023	0.768 ± 0.038	0.801 ± 0.033
	Worst	0.687 ± 0.028	0.692 ± 0.028	0.573 ± 0.044	0.59 ± 0.049
	Best	0.963 ± 0.009	0.926 ± 0.023	0.890 ± 0.035	0.946 ± 0.032

4 Discussion

Our study shows the usefulness of the transfer learning with deep CNNs for breast mass classification in US. The model based on InceptionResNetV2 CNN achieved AUC value of 0.860. Our results are in agreement with those reported in the previous studies on breast mass classification with deep learning [2–4], where the authors obtained AUC values in range from 0.79 to 0.90. In [13] a specific approach to transfer learning was applied, which included fine-tuning and modification of the InceptionV3 architecture and ImageNet dataset. The authors used an ensemble of deep models for classification and reported high AUC value of 0.960. In our case, we used the InceptionV3 model for transfer learning in a more standard way following the approach proposed in [2].

Classification performance of all three developed deep learning models was sensitive to B-mode image reconstruction modifications. The decrease in classification performance was significant for all models, with the largest decrease obtained for the models developed using features extracted from the VGG19 and InceptionV3 CNNs (AUC values of 0.592 and 0.584). The model trained based on InceptionResNetV2 features was less vulnerable to US image reconstruction method modification (AUC value of 0.687). Figure 5 shows that the adversarial examples are very similar visually to the B-mode images reconstructed using threshold level of 50 dB. In comparison to the previous studies investigating how to engineer successful adversarial attacks [9], we additionally explored the possibility of manipulating image pixel intensities to artificially improve breast mass classification. By modifying the B-mode image reconstruction method we improved the performance of all models and achieved AUC values of around 0.97.

Our study depicts several important issues related to the development of CADx systems using transfer learning with deep pre-trained CNNs. First of all, the image reconstruction procedures implemented in medical scanners should be taken into account during CADx system development. It is important to know how B-mode images were acquired and reconstructed. Classification errors may result from issues related to applied B-mode image reconstruction methods, such as using non-standard scanner settings. To improve performance and make deep learning models more robust it might be necessary to develop the models based on B-mode images acquired using different scanner settings. The second possibility is to always use the same image reconstruction algorithms and scanner setting for B-mode image acquisition. In our study we used a unique dataset of RF signals collected with a research US scanner. Regular clinical US scanners, however, usually don't have access to RF data, and such data are not stored in hospital databases. Researchers, who would like to develop deep learning models based on large sets of retrospectively collected B-mode images extracted from a hospital database should take into account what apparatus and procedures were used to scan the patients. Unfortunately, usually little is known about the applied B-mode image reconstruction algorithms implemented by different US scanner manufacturers.

There are several issues related to our approach, which should be addressed in future. First, to develop the models we used one of the most widely used, but relatively simple, transfer learning method. In this case the pre-trained deep CNNs were used as fixed feature extractors. It remains to be studied whether deep learning models developed from scratch would be similarly vulnerable to B-mode image reconstruction method modifications. Second, we only explored the possibility of fooling models based on the modification of compression threshold levels, but it is also possible to modify other parameters related to the B-mode image reconstruction method. For example, perturbations of B-mode image pixel intensities can also arise from setting different logarithm base for compression. Moreover, the texture of B-mode images depends on applied beamforming technique [27] and imaging frequency [25]. Nevertheless, in the case of our study it was sufficient to modify compression threshold values to significantly change classification performance of the deep learning models.

5 Conclusions

In this work we investigated the impact of B-mode image reconstruction method on breast mass classification with deep learning. By modifying B-mode image reconstruction method we were able to significantly decrease or increase classification performance of each deep learning classifier. We believe that our work is an important step towards the development of robust deep learning computer aided diagnosis systems.

Conflict of interest statement. The authors do not have any conflicts of interests.

References

1. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning. OSDI **16**, 265–283 (2016)
2. Antropova, N., Huynh, B.Q., Giger, M.L.: A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med. Phys. **44**, 5162–5171 (2017)
3. Byra, M.: Discriminant analysis of neural style representations for breast lesion classification in ultrasound. Biocybern. Biomed. Eng. **38**(3), 684–690 (2018)
4. Byra, M., et al.: 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. Byra, M., Nowicki, A., Wróblewska-Piotrzkowska, H., Dobruch-Sobczak, K.: Classification of breast lesions using segmented quantitative ultrasound maps of homodyned K distribution parameters. Med. Phys. **43**(10), 5561–5569 (2016)
6. Cheng, H.D., Shan, J., Ju, W., Guo, Y., Zhang, L.: Automated breast cancer detection and classification using ultrasound images: a survey. Pattern Recognit. **43**(1), 299–317 (2010). https://doi.org/10.1016/j.patcog.2009.05.012
7. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44**(3), 837–845 (1988)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
9. Finlayson, S.G., Kohane, I.S., Beam, A.L.: Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296 (2018)
10. Flores, W.G., de Albuquerque Pereira, W.C., Infantosi, A.F.C.: Improving classification performance of breast lesions on ultrasonography. Pattern Recognit. **48**(4), 1125–1136 (2015)
11. Giger, M.L., Karssemeijer, N., Schnabel, J.A.: Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annu. Rev. Biomed. Eng. **15**, 327–357 (2013)
12. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
13. Han, S., et al.: A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys. Med. Biol. **62**(19), 7714 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
15. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)
16. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427–436 (2015)
17. 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. **44**(11), 6105–6109 (2017)
18. Qi, X., et al.: Automated diagnosis of breast ultrasonography images using deep neural networks. Med. Image Anal. **52**, 185–198 (2019)

19. Robin, X., et al.: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform. **12**, 77 (2011)
20. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
24. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. **25**(1), 44 (2019)
25. Tsui, P.H., Zhou, Z., Lin, Y.H., Hung, C.M., Chung, S.J., Wan, Y.L.: Effect of ultrasound frequency on the nakagami statistics of human liver tissues. PLoS ONE **12**(8), e0181789 (2017)
26. Yap, M.H., et al.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Health Inform. **22**, 1218–1226 (2017)
27. Yu, X., Guo, Y., Huang, S.M., Li, M.L., Lee, W.N.: Beamforming effects on generalized Nakagami imaging. Phys. Med. Biol. **60**(19), 7513 (2015)