Cite as: Byra M, Dobruch-Sobczak K, Piotrzkowska-Wroblewska H, Klimonda Z, Litniewski J: Explaining a deep learning based breast ultrasound image classifier with saliency maps. J Ultrason 2022; 22: e70–e75. doi: 10.15557/JoU.2022.0013.

Submitted: 14.11.2021
Accepted: 24.01.2022
Published: 13.04.2022

Explaining a deep learning based breast ultrasound image classifier with saliency maps

Michał Byra[1], Katarzyna Dobruch-Sobczak[2], Hanna Piotrzkowska-Wroblewska[1], Ziemowit Klimonda[1], Jerzy Litniewski[1]

[1] *Department of Ultrasound, Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland*
[2] *Radiology Department II, Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland*

Correspondence: Michał Byra, Department of Ultrasound, Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5b, 02-106 Warsaw, Poland; e-mail: byra.michal@gmail.com

DOI: 10.15557/JoU.2022.0013

Keywords
deep learning; breast mass diagnosis; attention maps; explainability

Abstract

Aim of the study: Deep neural networks have achieved good performance in breast mass classification in ultrasound imaging. However, their usage in clinical practice is still limited due to the lack of explainability of decisions conducted by the networks. In this study, to address the explainability problem, we generated saliency maps indicating ultrasound image regions important for the network's classification decisions. **Material and methods:** Ultrasound images were collected from 272 breast masses, including 123 malignant and 149 benign. Transfer learning was applied to develop a deep network for breast mass classification. Next, the class activation mapping technique was used to generate saliency maps for each image. Breast mass images were divided into three regions: the breast mass region, the peritumoral region surrounding the breast mass, and the region below the breast mass. The pointing game metric was used to quantitatively assess the overlap between the saliency maps and the three selected US image regions. **Results:** Deep learning classifier achieved the area under the receiver operating characteristic curve, accuracy, sensitivity, and specificity of 0.887, 0.835, 0.801, and 0.868, respectively. In the case of the correctly classified test US images, analysis of the saliency maps revealed that the decisions of the network could be associated with the three selected regions in 71% of cases. **Conclusions:** Our study is an important step toward better understanding of deep learning models developed for breast mass diagnosis. We demonstrated that the decisions made by the network can be related to the appearance of certain tissue regions in breast mass US images.

Introduction

Breast cancer is the most frequent cancer in women worldwide(1). Ultrasound (US) imaging is commonly used by radiologists to characterize breast masses and conduct a diagnosis. However, accurate differentiation between malignant and benign breast masses requires knowledge about the characteristic US image features associated with the malignancy, which is why it is considered difficult. Malignant breast masses are commonly characterized by indistinct and highly variable contours as well as the presence of shadowing artifacts and calcifications. Various deep learning methods have been proposed over the last years to help with the breast mass classification(2,3). Deep convolutional neural networks (CNNs) can automatically process input US images to determine classification decisions. Although deep networks achieved excellent performance in breast mass classification, their usage in clinical practice is still limited due to the lack of model interpretability and explainability. Neural networks are commonly perceived as 'black-box' models, which raises questions about their applicability in medicine(4,5).

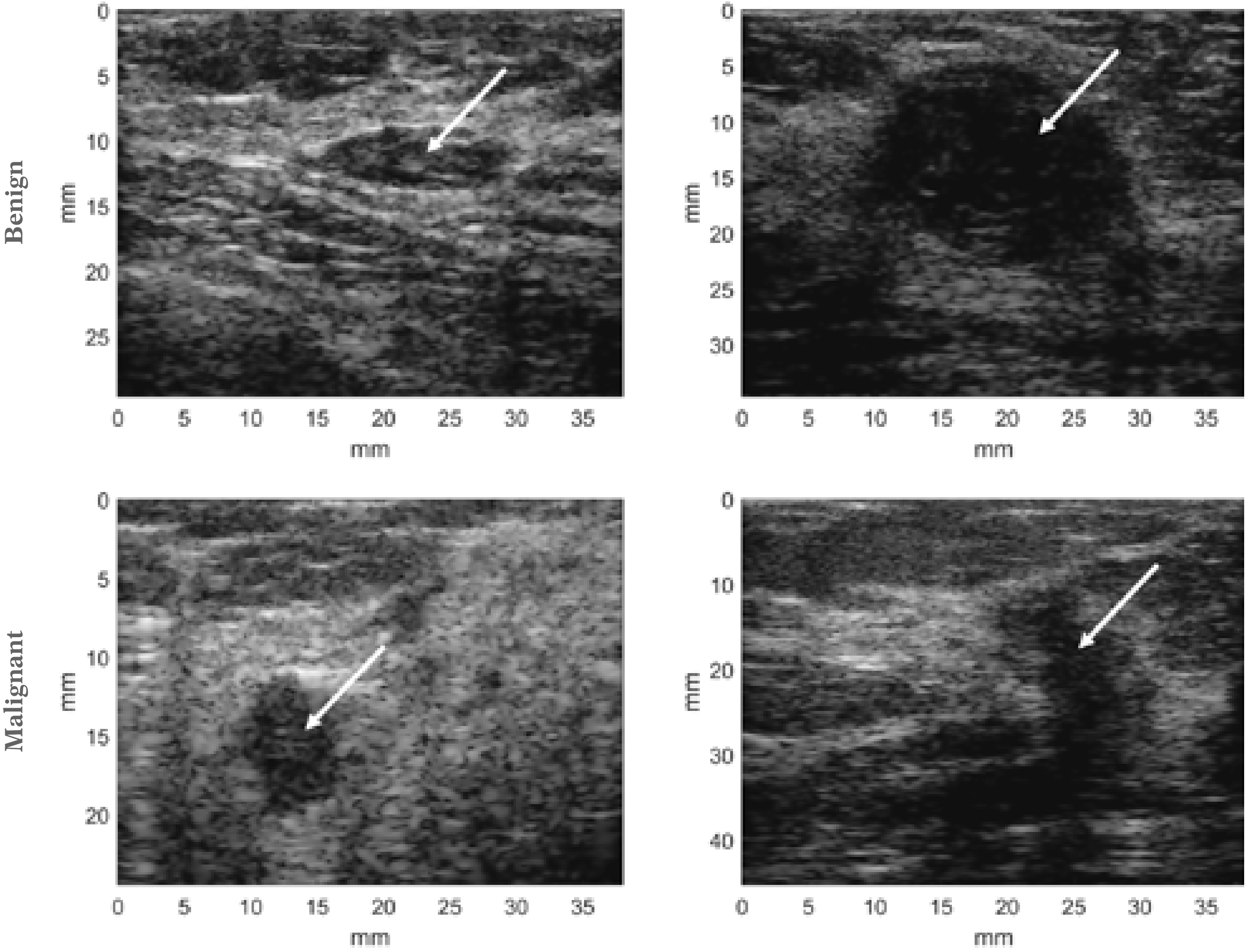

Fig. 1. *Exemplary US images presenting benign and malignant breast masses*

In this study, we address the problem of explainability of deep networks developed for breast mass classification in US imaging. First, we developed a CNN for breast mass classification using transfer learning. Second, we used the class activation mapping (CAM) technique to generate the network's saliency maps for each breast mass US image. Saliency maps indicate which image regions activated the network. Therefore, these maps can be used to visually explain the decisions conducted by the model. In this study, we divided the US images into three regions commonly adopted by radiologists to visually assess breast masses. Next, we utilized the pointing game metric to quantitatively associate the three regions with the saliency maps generated by the network[6].

Materials and methods

Dataset

The study was approved by the Institutional Review Board. All medical procedures involving human participants were performed in accordance with the guidelines stated in the Declaration of Helsinki and its later amendments or comparable ethical standards. US images were collected from 272 breast masses, of which 123 masses were malignant and 149 were benign. The average age of the patients diagnosed with malignant and benign masses was 54.4 and 40.8 years, respectively. Two orthogonal scans (longitudinal and transverse) were collected for each breast mass using the Ultrasonix SonixTouch Research US scanner (Ultrasonix Inc., Canada) equipped with the L14-5/38 linear probe operating at 10 MHz[7]. The radiologists who collected the data assessed each case using the Breast Imaging-Reporting and Data System (BI-RADS). A total of 88 masses corresponded to BI-RADS category 3, 123 to BI-RADS 4, and 61 to BI-RADS 5 according to the American College of Radiology Atlas and the guidelines of the Polish Ultrasound Society[8,9]. Lesions corresponding to the BI-RADS categories 4 and 5 were subjected to core needle biopsy. In the case of BI-RADS 3 masses, part of them were assessed using fine needle aspiration biopsy, while the remainder were followed up over a two-year period. Manual segmentations indicating areas of breast masses were prepared by an experienced radiologist. Exemplary US images are presented in Fig. 1. To develop the classification model,

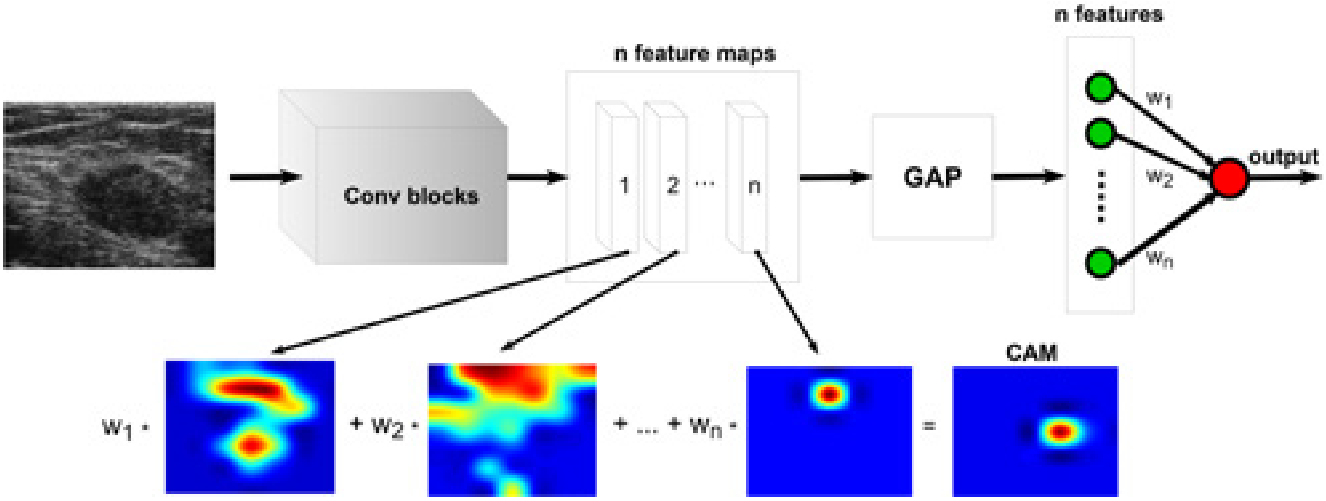

Fig. 2. *Scheme presenting the calculations of a saliency map. Weights of the linear dense classification layer are utilized to combine feature maps extracted before the global average pooling (GAP) layer*

patient data were divided into training and test sets with a 204/68 split. The ratio of malignant to benign masses, equal to approximately 34%, was maintained for each set.

Deep learning methods

To differentiate between malignant and benign breast masses, we used a recently proposed deep learning method(10). ResNet CNN pre-trained on the ImageNet dataset served as the backbone for the classification model(11,12). The last dense layer of the pre-trained network was replaced with a single output dense layer equipped with the sigmoid activation function suitable for the binary classification. The weights of the dense layer were randomly initialized. Transfer learning technique based on the scaling of deep representations was used to adjust the pre-trained network to process breast mass US images. More details about the deep learning model and the transfer learning technique utilized can be found in our previous paper(10). Binary cross-entropy and the Adam optimizer were used to train the network. The learning rate and batch size were set to 0.001 and 12, respectively. All US images were resized to the dimensions of 224x224 to match the original resolution of the ImageNet images used for the pre-training. Additionally, image augmentation was applied to generate more data for the training. The training of the network was terminated if no improvement with respect to the loss function on the test set was observed after 8 epochs.

Following the training, we used the CAM technique to generate image saliency maps (see Fig. 2). The CAM technique utilizes the weights of the last dense classification layer to combine feature maps extracted before the global average pooling (GAP) layer and generate a low resolution saliency map(13). Next, the saliency map is resized to match the resolution of the original US image. The saliency map indicates what image regions activated the network. In the case of the binary classification, the average value of the saliency map (the overall activation) is directly related to the classification decision made by the model(14). In this study, we coded the positive and negative activation regions with red and blue colors, respectively. A saliency map dominated by a red color would probably indicate a malignant breast mass, while a saliency map dominated by a blue color would correspond to a benign mass.

Evaluations

Classification performance was assessed using the receiver operating characteristic curve (ROC) and the area under the ROC curve (AUC). Accuracy, sensitivity, and specificity were calculated based on the point on the ROC curve that was the closest to the left upper corner of the curve(15). The pointing game metric was used to evaluate the saliency maps in a quantitative way(6). The aim of the pointing game score is to assess what image region is highlighted by the saliency map. In our case, we used the manual breast mass segmentations to pre-determine three regions in the US images:

1. The breast mass region, the appearance of which should naturally be taken into account by the network.
2. The peritumoral region, specified as a 5 mm ring around the breast mass, which was calculated with the morphological operations based on the breast mass manual segmentation. The appearance of this region is related to the characteristics of the mass boundary; blurry and indistinct margins are commonly encountered in malignant masses, while benign masses exhibit well-defined borders.
3. The region below the breast mass. Malignant masses commonly have a larger attenuation coefficient than benign masses, which makes the region below the malignant mass less bright in comparison to adjacent tissues, eventually causing the shadowing artifact. Therefore, the appearance of this region may also be important for the deep classifier.

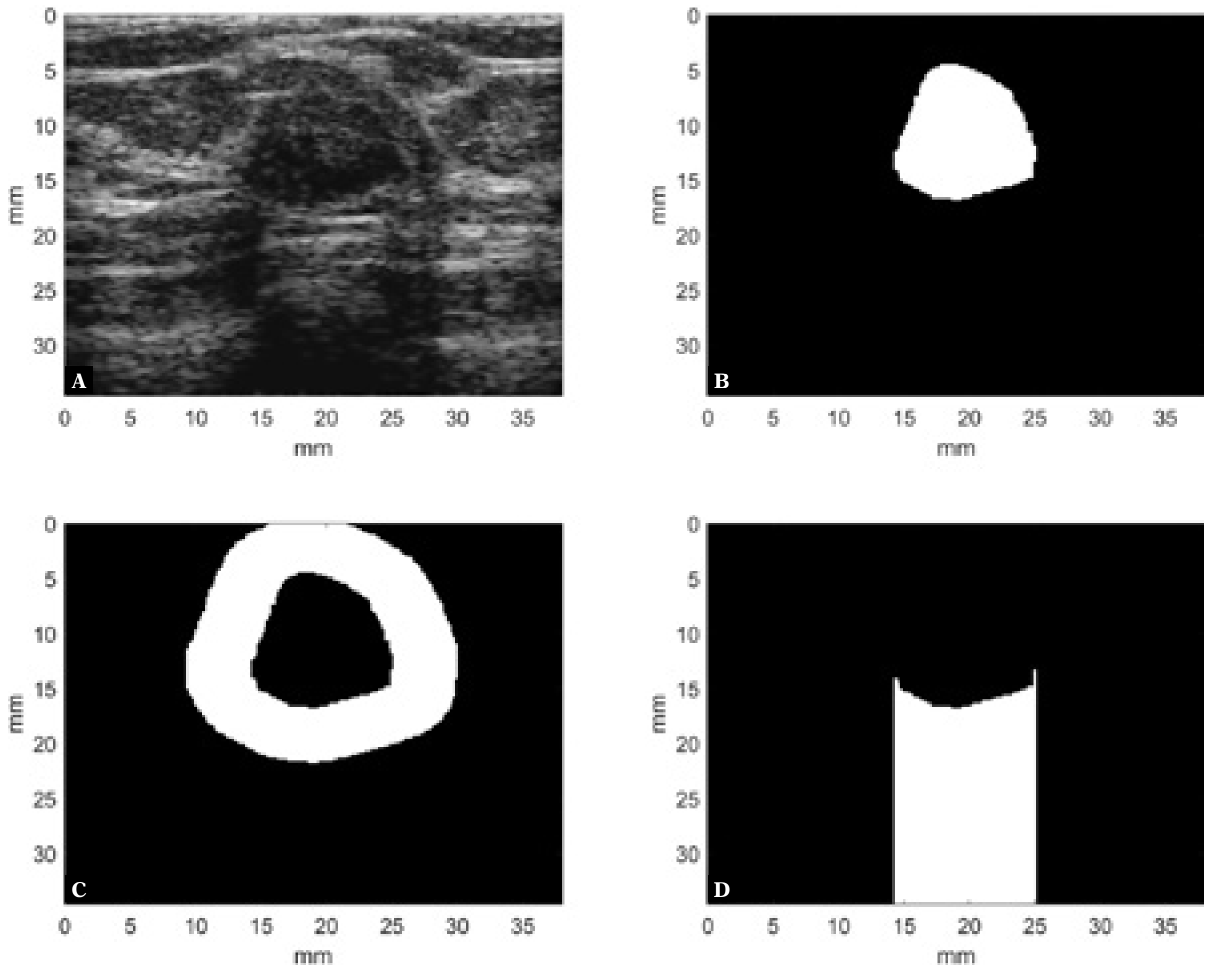

Fig. 3. A. *Exemplary US image and the three regions selected for the saliency map study:* **B.** *breast mass region,* **C.** *peritumoral region (mass boundary), and* **D.** *region below the breast mass*

Figure 3 presents a breast mass US image and the three regions selected to evaluate the saliency maps. In the case of the pointing game metric, we considered that the saliency map hit the region if the maximal or minimal (or both) point(s) in the saliency map was (were) contained within the particular region. As mentioned in the previous subsection, saliency maps in the case of the binary classification may show regions of both positive and negative activations. Therefore, to explain the decisions conducted by the model, it is important to assess both the location of the maximal and minimal activation areas.

Results

The classification performance of the implemented method is summarized in Tab. 1. The network achieved good performance, with AUC value and accuracy of 0.887, and 0.835, respectively. Tab. 2 presents the pointing game scores determined for the correctly classified cases from the test set. Here, we can see that in 71% of the cases the extreme points of the saliency maps corresponded to at least one of the three pre-selected regions. The points of extreme activation were related to the breast mass region, the peritumoral region, and the region below the mass in 34%, 38%, and 30% of cases, respectively. Notice that in some cases the maximum and minimum of the saliency

Tab. 1. *Breast mass classification performance of the deep learning model on the test set. AUC – area under the receiver-operating characteristic curve*

AUC	Accuracy	Sensitivity	Specificity
0.887 ± 0.015	0.835 ± 0.018	0.801 ± 0.025	0.868 ± 0.023

Tab. 2. *Pointing game scores obtained for the network's saliency maps and the three pre-defined regions. The results were calculated for the correctly classified cases from the test set*

Region	Percentage of accurate hits
Breast mass region	34%
Peritumoral region (boundary region)	38%
Region below the breast mass	30%
At least one of the above three regions	71%

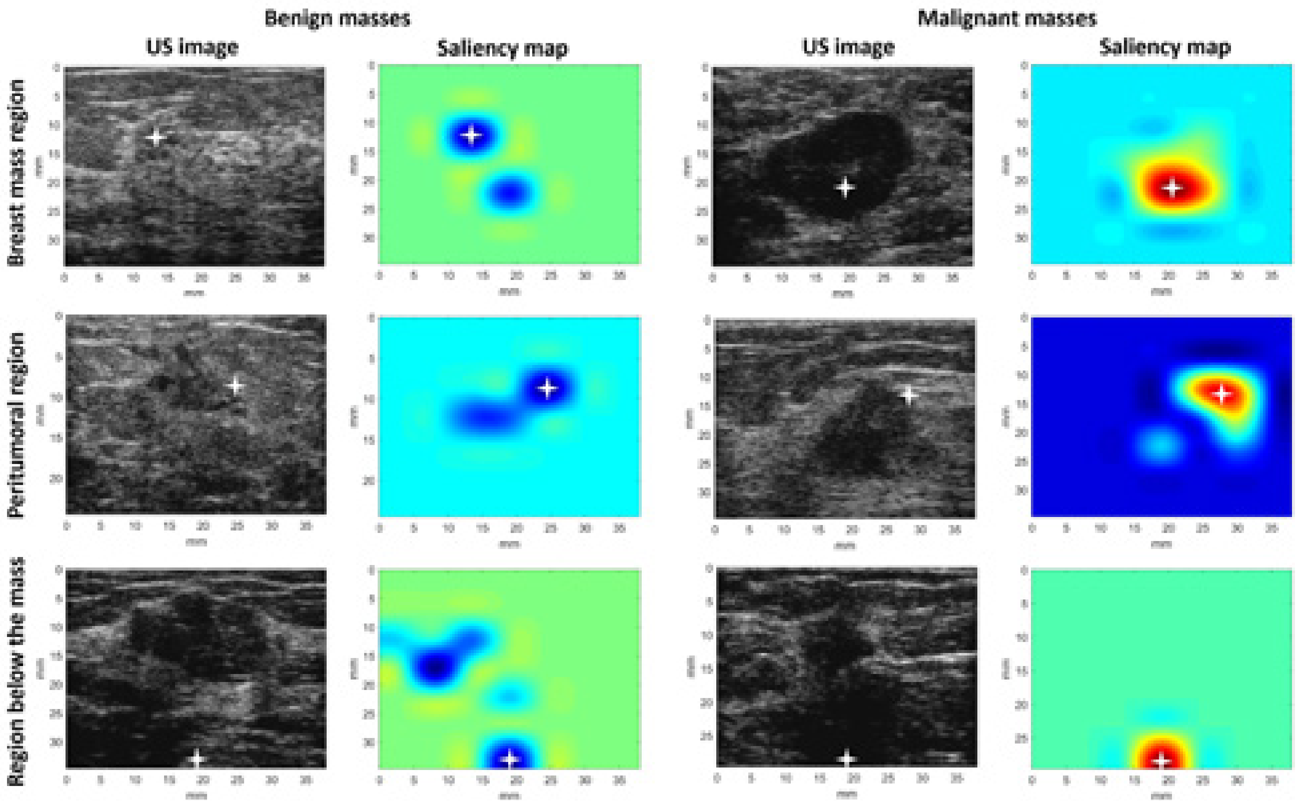

Fig. 4. *US images presenting benign and malignant breast masses and the corresponding saliency maps pointing out the three pre-determined regions in US images. The white cross indicates the extreme activation value of the saliency map responsible for the particular pointing game result*

map could correspond to two different regions, actually resulting in two pointing game hits. Exemplary class activation maps calculated for several masses are presented in Fig. 4.

Discussion

In this study, we developed a deep learning model for breast mass classification in US imaging. Moreover, we showed that the decisions conducted by the network could be visually explained. Regions of extreme activation in the saliency maps could be associated with the appearance of certain tissue regions in breast mass US images. We quantitatively assessed the saliency maps with the pointing game metric to find that the network's decisions could be related in 71% to the following three regions: the breast mass region, the peritumoral region, and the region below the breast mass. Our results revealed that the appearance of the breast mass boundary (peritumoral region) was the most frequently indicated by the network as important. This finding may be in an agreement with the review papers stating that the handcrafted mass boundary features are the better performing features for the breast mass classification(16). Nevertheless, the decisions of the network could also be associated with the appearance of the other two regions, depicting that the network could take into account the visual appearance of various US image regions.

There are several issues related to our study. First of all, we used only one technique to generate the saliency maps, but various other methods have been proposed recently, for example the GRAD-CAM or CAMERAS(17,18). While saliency maps generated with different algorithms should highlight similar tissue structures, it would be interesting to evaluate the usefulness and limitations of each technique. For example, the resolution of the saliency map may have an impact on the results obtained with the pointing game metric, which focuses only on the extreme points of the saliency map. As presented in Fig. 4, the activation area may be sometimes large enough to overlap with several tissue regions. For such cases, the conclusions drawn from the pointing game results may not be accurate. Second, saliency maps were calculated based on the ResNet CNN fine-tuned on a relatively small dataset of breast mass US images. It would be interesting to investigate whether the obtained saliency maps are universal in the case of the breast mass diagnosis or whether they perhaps depend on the utilized deep learning model. Moreover, we did not assess the relationship between the performance of the classifier and the pointing game results. Presumably, a network achieving low performance would probably provide noisy and meaningless saliency maps. Third, in the future it would be interesting to utilize saliency maps to improve classification performance. For example, it might be beneficial to suspend classification decisions if the saliency map does not highlight relevant tissue structures in the breast mass US image.

Conclusions

Our study is an important preliminary step toward a better understanding of deep learning models developed for breast mass diagnosis. We demonstrated that the decisions conducted by the deep network could be associated with the appearance of certain tissue regions in breast mass ultrasound images. In the future, we plan to perform additional experiments to further enhance the interpretability and explainability of deep learning models.

Conflict of interest

The authors do not report any financial or personal connections with other persons or organizations which might negatively affect the contents of this publication and/or claim authorship rights to this publication.

Acknowledgements

We acknowledge grant support from the National Science Center of Poland (2019/35/B/ST7/03792) and the National Centre for Research and Development (LIDER/24/0088/L-11/19/NCBR/2020).

Author contributions

Original concept of study: MB. Writing of manuscript: MB, KD-S, ZK, JL. Analysis and interpretation of data: MB, JL. Final acceptation of manuscript: MB, KD-S, JL. Collection, recording and/or compilation of data: MB, KD-S, HP-W, ZK, JL. Critical review of manuscript: MB, KD-S, HP-W, ZK, JL.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68: 394–424.
2. Ilesanmi AE, Chaumrattanakul U, Makhanov SS: Methods for the segmentation and classification of breast ultrasound images: a review. J Ultrasound 2021; 24: 367–382.
3. Kim J, Kim HJ, Kim C, Kim WH: Artificial intelligence in breast ultrasonography. Ultrasonography 2021; 40: 183–190.
4. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D: A survey of methods for explaining black box models. ACM Comput Surv 2018; 51: 93.
5. Adadi A, Berrada M: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018; 6: 52138–52160.
6. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S: Top-down neural attention by excitation backprop. Int J Comput Vis 2018; 126: 1084–1102.
7. Piotrzkowska-Wróblewska H, Dobruch-Sobczak K, Byra M, Nowicki A: Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. Med Phys. 2017; 44: 6105–6109.
8. ACR BI-RADS atlas: breast imaging reporting and data system. 5th ed. Reston; 2013. Available from: http://lib.ugent.be/catalog/rug01:002148922.
9. Jakubowski W, Dobruch-Sobczak K, Migda B: Standards of the Polish Ultrasound Society – update. Sonomammography examination. J Ultrason 2012; 50: 245–261.
10. Byra M: Breast mass classification with transfer learning based on scaling of deep representations. Biomed Signal Process Control 2021; 69: 102828.
11. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016: 770–778.
12. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L: ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition 2009: 248–255.
13. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A: Learning deep features for discriminative localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016: 2921–2929.
14. Byra M, Han A, Boehringer AS, Zhang YN, O'Brien Jr WD, Erdman Jr JW *et al.*: Liver fat assessment in multiview sonography using transfer learning with convolutional neural networks. J Ultrasound Med 2022; 41: 175–184.
15. Fawcett T: An introduction to ROC analysis. Pattern Recognit Lett 2006; 27: 861–874.
16. Gomez Flores W, Pereira WCDA, Infantosi AFC. Improving classification performance of breast lesions on ultrasonography. Pattern Recognit 2015; 48: 1121–1132.
17. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D: Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision 2017: 618–626.
18. Jalwana MAAK, Akhtar N, Bennamoun M, Mian A: CAMERAS: enhanced resolution and sanity preserving class activation mapping for image saliency. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021: 16327–16336.