BJR
© 2018 The Authors. Published by the British Institute of Radiology
https://doi.org/10.1259/bjr.20170576

Received: 08 August 2017 | Revised: 28 November 2017 | Accepted: 04 December 2017

Cite this article as:
Becker AS, Mueller M, Stoffel E, Marcon M, Ghafoor S, Boss A. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *Br J Radiol* 2018; **91**: 20170576.

## FULL PAPER

# Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study

**ANTON S BECKER, MD, MICHAEL MUELLER, M. Med, ELINA STOFFEL, B. Med, MAGDA MARCON, MD, SOLEEN GHAFOOR, MD and ANDREAS BOSS, MD, PhD**

Institute of Diagnostic and Interventional Radiology, University Hospital of Zurich, Zurich, Switzerland

Address correspondence to: Dr Anton S Becker
E-mail: *anton.becker@usz.ch*

The authors Anton S Becker and Michael Mueller contributed equally to the work

**Objective:** To train a generic deep learning software (DLS) to classify breast cancer on ultrasound images and to compare its performance to human readers with variable breast imaging experience.
**Methods:** In this retrospective study, all breast ultrasound examinations from January 1, 2014 to December 31, 2014 at our institution were reviewed. Patients with post-surgical scars, initially indeterminate, or malignant lesions with histological diagnoses or 2-year follow-up were included. The DLS was trained with 70% of the images, and the remaining 30% were used to validate the performance. Three readers with variable expertise also evaluated the validation set (radiologist, resident, medical student). Diagnostic accuracy was assessed with a receiver operating characteristic analysis.
**Results:** 82 patients with malignant and 550 with benign lesions were included. Time needed for training was 7 min (DLS). Evaluation time for the test data set were 3.7 s (DLS) and 28, 22 and 25 min for human readers (decreasing experience). Receiver operating characteristic analysis revealed non-significant differences ($p$-values 0.45–0.47) in the area under the curve of 0.84 (DLS), 0.88 (experienced and intermediate readers) and 0.79 (inexperienced reader).
**Conclusion:** DLS may aid diagnosing cancer on breast ultrasound images with an accuracy comparable to radiologists, and learns better and faster than a human reader with no prior experience. Further clinical trials with dedicated algorithms are warranted.
**Advances in knowledge:** DLS can be trained classify cancer on breast ultrasound images high accuracy even with comparably few training cases. The fast evaluation speed makes real-time image analysis feasible.

## INTRODUCTION

Ultrasound has been known to have the potential to diagnose breast lesions for more than 40 years.[1] In recent years, it has been demonstrated that the sensitivity for detecting breast cancer can be improved by using ultrasound in addition to mammography particularly in patients with dense breast tissue,[2,3] mainly in younger females.[2] Due to the development of new technologies like shear wave elastography or contrast enhanced ultrasound,[4,5] breast ultrasound is steadily gaining importance in the workup of females with suspected breast cancer. Although, ultrasound requires more of the radiologist's time[6] and is operator-dependent,[7] in contrast to mammography it does not entail the usage of ionizing radiation, provides a better soft-tissue contrast and offers the capability to guide a biopsy instrument in real time. Nevertheless, the addition of ultrasound in screening examination may still produce a high number of false positives, with some studies reporting a positive predictive value <5%.[8]

Computer-assisted detection (CAD) software has shown promising results in mammography[9] and is used in the clinical routine to improve the radiologist's sensitivity. In breast ultrasound, despite promising results in recent studies,[10–18] there is currently no clinically approved tool for routine use. Artificial neural networks have shown very promising results in the past few years for a wide range of tasks.[13,15,16,18] One class of algorithms in particular, called deep learning, has recently started to revolutionize quality control in industrial manufacturing, exhibiting at least human-like performance in defect detection and classification tasks.[19] One of the main drawbacks of deep learning

in medical imaging is the need for large amounts of high-quality training data, *i.e.* images with pixel-wise annotation and histological ground truth or long-term follow up. This issue may even be more pronounced in ultrasound when compared to mammography, since the former has a much lower resolution (approx. 200 *vs* 40 μm) and only depicts a small part of the total breast tissue. This is a shared problem with quality control in industrial manufacturing lines, where the number of training data for a particular production line is also very limited. Hence, we hypothesized that ultrasound image analysis might profit from the proprietary algorithms used in a generic, industrial-grade deep learning software .

Therefore, the purpose of this study was to train a generic DLS designed for industrial quality control to diagnose breast cancer on a limited set of ultrasound images and to compare its performance to human readers with variable breast imaging experience.

## METHODS AND MATERIALS

### Study population

The cantonal ethics committee of Zürich, Switzerland approved this retrospective study and waived the need for an informed consent for patients from the year 2014.

All patients undergoing breast ultrasound in 2014 in our hospital were reviewed for malignant or benign lesions. The breast ultrasound examination at our institution is highly standardized: All examinations are performed on a Logiq E9 Ultrasound Station with a 9L linear probe (GE Healthcare, Chicago, IL). The depth extends beyond the lesion of interest and the focus point is set on the lesion. For large lesions, more than one focus point may be used. For this study, only the B-images were used (*i.e.* no colour-Doppler or elastography data). Exclusion criteria were applied by Breast Imaging Reporting and Data System (BI-RADS) scores:[20] As a first step, we excluded all patients with normal breast ultrasound (BI-RADS 1) as well as all patients with lesions classified as clearly benign, except for patients with prior breast-conserving surgical treatment (BI-RADS 2 excluding scars). As a second step, all patients with neither radiological follow up of at least 24 months (breast ultrasound, mammography or breast MR) nor histopathologically proven lesion were excluded. We chose the rather conservative timeframe of 24 months to ensure the absence of even low-grade malignancies in the depicted lesions at the time of examination.

### Deep learning analysis

For the image analysis we used an industrial grade image analysis software (ViDi Suite v. 2.0; ViDi Systems Inc, Villaz-Saint-Pierre, Switzerland). The software uses state-of-the-art deep learning algorithms[21] to identify and categorize anomalies in image data. It is currently used in various industries for quality inspection *e.g.* in defect detection of metal surfaces, real time traffic analysis or appearance-based product identification. Though it is currently not approved for the routine clinical use, it has recently shown promising results for detecting malignancies in a dual-centre mammography study.[22] Deep learning or deep neural networks differ from conventional "shallow" neural networks. Deep neural networks contain three or more hidden layers not directly connected to the output neurons, which enables them to solve much more complex problems.[23] All computations were performed on a GeForce GTX 1080 graphics processor unit. All malignant lesions were labelled and contoured by two investigators in consensus (ASB and AB) for supervised training. A randomly chosen subset of the images ($n$ = 445, 70%) was used for the training of the software, and the remaining cases ($n$ = 192) were used to validate the resulting model in the training process. The probabilistic heatmaps generated by the software were used by the investigators to qualitatively assess the suspicious features as detected by the neural network; they were not shown during the readout.

### Human readout

The validation images ($n$ = 192) were presented in random order to two radiologists (Reader 1, SG, PGY-3 resident in diagnostic radiology and Reader 2, MMa, 3 years of experience in breast imaging) who were blinded to the clinical information as well as the study background or design. Additionally, a 4th year medical student (Reader 3, ES) was given the training images. The student had no prior clinical or research experience in breast or ultrasound imaging and did not receive specific instructions. The images were placed in two separate folders, one for benign and one for malignant lesions, and available to study once for a freely chosen amount of time ($n$ = 445). It was expected that the student, like the software, learns solely from the data. Subsequently, the student rated the validation cases in the same manner as the radiologists. All images were rated on a 5-point Likert-type scale for malignancy (roughly corresponding to the BI-RADS classification with 5 meaning >98% probability of breast cancer). The time needed for the complete readout was noted, and for the medical student the training was timed as well.

### Statistical analysis

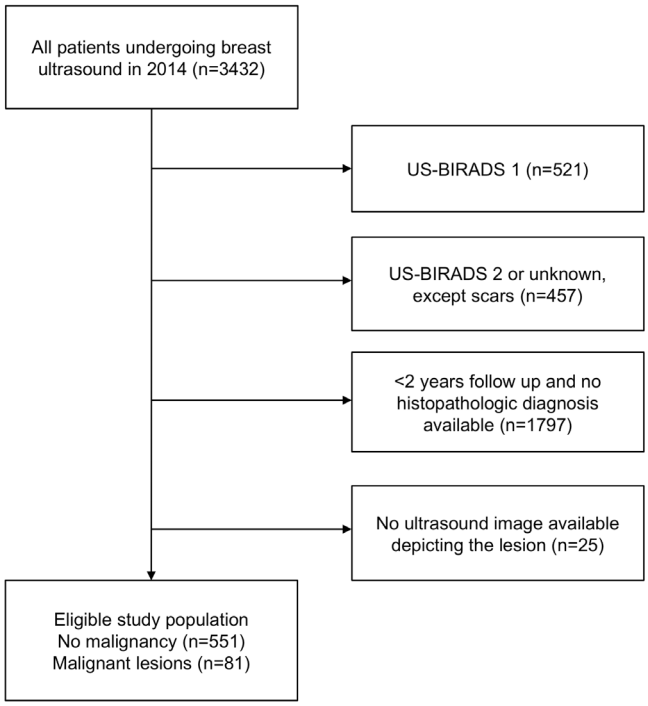
The statistical analysis was performed in R v.ion 3.3.1 (R Foundation for Statistical Computing, Vienna, Austria). Continuous variables were expressed as median and interquartile range, categorical variables as counts or percentages. Due to obvious differences in readout times between computer and humans, statistical testing was omitted. Interreader agreement was assessed pairwise with Lin's concordance correlation coefficient. To analyse the diagnostic performance, receiver operating characteristic analysis was performed for the computer test and validation data and the human readers. Diagnostic accuracy was expressed as the area under the receiver operating characteristic curve (AUC) and compared with DeLong's non-parametric test. The optimal cut-off (Youden Index) was determined and the resulting specificity, sensitivity, positive predictive value and negative predictive value were calculated. Sensitivities and specificites were compared using the McNemar test. A $p$ value < 0.05 was considered indicative of significant differences. All tests were two-tailed.

## RESULTS

### Study cohort

A total of ($n$) 3432 examinations were reviewed. Exclusion criteria were applied as defined and shown in Figure 1. From the remaining 657 patients, the ultrasound images saved in the examination were searched for the most representative image of the

Figure 1. Flowchart of the patient selection process.




described BI-RADS 3–6 lesion or scar (BI-RADS 2), respectively. During this step, another 25 patients had to be excluded due to lack of a suitable image. Two images of two different lesions were used in five patients: in two cases because of a bilateral malignancy and in three cases because of two independent benign lesions, which met the criteria described. The final population was comprised of 632 patients with one image for each of the 637 lesions. The eligible study cohort contains 82 patients with a malignancy or borderline lesion (84 lesions) and 550 patients without malignancies (553 benign lesions or scars).

The mean age of this cohort ($n$ = 632) was 53 ± 15 years (range 15–91 years). Of the 82 patients with malignancy, the most common histopathological diagnosis was invasive ductal carcinoma in 52 cases. Two of these patients had bilateral disease. Further histological subtypes are reported in Table 1. Of the 550 patients with benign lesions, three were included twice because of two different, biopsy-proven lesions, resulting in 553 images of benign lesions or scars included. 176 lesions in 173 patients were histopathologically proven, whereas for the other 377 no histopathological diagnosis was available, but only an inconspicuous (unremarkable) follow-up of at least 24 months. Table 2 shows the diagnoses of all benign lesions. In cases with no histopathological diagnosis available, the most probable diagnosis is listed, in general the first differential diagnosis of the initially examining radiologist. Of all included patients, 295 (46.7%) had undergone prior treatment of the breast, such as surgery, radiation therapy, or a combination therapy. The percentage of patients with prior treatment was smaller in the malignancy group ($n$ = 17, 21.0%) than in the one with benign lesions ($n$ = 278, 50.5%). The reason for this difference is the inclusion of post-surgical scars, which turned out to be benign in almost all the cases ($n$ = 209), while relapses of a prior breast carcinoma occurred in

Table 1. Histologies of the malignant and borderline lesions

| Histology | *n* (Training) | *n* (Validation) |
|---|---|---|
| Invasive ductal carcinoma | 39 | 13 |
| Invasive lobular carcinoma | 8 | 2 |
| Mucinous carcionoma | 1 | 0 |
| Tubular carcinoma | 1 | 0 |
| Ductal carcinoma *in situ* | 6 | 1 |
| Spindle cell sarcoma | 2 | 0 |
| Angiosarcoma | 1 | 0 |
| Lymphoma | 0 | 1 |
| Breast metastasis of a melanoma | 0 | 1 |
| Phyllodes tumour | 1 | 1 |
| Lobular intraepithelial neoplasia | 1 | 0 |
| Flat epithelial atypia | 1 | 0 |
| Papillary lesion | 1 | 1 |
| **Total** | **62** | **20** |

only 7 of the included cases (exclusively in patients with invasive ductal carcinomas). Benign lesions were slightly but significantly smaller than malignant lesions [12 mm (8–17 mm) *vs* 14 mm (10–22 mm); $p < 0.001$], which held true for all lesions with benign imaging characteristics and only follow-up as reference standard [8 mm (6–12 mm), $p < 0.001$] excluding scars [16 mm (12–19 mm), $p = 0.66$].

## Timing and interreader agreement

Training times for training set (445 images) were 7 min for the neural network [0.94 s image$^{-1}$ (sec im$^{-1}$)] and 48 min for the medical student (6.5 s im$^{-1}$). Readout times for the validation set (192 images) were 28 min (Reader 1, 8.8 s im$^{-1}$), 22 min (Reader 2, 6.9 s im$^{-1}$) and 25 min (Reader 3, 7.8 s im$^{-1}$). Processing time of the neural network for the test set was 3.7 s (0.0193 ± 0.0011 s im$^{-1}$). Interreader agreement between the human readers was best between the two more experienced readers 1 and 2 (0.56; [95% CI: 0.45–0.67]) and worst between the two less experienced readers 2 and 3 [0.35; (0.22–0.46)]. Interreader agreement between the neural network and the human readers was best between the software and the reader with intermediate experience [Reader 2, CCC = 0.49; (0.38–0.59)]. The full pair-wise comparison is given in Table 3.

## Diagnostic performance

The neural network's accuracy on the training set ($n$ = 445) was AUC = 0.96 [95% CI (0.92–0.99)]. The performance on the validation set ($n$ = 192) was AUC = 0.84 (0.75–0.93). Specificity and sensitivity were 80.4 and 84.2%, respectively. Diagnostic accuracy measured by AUC was not significantly different between the human readers (Reader 1: AUC = 0.89, Reader 2: AUC = 0.89 and Reader 3: AUC = 0.79) and the neural network ($p$ = 0.45–0.47), as depicted in the Receiver operating characteristic curve in Figure 2. As visualized in Table 4, between the human readers,

Table 2. Diagnoses of the benign lesions

| Diagnosis | Pathology | Follow up | *n* (training) | *n* (validation) |
|---|---|---|---|---|
| Fibroadenoma | 75 | 58 | 133 | 37 |
| Fibrosis | 68 | 1 | 69 | 8 |
| Cyst | 15 | 54 | 69 | 19 |
| Fat necrosis | 4 | 27 | 31 | 7 |
| Adenosis | 4 | 0 | 4 | 1 |
| Normal breast tissue | 2 | 0 | 2 | 1 |
| Duct ectasia | 2 | 2 | 4 | 2 |
| Scar | 1 | 208 | 209 | 69 |
| Abscess | 1 | 0 | 1 | 0 |
| Oil cyst | 1 | 0 | 1 | 0 |
| Hamartoma | 1 | 0 | 1 | 0 |
| Fat lobe | 1 | 5 | 6 | 1 |
| Usual ductal hyperplasia | 1 | 0 | 1 | 0 |
| Lymph node | 0 | 7 | 7 | 4 |
| Focal dense breast tissue | 0 | 3 | 3 | 1 |
| Haematoma | 0 | 2 | 2 | 0 |
| Seroma | 0 | 2 | 2 | 0 |
| Atheroma | 0 | 1 | 1 | 0 |
| Lipoma | 0 | 1 | 1 | 0 |
| Varicose vein | 0 | 1 | 1 | 0 |
| Benign (not further specified) | 0 | 5 | 5 | 5 |
| **Total number of lesions** | **176** | **377** | **398** | **155** |
| **Total number of patients** | **173** | **377** | **396** | **154** |

there was a significant trend of better performance with increased experience, especially for the specificity (89.0, 82.7 and 72.8%), but also for the AUC (0.89, 0.89 and 0.79) and for the sensitivity (84.2, 84.2 and 73.7%).

### Features on the heatmaps

The neural network rated post-operative changes more often as malignant than the human readers (Figure 3), especially in cases with large areas of acoustic shadowing. The differentiation of other features such as size or texture might have led to the few false negatives as illustrated in Figure 4, which shows a quite well defined, small carcinoma with partly well defined granular internal texture. On the other hand, the neural network was excellent and in several cases superior to the readers in correctly classifying small benign lesions (Figure 5) and voluminous malignant lesions (Figure 6). Interestingly, the neural network classified the only lymphoma and the only male patient (Figure 6) correctly in contrast to the two more experienced readers. Notable are also the cases where the neural network classified benign lesions of patients with no prior surgery correctly, even if the images might have had some aspects of post-operative changes (Figures 7 and 8).

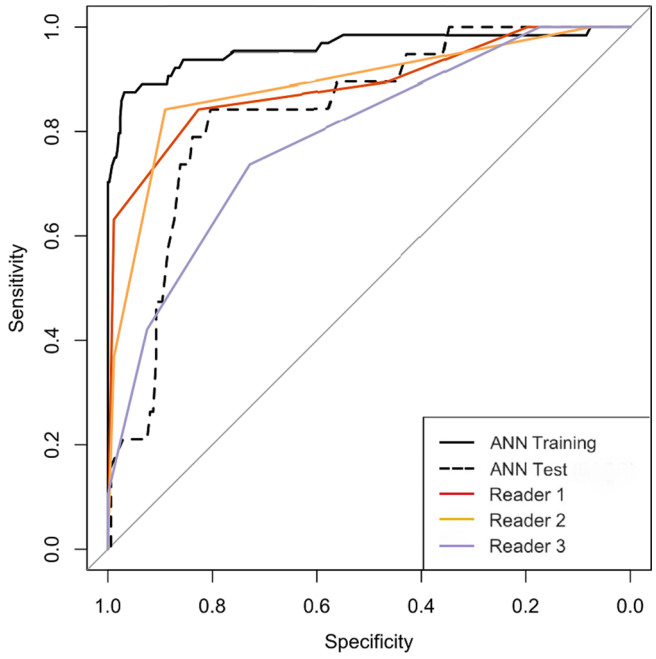
## DISCUSSION

In the present study, we directly compared the diagnostic performance of an industrial grade artificial neural network for image analysis with human readers. We found that the neural network, trained on only a few hundred cases, exhibited comparable accuracy to the reading of a radiologist. There was a tendency for the neural network to perform better than a

Table 3. Pair-wise interreader agreement measured by the concordance correlation coefficient

| | Neural network | Reader 1 | Reader 2 |
|---|---|---|---|
| Reader 1 | 0.33 [95% CI (0.19–0.47)] | | |
| Reader 2 | 0.49 [95% CI (0.37–0.59)] | 0.57 [95% CI (0.45–0.67)] | |
| Reader 3 | 0.20 [95% CI (0.08–0.32)] | 0.46 [95% CI (0.34–0.56)] | 0.35 [95% CI (0.22–0.47)] |

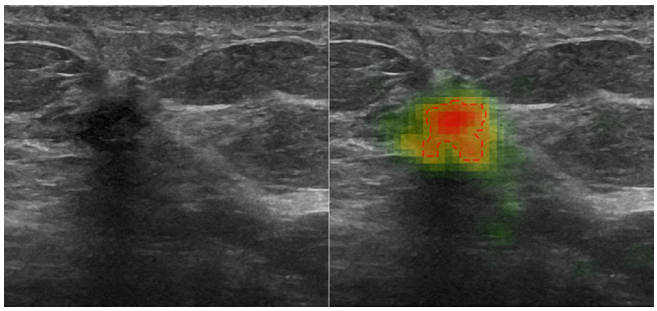
Figure 2. Receiver operating characteristic curve of the whole study population (black solid) as well as the test data set (black dashed) and the performance of the human readers on the test cohort (red and orange for the radiologists, purple for the medical student). AUC for the software were 0.96 for the training set, 0.84 for the validation set, and for the readers (validation only) 0.89, 0.89 and 0.79, respectively. AUC, area under the receiver operating characteristic curve.




medical student, who was trained with the same training data set.

The potential of neural networks to outperform humans has been described in other fields like handwritten digit recognition[24] or surface texture classification.[25] The ratings of the neural network in our study were most similar to the human reader with intermediate experience, indicating that the versatility of the software still comes at the cost of being outperformed by more seasoned radiologists. Also, it has to be kept in mind that in the clinical setting the radiologist does not only classify images, but examines the patient: This entails careful examination of prior examinations, family/personal history and corresponding other modalities (mammography or MRI), and lastly communicating the results to referring physicians and the patient in an

Figure 3. A 71-year-old female with scar after segmentectomy of the right breast. This scar was originally classified as BI-RADS 3, before it was down-staged after stable follow-up (total follow-up: 28 months). Both the neural network (0.95, cut-off 0.69) and the two radiologists (4/5 and 5/5) rated the lesion false positive as probably malignant. BI-RADS, Breast Imaging Reporting and Data System.



appropriate manner. However, at an evaluation time in the order of milliseconds per image, the software in its current state may serve as a visual aid for inexperienced physicians. The feasibility of a real-time ultrasound analysis with deep learning has already been demonstrated in fetal ultrasound.[26] We have illustrated a potential use in the online supplementary movie (Supplementary Video 1, Supplementary material available online) with two retrospectively analysed ultrasound movies of a benign adenosis (left) and an invasive ductal carcinoma (right).

Interestingly, the software showed remarkable generalizability in that it identified the only lymphoma in the only male patient as a malignant lesion.[27] The software was not confused by the two factors which were completely missing in the training set. On the other hand, this malignoma was missed by the two radiologists but not by the student—probably because the radiologists were actively looking for primary female breast cancer (a case of inattentional blindness)[28]. Nevertheless, it may be argued that this would certainly not have happened in the real clinical setting, *i.e.* in conjunction with the patient history and physical examination.

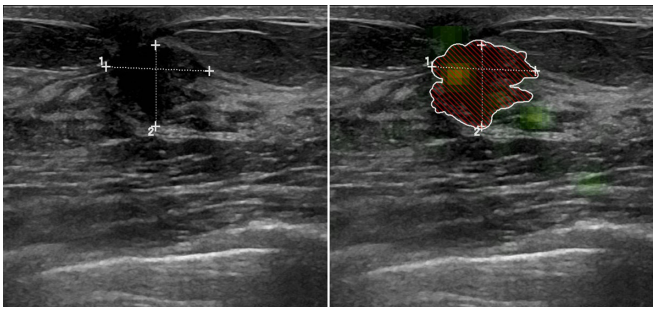
Compared to mammography, ultrasound images exhibit a lower spatial resolution per image, which however does not result in an inferior performance of the machine learning algorithm.[22] One reason for this is presumably the better soft tissue contrast of ultrasound, so the images contain more relevant information

Table 4. Results of the ROC analysis

| | AUC (95% CI) | Specificity (%) | Sensitivity (%) | PPV (%) | NPV (%) |
|---|---|---|---|---|---|
| Neural network (training) | 0.96 (0.92–0.99) | 96.9 | 87.5 | 82.4 | 97.9 |
| Neural network (validation) | 0.84 (0.75–0.93) | 80.3 | 84.2 | 32.0 | 97.9 |
| Reader 1 | 0.89 (0.79–0.98) | 89.0 | 84.2 | 45.7 | 98.1 |
| Reader 2 | 0.89 (0.79–0.98) | 82.7 | 84.2 | 34.8 | 97.9 |
| Reader 3 | 0.79 (0.69–0.89) | 72.8[a] | 73.7[a] | 23.0 | 96.2 |

AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value.
[a]$p < 0.05$ compared to both other readers and the NN.

Figure 4. One of the rare examples of the false negatives, where the human readers were superior to the neural network in detecting malignancy. A 41-year-old female with a palpable mass in her left breast initially rated as BI-RADS 5 lesion and later confirmed malignant (invasive ductal carcinoma). While the neural network rated the lesion as rather benign (0.46, cut-off 0.69), the two readers with clinical experience classified it as probably malignant (4/5) and the medical student as undetermined (3/5). BI-RADS, Breast Imaging Reporting and Data System.



about the (soft tissue) tumours despite containing a lower absolute number of pixels. Moreover, since mammographies are projection images, each pixel actually represents the integral of the three dimensional space in the projection axis, meaning that the relevant information about the pathology can be "contaminated" by overlaid physiological structures. Quite contrary, the image in ultrasound represents a single slice of the 3D space, selected by the examining physician with the focus on the lesion of interest (higher "lesion-to-background" ratio).

Usually, neural networks require large amounts of training data in the order of millions of images, especially when the resolution is low. Nonetheless, our study shows that a human-like performance can already be achieved with several hundred images. One implication may be that radiology software in the future could be tailored to the patient population which

Figure 5. A 60-year-old female with an initially BI-RADS 4 classified lesion of the right breast, which turned out to be a cyst after biopsy. The two radiologists rated the lesion the same as the radiologist performing the examination had done, as somewhat between indifferent and rather malignant (3 and 4/5), while the medical student rated the lesion as rather benign (2/5). The neural network classified the lesion correctly as benign and could have prevented the unnecessary biopsy (0.23, cut-off 0.69).

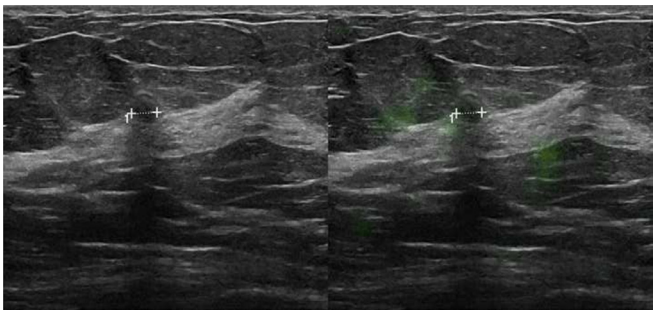
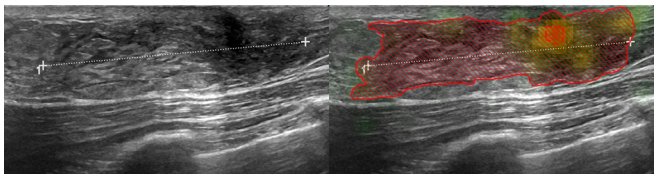


Figure 6. The only male patient (58 years old) and the only lymphoma included in the study population were rated as rather benign by the two radiologists (2/5), but as malignant by the neural network (0.78, cut-off 0.69). Interestingly, the medical student also correctly diagnosed the lesion as potentially malignant (4/5).



a particular hospital is serving, and would not need to be an "out-of-the-box-blend" from other cohorts with the resulting trade-off in diagnostic performance. The heatmap would function as a radiological textbook written for the particular institution/patient population, helping the radiologist in his decision making process.

Although the current work used a supervised training approach, *i.e.* the lesions were marked manually pixel-wise in the images, one could imagine that in the future, images and radiology as well as pathology reports could be extracted from a database (*e.g.* PACS) and the neural network could train itself with the available data in an unsupervised, or semi-supervised fashion.[29]

A number of limitations of this study need to be acknowledged. First, a large proportion of patients had to be excluded due to our stringent inclusion criteria. However, this was necessary to avoid training of the neural network on data with poor reference standard and thus obtain a falsely high or low performance. Also, this is a single-centre study with only a few hundred lesions, and a large part of the benign lesions were scars. This is another major limitation, since a classifier trained on such a large proportion of scars may misdiagnose cancerous lesions with similar characteristics if applied

Figure 7. A 60-year-old female with a lesion initially rated as BI-RADS 4, later confirmed as biopsy-proven fibrosis of the left breast and 28 months of unsuspicious follow-up. All the human readers rated the lesion as probably malignant (4 or 5/5). Only the neural network classified the lesion correctly as benign (0.38, cut-off 0.69). This is one of the examples where the neural network could have prevented an unnecessary biopsy.

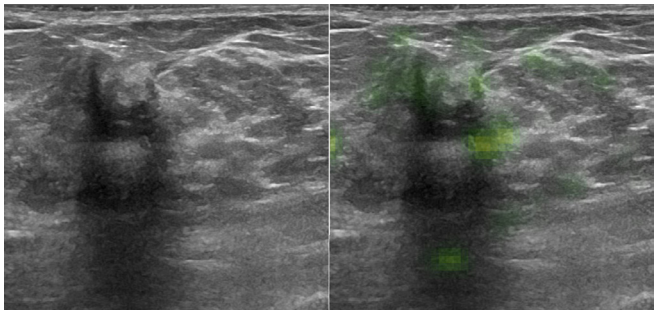
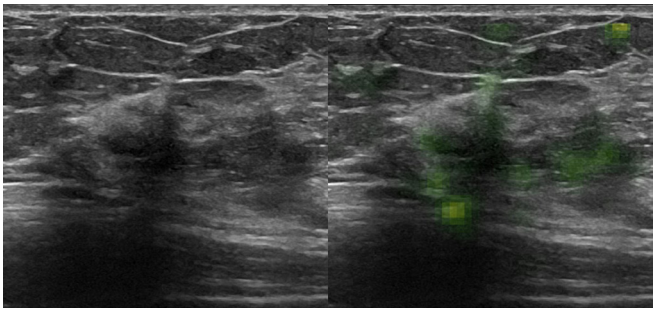


Figure 8. A 55-year old female with an initially BI-RADS 4 classified lesion of the left breast, for which biopsy showed adenosis and no sign of malignancy. It was correctly classified as benign by the neural network (0.52, cut-off 0.69), which might have rendered the biopsy unnecessary.



in another population, *e.g.* a healthy screening cohort . The results may thus not entirely be transferrable to other patient populations—which, as discussed above, can also be seen as a strength. As evident by the high number of patients with a prior procedure included, there was a high proportion of referred patients with a prior history of cancer or surgery. Combined with the retrospective study design, we acknowledge an inherent selection bias. Prospective, multicentre studies should be performed to validate our results. Second, our sample size is rather small regarding the retrospective design. Restricting the sample size allowed us to demonstrate the robustness of the software on a small training dataset, which in itself is an important result given the large amounts of data used in other studies.

In conclusion, our retrospective, single-centre study demonstrates that a generic deep-learning software for industrial image analysis can diagnose breast cancer in breast ultrasound images with a high accuracy, comparable to human readers, at a speed that would allow real-time analysis during an ultrasound examination. These results warrant further investigation with dedicated algorithms. The software learns better and faster than a human reader with no prior experience, given the same amount of training data.

## Advances in knowledge

(1) DLS for industrial quality control can detect anomalies in breast ultrasound with high diagnostic accuracy (AUC 0.84), comparable to radiologists (AUC 0.88).
(2) The software learns faster and better than a medical student with no prior experience in breast imaging (AUC 0.79) and its reading is most similar to a radiology resident (CCC = 0.49).
(3) The speed of the software in the order of milliseconds per image would allow real-time analysis during an ultrasound examination.

## REFERENCES


1. Cole-Beuglet C, Beique RA. Continuous ultrasound b-scanning of palpable breast masses. *Radiology* 1975; **117**: 123–8. doi: https://doi.org/10.1148/117.1.123
2. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology* 2002; **225**: 165–75. doi: https://doi.org/10.1148/radiol.2251011667
3. Brem RF, Lenihan MJ, Lieberman J, Torrente J. Screening breast ultrasound: past, present, and future. *AJR Am J Roentgenol* 2015; **204**: 234–40. doi: https://doi.org/10.2214/AJR.13.12072
4. Giannotti E, Vinnicombe S, Thomson K, McLean D, Purdie C, Jordan L, et al. Shear-wave elastography and greyscale assessment of palpable probably benign masses: is biopsy always required? *Br J Radiol* 2016; **89**: 20150865: 20150865: . doi: https://doi.org/10.1259/bjr.20150865
5. Xiao XY, Chen X, Guan XF, Wu H, Qin W, Luo BM. Superb microvascular imaging in diagnosis of breast lesions: a comparative study with contrast-enhanced ultrasonographic microvascular imaging. *Br J Radiol* 2016; **89**: 20160546. doi: https://doi.org/10.1259/bjr.20160546
6. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Böhm-Vélez M, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *JAMA* 2008; **299**: 2151–63. doi: https://doi.org/10.1001/jama.299.18.2151
7. Skaane P, Olsen JB, Sager EM, Abdelnoor M, Berger A, Kullmann G, et al. Variability in the interpretation of ultrasonography in patients with palpable noncalcified breast tumors. *Acta Radiol* 1999; **40**: 169–75. doi: https://doi.org/10.3109/02841859909177733
8. Tice JA, Ollendorf DA, Lee JM, and Pearson SD. The comparative clinical effectiveness and value of supplemental screening tests following negative Mammography in women with dense breast tissue. Institute for Clinical & Economic Review. 2014.
9. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12, 860 patients in a community breast center. *Radiology* 2001; **220**: 781–6. doi: https://doi.org/10.1148/radiol.2203001282
10. Moon WK, Lo CM, Chang JM, Huang CS, Chen JH, Chang RF. Quantitative ultrasound analysis for classification of BI-RADS category 3 breast masses. *J Digit Imaging* 2013; **26**: 1091–8. doi: https://doi.org/10.1007/s10278-013-9593-8
11. Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. *Acad Radiol* 2004; **11**: 272–80. doi: https://doi.org/10.1016/S1076-6332(03)00719-0
12. Uniyal N, Eskandari H, Abolmaesumi P, Sojoudi S, Gordon P, Warren L, et al. Ultrasound RF time series for classification of breast lesions. *IEEE Trans Med Imaging* 2015; **34**: 652–61. doi: https://doi.org/10.1109/TMI.2014.2365030
13. Singh BK, Verma K, Thoke AS. Adaptive gradient descent backpropagation for classification of breast tumors in ultrasound imaging. *Procedia Comput Sci* 2015; **46**:

1601–9. doi: https://doi.org/10.1016/j.procs.2015.02.091
14. Shi J, Zhou S, Liu X, Zhang Q, Lu M, Wang T. Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. *Neurocomputing* 2016; **194**: 87–94. doi: https://doi.org/10.1016/j.neucom.2016.01.074
15. Shan J, Alam SK, Garra B, Zhang Y, Ahmed T. Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods. *Ultrasound Med Biol* 2016; **42**: 980–8. doi: https://doi.org/10.1016/j.ultrasmedbio.2015.11.016
16. Ardakani AA, Gharbali A, Mohammadi A. Classification of breast tumors using sonographic texture analysis. *J Ultrasound Med* 2015; **34**: 225–31. doi: https://doi.org/10.7863/ultra.34.2.225
17. Chen CM, Chou YH, Han KC, Hung GS, Tiu CM, Chiou HJ, et al. Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks. *Radiology* 2003; **226**: 504–14. doi: https://doi.org/10.1148/radiol.2262011843
18. Kim SM, Han H, Park JM, Choi YJ, Yoon HS, Sohn JH, et al. A comparison of logistic regression analysis and an artificial neural network using the BI-RADS lexicon for ultrasonography in conjunction with introbserver variability. *J Digit Imaging* 2012; **25**: 599–606. doi: https://doi.org/10.1007/s10278-012-9457-7
19. Sun J, Wyss R, Steinecker A, Glocker P. Automated fault detection using deep belief networks for the quality inspection of electromotors. *Tm - Technisches Messen* 2014; **81**: 255–63. doi: https://doi.org/10.1515/teme-2014-1006
20. Sickles E, D'Orsi C, Bassett L, Appleton C, Berg W, Burnside E. *ACR bi-rads atlas, breast imaging reporting and data system*. Reston, VA: American College of Radiology; 2013. pp. 39–48.
21. Bengio Y. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning* 2009; **2**: 1–127. doi: https://doi.org/10.1561/2200000006
22. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest Radiol* 2017; **52**: 434-440. doi: https://doi.org/10.1097/RLI.0000000000000358
23. Carlsson G, Ishkhanov T, de Silva V, Zomorodian A. On the local behavior of spaces of natural images. *Int J Comput Vis* 2008; **76**: 1–12. doi: https://doi.org/10.1007/s11263-007-0056-x
24. Chen L, Wang S, Fan W, Sun J, Naoi S, and IEEE. 2015. Beyond human recognition: a cnn-based framework for handwritten character recognition. In: pattern recognition (acpr), 2015. 3rd iapr asian conference on. 695–9.
25. Kerr E, McGinnity TM, Coleman S, and IEEE. 2014. Material classification based on thermal and surface texture properties evaluated against human performance. In: control automation robotics & vision (icarcv), 2014. 13th international conference on. 444–9.
26. Baumgartner CF, Kamnitsas K, Matthew J, Fletcher TP, Smith S, Koch LM, et al. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Trans Med Imaging* 2017; **36**: 2204–15. doi: https://doi.org/10.1109/TMI.2017.2712367
27. Surov A, Holzhausen HJ, Wienke A, Schmidt J, Thomssen C, Arnold D, et al. Primary and secondary breast lymphoma: prevalence, clinical signs and radiological features. *Br J Radiol* 2012; **85**: e195–e205. doi: https://doi.org/10.1259/bjr/78413721
28. Drew T, Võ ML, Wolfe JM. The invisible gorilla strikes again: sustained inattentional blindness in expert observers. *Psychol Sci* 2013; **24**: 1848–53. doi: https://doi.org/10.1177/0956797613479386
29. Hofmanninger J, and Langs G. Mapping visual features to semantic profiles for retrieval in medical imaging. CVPR. 2015; 457–65.