Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

MSE Medical Engineering

Master Thesis

Detection and Classification of Lesions in Breast Ultrasound using a Deep Convolutional Neural Network

Carlotta Ruppert

Brühlbergstrasse 18

8400 Winterthur

Submission	28.02.2022
Advisor	Prof. Dr. Mathias Bonmarin
External Supervisor	Dr. Alexander Ciritsis

Abstract

Purpose: The aim of this study was to investigate the potential of a deep convolutional neural network (dCNN) to (i) detect (ii) segment and (iii) classify lesions in conventional breast ultrasound images in accordance with the Breast Imaging Reporting and Data System (BI-RADS), mimicking the human decision-making process for subsequent clinical procedures.

Materials and Methods: 3278 conventional breast ultrasound images from 1078 individual patients depicting lesions were manually segmented and classified by two radiologists according to the BI-RADS standard. A U-Net based multiclass segmentation network was trained with 2510 images and validated with a dataset of 768 images. The performance of the network on detection (precision, recall, f-score), segmentation (IoU) and classification (confusion matrices) was evaluated on a final test dataset consisting of 154 images which was not part of the training and validation dataset. The performance of the dCNN compared to human readers was quantified and evaluated using interrater agreement (Cohen's Kappa).

Results: For lesion detection, the network reached 70% (65%) precision and 70% (81%) recall with respect to the annotating radiologist 1 (2). In comparison, the precision and recall scores between two radiologists amounted to 69% and 86%, respectively. The networks segmentation accuracy of lesion shapes yielded an IoU of 75.1% (74.5%) with respect to radiologist 1 (2). The interrater agreement for lesion classification between the dCNN and radiologist 1 (2) was substantial (moderate) for 3 classes and almost perfect (substantial) for binary classification (benign and malignant). In comparison, the interrater agreement between the radiologists was measured as moderate to substantial for 3 and 2 classes, respectively.

Conclusions: In this study we demonstrated that a dCNN can be successfully trained to (i) detect, (ii) segment and (iii) classify lesions in accordance with the BI-RADS classification system in conventional breast ultrasound images. The performance of the dCNN is comparable to the performance of radiologists with more than two years' experience in breast imaging. Our dCNN can serve as an observer-independent guide for subsequent clinical procedures, contribute to standardization of BI-RADS classes and potentially prevent unnecessary biopsies or delayed treatment by reducing false-positive and false-negative diagnoses.

Table of Contents

1. Introduction

With more than 2.3 million annual cases worldwide, breast cancer is the most frequent cancer and leading cause of death from cancer among women[1,2]. Early diagnosis of breast cancer is key to increase the chance of the patient's survivability[3,4]. The most common and widely used modality for breast cancer screening remains mammography[5]. Although mammography has an overall sensitivity of approximately 85% [6,7], sensitivity decreases drastically in women with dense breast[8,9]. Since dense breast tissue is an additional risk factor for breast cancer[10], the importance of alternative screening methods becomes apparent.

Breast ultrasonography with its high sensitivity may be a viable alternative to mammography for women with dense breasts or who are afraid to undergo the discomfort caused by mammography[11,12]. However, the increased sensitivity is accompanied by increased false-positive rates[13]. Furthermore, the quality of breast ultrasound interpretation is highly observer-dependent and requires well-trained and experienced radiologists[14].

To combat this variation in interpretation, the American College of Radiology (ACR) released the Breast Imaging Reporting and Data System (BI-RADS) which helps guiding radiologists in their decision-making process and therefore acts as a standard. The atlas defines seven lesion classifications: incomplete (0), negative (1), benign (2), probably benign (3), suspicious for malignancy (4), highly suggestive for malignancy (5), and known biopsy-proven malignancy (6). According to each classification, different clinical procedures should be followed. BI-RADS class 1 to 2 do not require any further actions, whereas a three-to-six-month follow-up examination is recommended for probably benign lesions (BI-RADS 3). For a BI-RADS 4 or higher lesion a biopsy is recommended, which is then analyzed and possibly determines the lesion to be a BI-RADS 6 lesion[15]. In spite of the BI-RADS classification system, radiologic assessment is highly subjective, with high variability in inter- and intrareader agreement[16]. A computer-aided diagnostic (CADx) system for observer-independent classification of breast lesions according to the BI-RADS catalogue could help standardizing classifications further. Segmentation as the most detailed form of object detection is a desired feature of such a CADx system.

Promising preliminary results in the automated classification of breast lesions in ultrasound[17–20] combined with recent advancements made in segmentation architectures developed for biomedical image processing[21] raise expectations for fully automatic lesion segmentation and classification CADx systems.

The aim of this study is to investigate a dCNN that combines fully automatic segmentation and classification of breast lesions according to the ACR BI-RADS catalog in conventional breast ultrasound, to evaluate its performance compared to human readers and to research its potential to serve as an observer-independent CADx system that guides radiologists in their decision-making process.

2. Materials and Methods

2.1. Database Search

The training, validation and test dataset used for training and evaluation of the dCNN was created in 4 steps (Figure 1). The institutional Radiological Information System (RIS) was queried for patient reports receiving conventional breast ultrasound imaging containing the keyword "BI-RADS" for the time period of January 2013 to December 2015. This resulted in a total of 7480 radiological patient reports. Additionally, the Picture Archiving and Communication System (PACS) was queried for all conventional breast ultrasound examination acquired in the same time period. The PACS query resulted in a total of 135839 breast ultrasound images. Subsequently, the images were linked to the corresponding patient reports via patient ID and study date, resulting in 70944 images of 4542 patients, thereby filtering for images that were acquired during a lesion screening.

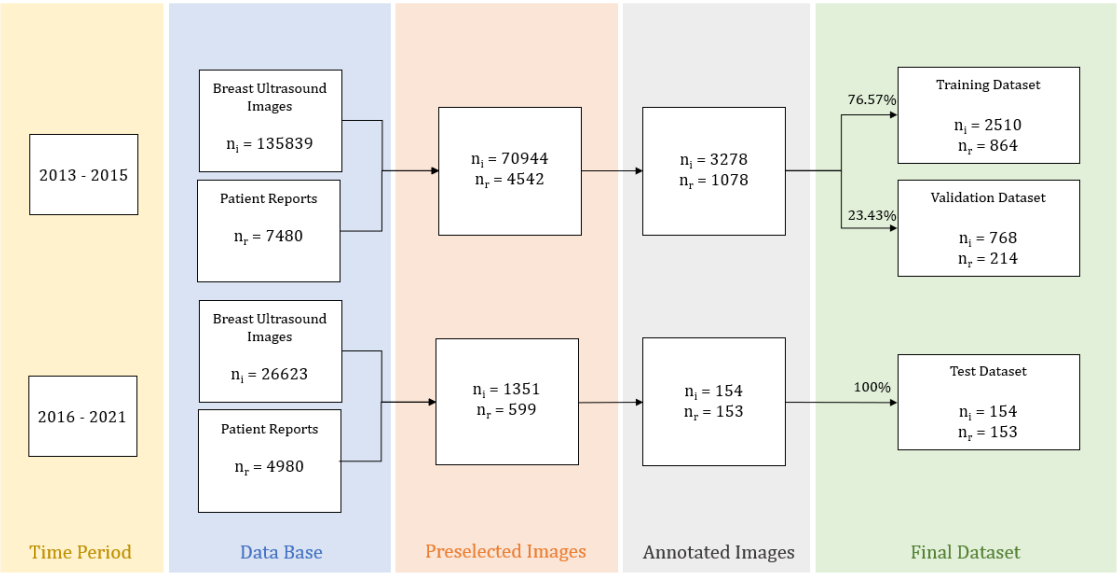


Figure 1: Creation of the Training, Validation and Test Dataset in 4 steps: 1. Extraction of a number of breast ultrasound images n_i and a number of patient reports containing the word "BI-RADS" n_r from the PACS data base. 2. Filtering images by assigning them to patient BI-RADS reports and in case of the test dataset pre-balancing the image classes according to the BI-RADS classes given in the report. 3. Labeling the images. 4. Splitting the data in a training, validation, and test dataset.

2.2. Data Labeling

In order to ease the workflow of the radiologists and ensure high-quality ground-truth labels we developed a custom graphical user interface (Figure 2) using the PyQt5 Python library, QMLCreator and Python programming. The labeling process included segmentation of lesions and their classification into BI-RADS classes 2 to 5. BI-RADS classes 1 and 6 are neglected since BI-RADS class 1 translates to healthy breast tissue in which case no lesion would be added and BI-RADS class 6 defines a malignant lesion that was pathologically confirmed; information that cannot be extracted from a mere image. The labeled BI-RADS classes 2 to 5 were further combined into three different classes defined as follows: “benign” (BI-RADS 2) (Figure 3a), “probably benign” (BI-RADS 3) (Figure 3b), and “(highly) suspicious for malignancy” (BI-RADS 4 /5) (Figure 3c). In total 3278 breast ultrasound images of 1078 patients were labeled. Images depicting healthy breast tissue were excluded during the labeling process.

Figure 2: Graphical User Interface used to select relevant images, segment lesions, annotate BI-RADS classes and keywords.

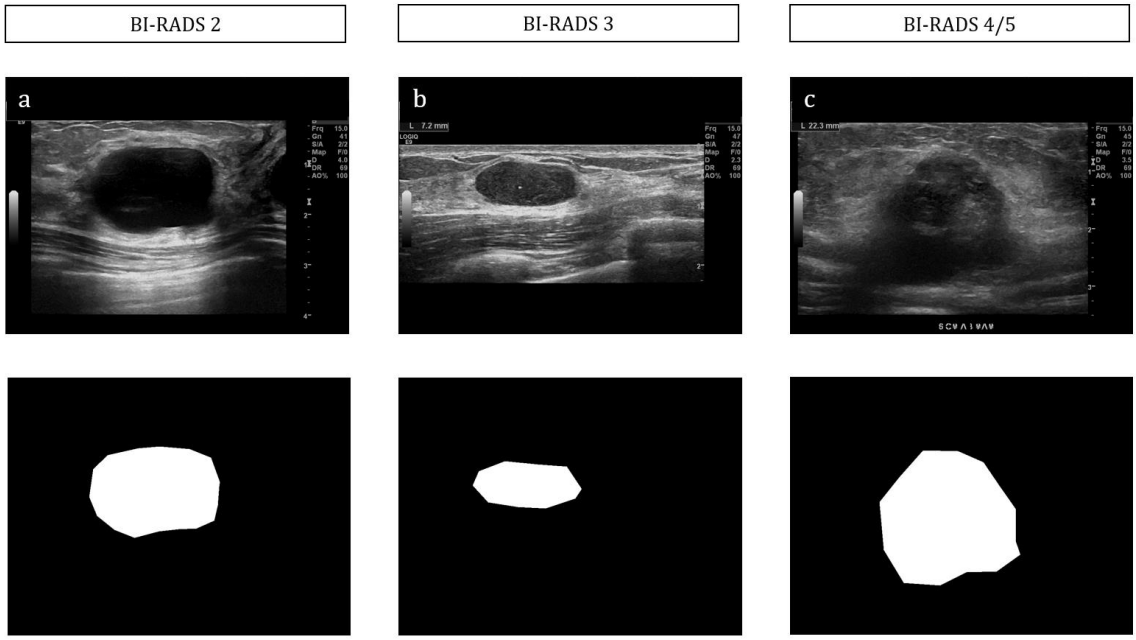


Figure 3: Ultrasound images of the three defined classes "benign" (BI-RADS 2)(a), "probably benign" (BI-RADS 3)(b) and "(highly) suspicious for malignancy" (BI-RADS 4/5)(c) with their respective segmented masks.

2.3. Training and Validation Dataset

For training and validation, the dataset was split into two parts; 76.57% of the images are used to train of the dCNN and 23.43% to evaluate the model. We ensured that images from a given patient screening do not appear both in training and validation datasets. The training dataset was unbalanced with respect to images depicting "benign" lesions (1387), "probably benign" lesions (934) and lesions that are "(highly) suspicious for malignancy" (189). Conversely, the validation dataset is balanced with 256 images per class as listed in Table 1.

Table 1: Number of images per class used for training and validation of the neural network.

Class	Training Dataset	Validation Dataset
benign (BI-RADS 2)	1387	256
probably benign (BI-RADS 3)	934	256
(highly) suspicious for malignancy (BI-RADS 4/5)	189	256

2.4. Test Dataset

We generated a test dataset using a similar procedure as outlined for the training and validation dataset (Figure 1). 26623 images and 4980 patient reports from between January 2016 and March 2021 were downloaded from the PACS and RIS database. Using the BI-RADS class given in the patient reports, the images were then pre-balanced according to the three introduced classes "benign", "probably benign" and "(highly) suspicious for malignancy". Two radiologists from different institutions with two and three years' experience in breast imaging, respectively, annotated a total of 154 images from 153 patients.

2.5. Training of the dCNN

The calculations were operated on a consumer-grade computer (Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, NVIDIA GeForce RTX 3070 GPU), running the operating system Ubuntu Linux 20.04 with Tensorflow 2.7.0 (Google Brain Team) and Keras 2.7.0 (Massachusetts Institute of Technology). A multiclass segmentation network based on the U-Net architecture[21] was implemented (Appendix A). The network receives a 256x256 grayscale image as input and predicts the segmentation masks of four classes: the three classes previously defined (BI-RADS 2, BI-RADS 3, BI-RADS 4/5) and additionally, an image-background class (Figure 4). The model was trained with a batch size of 4 for 35 epochs using a weighted categorical cross entropy loss to mitigate class imbalance. The model was optimized using the Adam optimizer with an initial learning rate of 0.001 which is reduced by a factor of 10 each time the validation loss stalled for longer than 5 epochs.

2.6. Evaluation metrics and post-processing

During training the model was evaluated using averaged accuracy for the prediction of each pixel as well as intersection over union (IoU), a metric, which is defined by the area of intersection of prediction and ground truth divided by the area of union of prediction and ground truth and is used to measure accuracy of object detection and segmentation. Due to heavy class imbalance (background pixels yield to 98.23% of the data) these metrics only provide limited information about the desired

performance of the model. Thus, additional metrics and post-processing steps were introduced to evaluate the performance on validation and test data.

Figure 4 visualizes the prediction output, post-processing steps needed for sensible evaluation of shape and class predictions. To obtain a robust output for segmented lesions the predicted background mask is inverted and filtered with a threshold of 100, leaving only segmented areas corresponding to detected lesions. In turn, each predicted lesion was compared to every ground truth lesion annotated by radiologists and assigned using the IoU. A predicted lesion is interpreted as true positive if IoU ≥ 0.5. Next, to evaluate segmentation accuracy of detected lesions IoU was calculated for all correctly detected lesions and for detected lesions of each defined class “benign” (BI-RADS 2), “probably benign” (BI-RADS 3) and “suspicious for malignancy” (BI-RADS 4/5). True positive, false positive and false negative lesion predictions were counted to compute standard metrics for evaluating model performance in the field of object detection:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Subsequently, for each true positive lesion, the predicted class of the model is determined as follows: The U-Net architecture predicts a class for each pixel; for each lesion, all predicted pixel classes are summed up and the maximum predicted class is set as the lesion class. We generated confusion matrices to evaluate classification accuracy of our dCNN and two human readers.

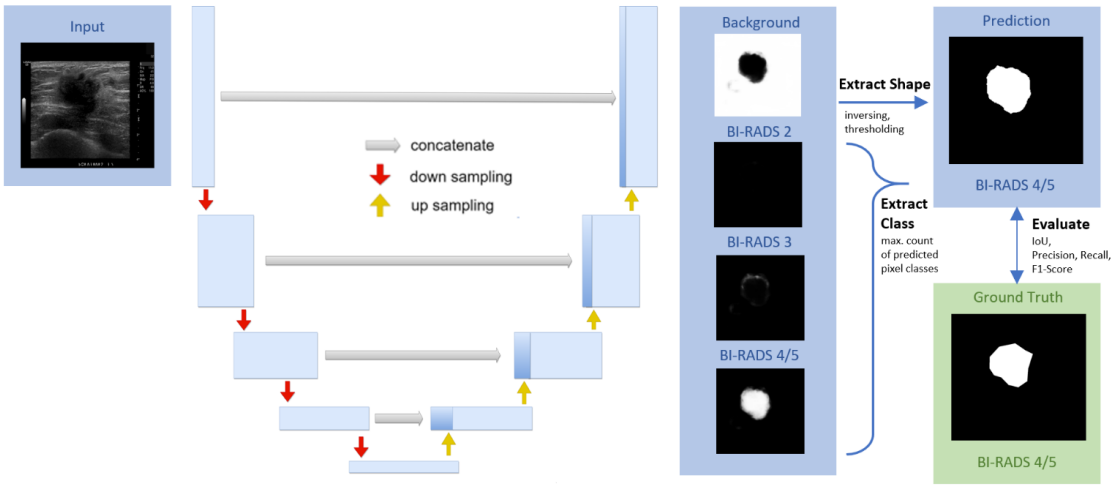


Figure 4: Multiclass segmentation prediction, postprocessing and evaluation pipeline. The U-Net based architecture consisting of an encoder and decoder structure receives a 256x256 breast ultrasound image as input and predicts segmentation masks for four classes: "background" (BI-RADS 1), "benign" (BI-RADS 2), "probably benign" (BI-RADS 3) and "(highly) suspicious for malignancy" (BI-RADS 4/5). The background prediction is inverted and filtered with a threshold to extract robust shapes of lesions. The class is extracted by counting the predicted pixel classes inside the segmented lesion. The extracted shape is compared to the ground truth using an IoU score. True positive, false negative and false positive predictions for lesions are used to calculate precision, recall and f-score.

2.7. Statistical analysis

Statistical analysis was performed using the Python scikit-learn library. Interrater and intrarater reliability between the dCNN and two radiologists were assessed by computing Cohen's kappa κ, a robust statistic to evaluate agreement of different readers. The Kappa results are interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement[22].

3. Results

3.1. Training and validation dataset

After 35 epochs of training, the network achieved a training and validation loss of 0.05 and 0.52, respectively. Moreover, the validation accuracy and IoU score amounted to 0.98 and 0.95, respectively.

3.1.1. Detection

A precision of 68.7%, recall of 84.3% and f-score of 75.7% were computed from true positive, false positive and false negative detected lesions on the validation dataset (Table 2).

3.1.2. Segmentation

An IoU score of 73.7% was measured for all correctly detected lesions in the validation dataset. Segmentation accuracy of the three introduced classes varied between 71.4% and 77.2% (Table 2).

3.1.3. Classification

We measured an average classification accuracy (normalized by the number of detected examples per class) of 78.76% on the validation dataset.

Table 2: Validation and test segmentation evaluation results. 0 = BI-RADS 2, 1 = BI-RADS 3, 2 = BI-RADS 4/5.

Detection and Segmentation Evaluation										
	Precision	Recall	F1	TP	FP	FN	IoU	IoU (0)	IoU (1)	IoU (2)
Validation										
dCNN / GT	68.7%	84.3%	75.7%	646	120	294	73.7%	**77.2%**	71.9%	71.4%
Testing										
dCNN / Reader 1	**70.3%**	70.7%	70.5%	113	27	64	75.1%	75.8%	72.3%	73.7%
dCNN / Reader 2	65.1%	81.4%	72.4%	**123**	51	**54**	74.5%	76.4%	74.1%	69.1%
Reader 1/ Reader 2	69%	**85.7%**	**76.4%**	120	**20**	**54**	**76.7%**	77.1%	**77.3%**	**75.1%**

3.2. Test dataset

3.2.1. Detection

We evaluated detection accuracy on the test dataset by testing the predictions of the dCNN against the annotations of radiologist 1 (2) and by comparing the annotations of both radiologists. The precision of the dCNN is of 70.3% (65.1%) with respect to radiologist 1 (2) compared to the precision score of 69% between the two radiologists (Table 2). The recall of the dCNN with respect to radiologist 1 (2) is of 70.7% (81.4%) and is slightly lower than the recall of 85.7% between the two radiologists. We observe that recall is always higher than precision.

3.2.2. Segmentation

The dCNN yielded an IoU of 75.1% (74.5%) with respect to radiologist 1 (2), which is only slightly below the interrater IoU of 76.7% between the two radiologists.

3.2.3. Classification

The average classification accuracy (normalized by the number of detected examples per class) of the dCNN amounted to 79.8% (65.8%) when tested against radiologist 1 (2) and is comparable to the measured accuracy between the radiologists

of 77.7% (Figure 5). Compared to the radiologist 1, the dCNN classified 96% of all lesions labeled as BI-RADS 4/5 correctly. More variation is visible in the classification between BI-RADS class 2 and 3; 23% and 37% of lesions annotated as BI-RADS 2 are classified as BI-RADS 3 by the network, for radiologist 1 and 2, respectively.

We computed the interrater reliability scores for the previously defined 3 classes and additionally for the binary classification of benign and malignant lesions (BI-RADS 2/3 and BI-RADS 4/5) (Table 3). The classification interrater agreement between the dCNN and radiologist 1 (2) was substantial (moderate) for three classes and almost perfect (substantial) for binary classification. In comparison, the radiologist's interrater agreement was measured as moderate and substantial for 3 and 2 classes, respectively. When the dCNN classifications were compared to the consensus of both radiologists, Cohen's Kappa read 0.719 (substantial) and 0.968 (almost perfect) for 3 and 2 classes, respectively.

Table 3: Interrater reliability scores using Cohen's Kappa

Three Classes				
Reader 1	**Reader 2**	**κ**	**linearly weighted κ**	**quadratically weighted κ**
dCNN	Radiologist 1	0.678	0.759	0.839
dCNN	Radiologist 2	0.458	0.594	0.717
Radiologist 1	Radiologist 2	0.599	0.710	0.794
dCNN	Consensus	0.719	0.822	0.897
Two Classes				
Reader 1	**Reader 2**	**κ**		
dCNN	Radiologist 1	0.894		
dCNN	Radiologist 2	0.751		
Radiologist 1	Radiologist 2	0.737		
dCNN	Consensus	0.968		

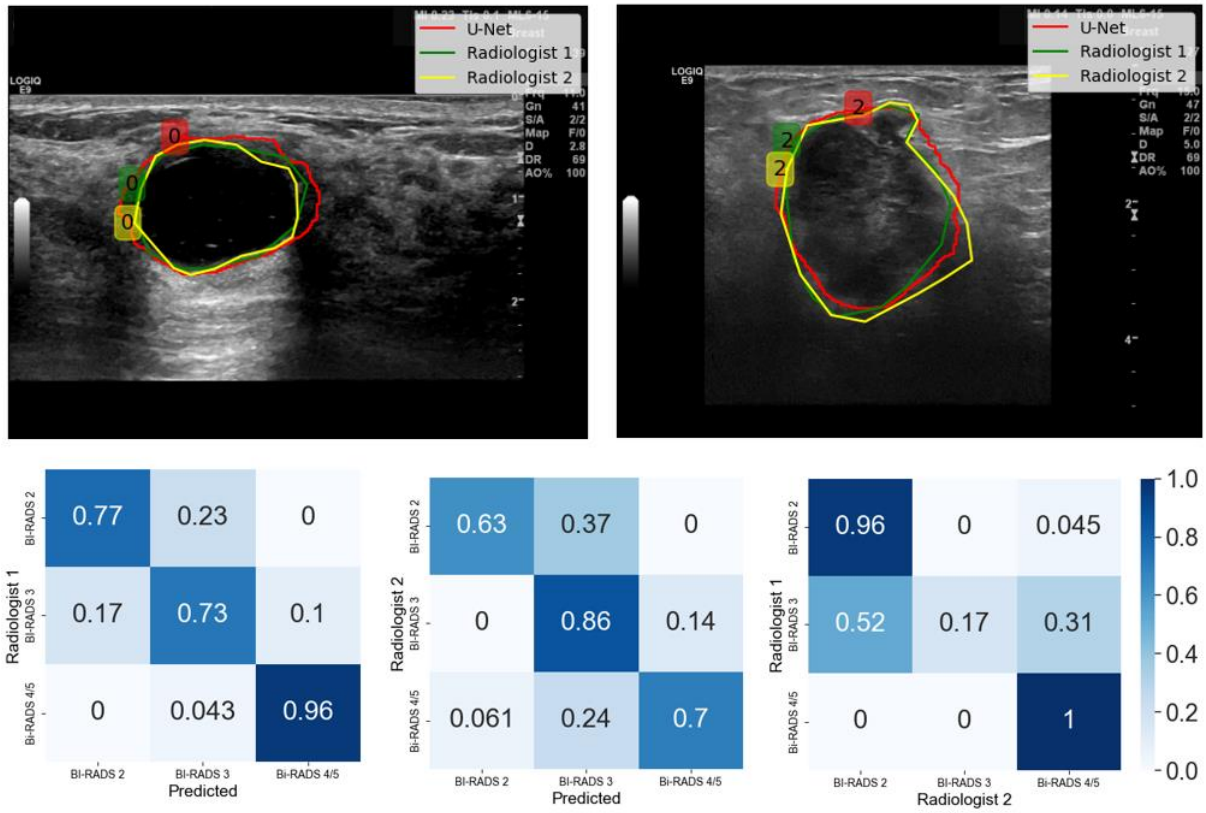


Figure 5: Comparison of class predictions of the neural network and two radiologists. Two visual examples show correctly segmented and classified lesions (class 0 corresponds to BI-RADS 2 and class 2 translates to BI-RADS 4/5). The three confusion matrices show the predicted classes versus the labels of radiologist 1 and 2 and finally the class labels of both radiologists (from left to right).

4. Discussion

In the present study, we showed that a U-Net based deep convolutional neural network (dCNN) can be successfully trained to (i) detect, (ii) segment and (iii) classify lesions in breast ultrasound imaging in accordance with the BI-RADS standard. We demonstrated that our model's predictions are comparable to those of two experienced radiologists in terms of detection, segmentation and classification accuracy of lesions in conventional breast ultrasound. For lesion detection our network reached 70% (65%) precision and 70% (81%) recall with respect to the annotations of radiologist 1 (2), which is comparable to the precision and recall scores of 69% and 86% between the two radiologists. This is particularly impressive since the metric used to identify true and false detections penalizes the network's predictions: the network's segmentation often consists of multiple closely adjacent areas, which may be misinterpreted as false positives and false negatives and contributes to lower precision and recall scores. Moreover, we showed that the segmentation accuracy of the network with an IoU of 75.1% (74.5%) with respect to the ground truth annotated by radiologist 1 (2) lies within 5% of the segmentation accuracy measured between the radiologists of 76.6%. Mean classification accuracy of our model yields to 79.8% (65.8%) with respect to radiologist 1 (2) and can therefore compete with the measured accuracy between both radiologists of 77.7%.

The BI-RADS classification is not only used by radiologists to describe their findings, but also to guide their choice of subsequent clinical action or the lack thereof. With our approach, lesion BI-RADS classes were split according to the consequent clinical measures taken: in case of a "benign" (BI-RADS 2) lesion no further action is needed, a follow-up examination in three to six months is planned for a "probably benign" (BI-RADS 3) lesion and a biopsy is recommended for a lesion which is "(highly) suspicious for malignancy" (BI-RADS 4/5). Therefore, by classifying lesions our approach is mimicking the decision-making process of a radiologist. Despite the ACR BI-RADS standardization, a reader's experience and workload contribute to their quality in radiological decision-making. A study showed that 29% of breast-density BI-RADS assessments were categorized differently after short-term reimaging[23]. We observed a similar discrepancy in this study in the categorization of BI-RADS lesions, in particular between BI-RADS 2 and 3 classification. Two radiologists with different levels of experience achieve moderate agreement (κ = 0.599) on our test data, mostly due to disagreement between BI-RADS 2 and 3 categorization. We presented 85 images, previously labeled as BI-RADS 2, twice to

the same reader which resulted in a moderate intrareader agreement (κ = 0.538). This emphasizes the difficulty of consistent classification of breast lesions in ultrasound and the need for further standardization, possibly achievable by using a CADx system, which supports the decision-making process. Our U-Net based architecture was able to achieve substantial (moderate) agreement with radiologist 1 (2). We observe a better agreement between network predictions and radiologist 1 (κ = 0.678) compared to the agreement between network predictions and radiologist 2 (κ = 0.458). We attribute this effect to the fact that radiologist 1 labeled most of the data the network was trained on. Note, that the low interrater agreement between network and radiologist 2 is also mostly caused by disagreement in BI-RADS 2/3 classifications. Misclassifications between benign (BI-RADS 2/3) and malignant (BI-RADS 4/5) lesions can have much more drastic consequences for affected false-positive diagnosed patients, who have to go through the discomfort of unnecessary biopsies and short-term distress[13,24] and affected false-negative diagnosed patients, who's survival-essential therapy might be delayed[13,24]. Therefore, we additionally evaluated interrater agreement when combining BI-RADS 2 and 3 in a single class, which leads to a binary classification of whether a biopsy is required or not. For binary lesion classification our model outperformed interrater agreement between both human readers (κ = 0.737) by reaching almost perfect agreement (κ = 0.894) with radiologist 1 and substantial (κ = 0.751) with radiologist 2. Our results show that the use of CADx software such as our U-Net architecture could be key to decrease false-positive and false-negative rates and their effects, thereby maximizing the potential of ultrasound breast-screening.

In recent years, many studies have investigated the classification of lesions in ultrasonography with the use of machine and deep learning. Deep convolutional neural networks as well as machine learning classifiers such as LDA, SVM and decision trees have been successfully used to classify breast lesions into the binary categories benign and malignant[17–19]. Moreover, automatic binary segmentation of lesions in ultrasounds using U-Net based architectures has proven to be successful[25,26]. Vakanski et. al. introduced a promising U-Net based architecture enriched by attention blocks that successfully segmented lesions in breast ultrasound images (0.955 AUC-ROC)[26]. Finally, an approach combining detection and classification of lesions according to the ACR BI-RADS catalog was carried out by Ciritsis et. al. using a sliding window approach[20]. Our approach is similar to that of Ciritsis et. al.[20] regarding the subdivision of the BI-RADS classes into recommendations for consequent

clinical procedures. However, instead of a computationally expensive sliding window approach for lesion detection our architecture segments lesions, which is less expensive and provides more detailed information about lesion shape. In the future, our model may be used to measure lesions using the accurate segmentation masks it provides. We introduced an end-to-end approach for detection, segmentation, classification and recommendation for clinical actions, thereby combining the aims of preceding works. Furthermore, our approach differs from most lesion classification works in terms of acquiring ground truth, which we did not obtain pathologically, but by radiologists annotating data following their standard clinical procedure without access to histological results of biopsies. This ensures firstly that the ultrasound image in question depicts the lesion and respective features of the given BI-RADS class, and secondly contributes to the development of a radiologist-mimicking dCNN.

In the following some limitations are discussed. Radiologists had limited access to patient reports and history while annotating training and validation data, but test data was annotated completely blinded for the fair and objective comparison between radiologists and the dCNN. This does not match the real-life clinical workflow, where radiologists do not only base their opinion on image data of a single modality but include patient and family history as well as previous screenings and examinations. Therefore, we can assume that the quality of lesion classification of the two human readers in real-life clinical settings would increase and result in higher interrater and intrarater agreements. Moreover, training, validation and test data while acquired over several years was extracted from the PACS database of a single institution leading to no variety in ultrasound vendors. Pre-balancing the dataset according to the BI-RADS classification given in the corresponding patient reports led to a small test dataset of only 154 images which limits statistical robustness. Finally, 80% of the training data was annotated by a single radiologist who also participated in the annotation of test data (radiologist 1). In order to achieve an observer-independent standardization of BI-RADS classification, the number of annotators, ultrasound vendors and clinical institutions should be maximized. Nonetheless, it should be underlined that our model was able to moderately generalize which is shown by the higher interrater agreement for two classes between dCNN and radiologist 2 compared with the interrater agreement of both radiologists. Moreover, this effect shows the potential of calibrating the architecture to regional differences in interpretation of the BI-RADS classes. Furthermore, the classification of lesions could be improved in the future by using descriptive tags radiologists use to

justify the given BI-RADS class and describe lesions in radiological reports. The tags describe shape, orientation, echo patterns, dorsal acoustic features and the nature of the internal structure and border of the lesion and are given in the ACR BI-RADS catalogue. We propose an explainable AI approach using a dCNN image classification algorithm to predict tags, which then can be used as in the real-life clinical workflow to determine the BI-RADS classification.

In conclusion, we demonstrated that a dCNN can be successfully trained to (i) detect, (ii) segment and (iii) classify lesions in conventional breast ultrasound comparably to experienced radiologists and in accordance with the BI-RADS classification system and can therefore serve as an observer-independent guide for subsequent clinical procedures. We showed that the use of our dCNN can contribute to further standardization in the interpretation of BI-RADS classification and has the potential to reduce false-positive and false-negative rates, thereby avoiding unnecessary biopsies and prevent vital treatment from being delayed. Moreover, our model could act as a learning-tool for prospective radiologists or be integrated in ultrasound machines commercialized for gynecologists who often perform ultrasound breast screenings but lack the experience of radiologists working in hospitals.

References

1. Breast cancer. Accessed February 24, 2022. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

2. 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(6):394-424. doi:10.3322/caac.21492

3. Tabár L, Vitak B, Chen THH, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology*. 2011;260(3):658-663. doi:10.1148/radiol.11110469

4. Recommendations on screening for breast cancer in average-risk women aged 40–74 years. *CMAJ Can Med Assoc J*. 2011;183(17):1991-2001. doi:10.1503/cmaj.110334

5. Lång K, Andersson I, Zackrisson S. Breast cancer detection in digital breast tomosynthesis and digital mammography-a side-by-side review of discrepant cases. *Br J Radiol*. 2014;87(1040):20140080. doi:10.1259/bjr.20140080

6. Bock K, Borisch B, Cawson J, et al. Effect of population-based screening on breast cancer mortality. *Lancet Lond Engl*. 2011;378(9805):1775-1776. doi:10.1016/S0140-6736(11)61766-2

7. Tabár L, Fagerberg CJ, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet Lond Engl*. 1985;1(8433):829-832. doi:10.1016/s0140-6736(85)92204-4

8. Mandelson MT, Oestreicher N, Porter PL, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. *J Natl Cancer Inst*. 2000;92(13):1081-1087. doi:10.1093/jnci/92.13.1081

9. Boyd NF, Martin LJ, Yaffe MJ, Minkin S. Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Res BCR*. 2011;13(6):223. doi:10.1186/bcr2942

10. Kerlikowske K, Miglioretti DL, Vachon CM. Discussions of Dense Breasts, Breast Cancer Risk, and Screening Choices in 2019. *JAMA*. 2019;322(1):69-70. doi:10.1001/jama.2019.6247

11. Berg WA, Zhang Z, Lehrer D, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA*. 2012;307(13):1394-1404. doi:10.1001/jama.2012.388

12. Poulos A, Llewellyn G. Mammography discomfort: a holistic perspective derived from women's experiences. *Radiography*. 2005;11(1):17-25. doi:10.1016/j.radi.2004.07.002

13. Lee JM, Arao RF, Sprague BL, et al. Performance of Screening Ultrasonography as an Adjunct to Screening Mammography in Women Across the Spectrum of Breast Cancer Risk. *JAMA Intern Med*. 2019;179(5):658-667. doi:10.1001/jamainternmed.2018.8372

14. Yap MH, Pons G, Marti J, et al. Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. *IEEE J Biomed Health Inform*. 2018;22(4):1218-1226. doi:10.1109/JBHI.2017.2731873

15. American College of Radiology, ed. *ACR BI-RADS®-Atlas Der Mammadiagnostik*. Springer Berlin Heidelberg

16. Lee HJ, Kim EK, Kim MJ, et al. Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. *Eur J Radiol*. 2008;65(2):293-298. doi:10.1016/j.ejrad.2007.04.008

17. Tanaka H, Chiu SW, Watanabe T, Kaoku S, Yamaguchi T. Computer-aided diagnosis system for breast ultrasound images using deep learning. *Ultrasound Med Biol*. 2019;45:S4. doi:10.1016/j.ultrasmedbio.2019.07.426

18. Fleury E, Marcomini K. Performance of machine learning software to classify breast lesions using BI-RADS radiomic features on ultrasound images. *Eur Radiol Exp*. 2019;3(1):34. doi:10.1186/s41747-019-0112-7

19. Yap MH, Goyal M, Osman FM, et al. Breast ultrasound lesions recognition: end-to-end deep learning approaches. *J Med Imaging*. 2018;6(1):011007. doi:10.1117/1.JMI.6.1.011007

20. Ciritsis A, Rossi C, Eberhard M, Marcon M, Becker AS, Boss A. Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. *Eur Radiol*. 2019;29(10):5458-5468. doi:10.1007/s00330-019-06118-7

21. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Lecture Notes in Computer Science. Springer International Publishing; 2015:234-241. doi:10.1007/978-3-319-24574-4_28

22. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica*. 2012;22(3):276-282.

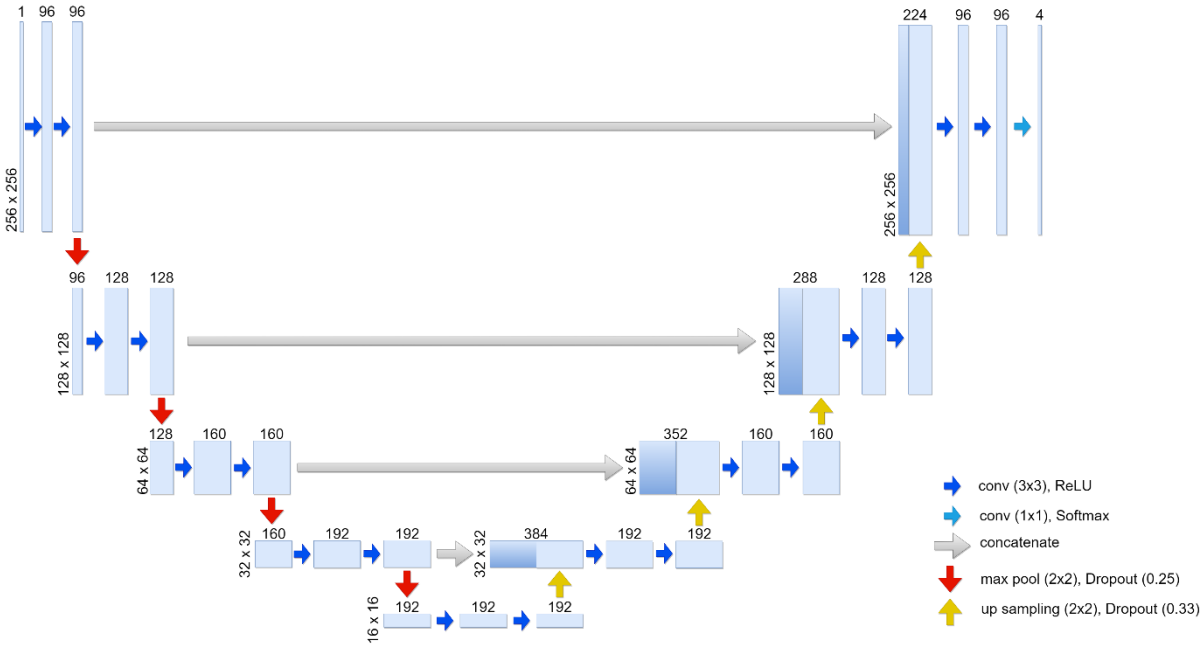
23. Kim WH, Moon WK, Kim SM, et al. Variability of breast density assessment in short-term reimaging with digital mammography. *Eur J Radiol*. 2013;82(10):1724-1730. doi:10.1016/j.ejrad.2013.05.004

24. Aro AR, Pilvikki Absetz S, van Elderen TM, van der Ploeg E, van der Kamp LJT. False-positive findings in mammography screening induces short-term distress — breast cancer-specific concern prevails longer. *Eur J Cancer*. 2000;36(9):1089-1097. doi:10.1016/S0959-8049(00)00065-4

25. Almajalid R, Shan J, Du Y, Zhang M. Development of a Deep-Learning-Based Method for Breast Ultrasound Image Segmentation. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. ; 2018:1103-1108. doi:10.1109/ICMLA.2018.00179

26. Vakanski A, Xian M, Freer PE. Attention-Enriched Deep Learning Model for Breast Tumor Segmentation in Ultrasound Images. *Ultrasound Med Biol*. 2020;46(10):2819-2833. doi:10.1016/j.ultrasmedbio.2020.06.015

Appendix



Appendix A: Multiclass U-Net based Architecture. The encoder structure receives an input gray-scale image of size 256x256 as input and consists of 4 encoder blocks with the filter sizes 96, 128, 160 and 192. Each block is built from two 2D convolutional layers with a kernel size of 3x3 and ReLU activation and is followed by batch normalization, max pooling for encoding and dropout with a dropout rate of 0.25. The bottom of the U-shaped architecture consists of the same layers and properties as before but is followed by a 2D up sampling layer and a slightly higher dropout rate of 0.33. It holds the number of activation maps at 192. The decoder structure is built symmetrically to the encoder structure but receives two inputs: the output of its counter-encoder-part and the output of the preceding decoder block. This way, information of different resolutions can be maintained. The last decoder block is followed by a final 2D convolutional layer with a kernel size of 1x1 and a filter size of 4 with Softmax activation resulting in 4 output maps.

Declaration of Independence

By submitting this work, the student assures that he/she wrote the work independently and without outside help (in the case of team work, the work of the other team members does not count as outside help).

The undersigned student declares that all cited sources (including websites) are correctly identified in the text or appendix, i.e. that the present work does not contain any plagiarism, i.e. no parts that are partially or completely taken from someone else's text or someone else's work under specification of their own authorship or without citing the source.

Place, date 02/28/2022, Winterthur

Signature student  .....

Selbständigkeitserklärung

Mit der Abgabe dieser Arbeit versichert der/die Studierende, dass er/sie die Arbeit selbständig und ohne fremde Hilfe verfasst hat (Bei Teamarbeiten gelten die Leistungen der übrigen Teammitglieder nicht als fremde Hilfe).

Der/die unterzeichnende Studierende erklärt, dass alle zitierten Quellen (auch Internetseiten) im Text oder Anhang korrekt nachgewiesen sind, d.h. dass die vorliegende Arbeit keine Plagiate enthält, also keine Teile, die teilweise oder vollständig aus einem fremden Text oder einer fremden Arbeit unter Vorgabe der eigenen Urheberschaft bzw. ohne Quellenangabe übernommen worden sind.

Ort, Datum 28.02.2022, Winterthur

Unterschrift Studierende/r