Deep Learning in Mammography
Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer

Anton S. Becker, MD, Magda Marcon, MD, Soleen Ghafoor, MD, Moritz C. Wurnig, MD, MSc, Thomas Frauenfelder, MD, and Andreas Boss, MD, PhD

Objectives: The aim of this study was to evaluate the diagnostic accuracy of a multipurpose image analysis software based on deep learning with artificial neural networks for the detection of breast cancer in an independent, dual-center mammography data set.

Materials and Methods: In this retrospective, Health Insurance Portability and Accountability Act-compliant study, all patients undergoing mammography in 2012 at our institution were reviewed (n = 3228). All of their prior and follow-up mammographies from a time span of 7 years (2008–2015) were considered as a reference for clinical diagnosis. After applying exclusion criteria (missing reference standard, prior procedures or therapies), patients with the first diagnosis of a malignoma or borderline lesion were selected (n = 143). Histology or clinical long-term follow-up served as reference standard. In a first step, a breast density- and age-matched control cohort was selected (n = 143) from the remaining patients with more than 2 years follow-up (n = 1003). The neural network was trained with this data set. From the publicly available Breast Cancer Digital Repository data set, patients with cancer and a matched control cohort were selected (n = 35 × 2). The performance of the trained neural network was also tested with this external data set. Three radiologists (3, 5, and 10 years of experience) evaluated the test data set. In a second step, the neural network was trained with all cases from January to September and tested with cases from October to December 2012 (screening-like cohort). The radiologists also evaluated this second test data set. The areas under the receiver operating characteristic curve between readers and the neural network were compared. A Bonferroni-corrected P value of less than 0.016 was considered statistically significant.

Results: Mean age of patients with lesion was 59.6 years (range, 35–88 years) and in controls, 59.1 years (35–83 years). Breast density distribution (A/B/C/D) was 21/59/42/21 and 22/60/41/20, respectively. Histologic diagnoses were invasive ductal carcinoma in 90, ductal in situ carcinoma in 13, invasive lobular carcinoma in 13, mucinous carcinoma in 3, and borderline lesion in 12 patients. In the first step, the area under the receiver operating characteristic curve of the trained neural network was 0.81 and comparable on the test cases 0.79 (P = 0.63). One of the radiologists showed almost equal performance (0.83, P = 0.17), whereas 2 were significantly better (0.91 and 0.94, P < 0.016). In the second step, performance of the neural network (0.82) was not significantly different from the human performance (0.77–0.87, P > 0.016); however, radiologists were consistently less sensitive and more specific than the neural network.

Conclusions: Current state-of-the-art artificial neural networks for general image analysis are able to detect cancer in mammographies with similar accuracy to radiologists, even in a screening-like cohort with low breast cancer prevalence.

Key Words: mammography, breast cancer, artificial neural network, artificial intelligence, machine learning, deep learning, diagnostic accuracy

(*Invest Radiol* 2017;52: 434–440)

Despite recent advances in breast ultrasound and magnetic resonance imaging, digital mammography still constitutes the mainstay in diagnostic breast imaging around the world.[1] However, the diagnostic performance of experienced radiologists to detect the index cancer in a large screening study was only moderate, particularly in women with heterogeneously or extremely dense breasts.[2] Recently, mammography screening programs have been criticized due to the high recall rate and high rate of false-positives resulting in unnecessary biopsies.[3,4] Although the increasing use of adjunctive new techniques such as tomosynthesis may hopefully help to decrease recall rates,[5] further efforts are necessary to improve the diagnostic accuracy in breast imaging,[6] considering not only monetary costs but primarily the substantial psychological burden for women receiving false-positive findings[7] even years after cancer has been ruled out.[8]

Interpreting the different patterns in mammography is challenging and requires a high level of specialization, routine, and experience. Mammograms, being single-slice projection images, represent an ideal target to train artificial neural networks (ANNs) to acquire such routine and experience. In the past, there were 2 main obstacles for the implementation of ANNs in the clinical routine. First, the training of a deep learning with ANN (dANN) capable of analyzing high-resolution mammograms requires large computing power. In recent years, it has become feasible to run such complicated dANN computations on a conventional radiology reporting workstation with a consumer-grade graphics processor unit. Consequently, specialized deep learning models show promising first results in the evaluation of diagnostic images.[9] Second, ANNs in general need very large data sets for training. Unfortunately, assembling high-quality data sets, that is, with histological proof and pixel-wise annotation of the ground truth in the image, is a tedious and time-intensive task. On the other hand, large but low-quality data sets result in a "garbage-in, garbage-out" symptomatology. The same problem also arises in quality control of industrial manufacturing lines with a limited amount of specific training data for 1 facility or product. One advantage of dANN over other algorithms is the generalizability to other domains, that is, dANN-based image analysis software for quality control in industrial manufacturing should also be able to learn to analyze medical images.

The purpose of this retrospective cohort study was to evaluate the diagnostic accuracy of a multipurpose image analysis software based on dANN for the detection of breast cancer in mammography in an independent, dual-center data set.

MATERIALS AND METHODS

Study Population

Internal Cohort

This retrospective, Health Insurance Portability and Accountability Act-compliant study was approved by the local ethics committee, who

Received for publication December 7, 2016; and accepted for publication, after revision, January 6, 2017.

From the Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, Zurich, Switzerland.

Conflicts of interest and sources of funding: none declared.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.investigativeradiology.com).

Correspondence to: Anton S. Becker, MD, Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, Raemistrasse 100, 8091 Zurich, Switzerland. E-mail: anton.becker@usz.ch.

ISSN: 0020-9996/17/5207–0434
DOI: 10.1097/RLI.0000000000000358

waived the need for informed consent. All patients undergoing mammography in the year 2012 at our institution were reviewed (n = 3228) for lesions with therapeutic consequence (malignancy as well as borderline lesions). All of their prior and follow-up mammograms from a time span of 7 years (2008–2015) were considered as a reference for clinical diagnosis. Patients who had undergone any kind of surgical intervention before the first mammogram at our institution were excluded (n = 972). This stringent criterion was applied to avoid training the dANN for postinterventional changes and obtaining a falsely high diagnostic performance. In all cases with no malignancy, patients who had less than 2 years of follow-up examinations were excluded (n = 1101). Two cases of breast cancer were excluded due to missing histopathological report. Two patients with another malignancy with metastasis in and direct invasion of the breast were excluded. Age and breast density according to the ACR BI-RADS lexicon 2013[10] were systematically recorded for all included patients. The eligible study population is composed of 1003 patients without breast cancer and 143 patients with histology-proven invasive breast cancer or another lesion (borderline lesion or carcinoma in situ) with therapeutic consequence as shown in the flowchart in Figure 1. Diagnosis by histopathology was established within 3 weeks in all cases. Before image analysis, 4 views from the pathologic group had to be removed: 2 in which the pathology was outside of the field of view and 2 with a visible implanted pectoral pacemaker or port. One hundred thirty-seven patients had unilateral and 6 had bilateral disease.

External Cohort

An external test cohort was obtained from the publicly available Breast Cancer Digital Repository (BCDR) data sets of digital mammograms.[11–13] All 35 available patients with biopsy-proven malignant lesions were used. An age- and breast density–matched control cohort (n = 35) was selected from patients without malignant lesions with the nearest neighbor algorithm.[14] The complete list of BCDR cases is given in Supplementary Table 1 , Supplemental Digital Content 1, http://links.lww.com/RLI/A310.

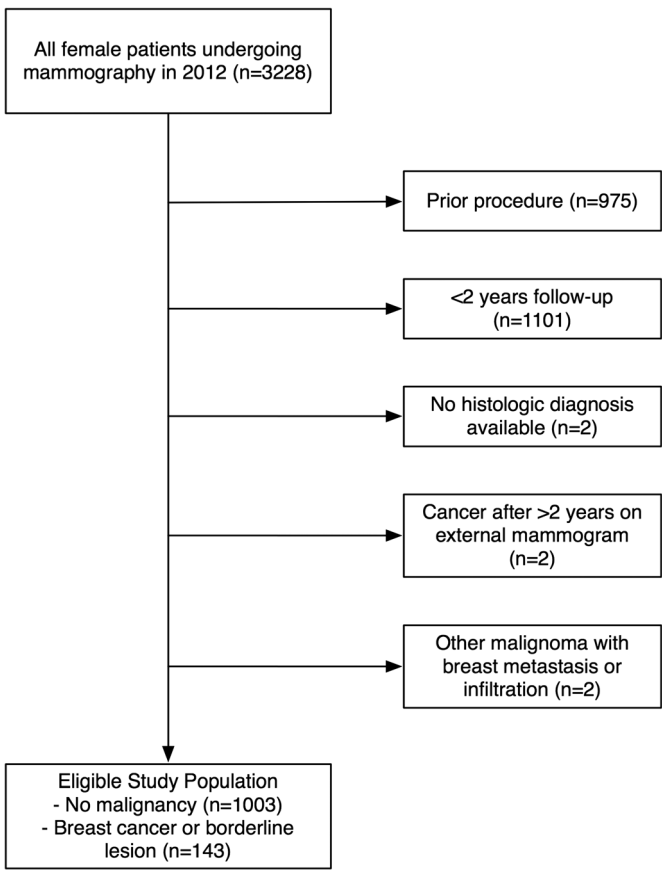

FIGURE 1. Flowchart of the patient selection process.

Study Design

Study 1: Performance of the ANN in a Cohort With High Frequency of Breast Cancer

The 143 patients with known breast cancer from the internal cohort and 143 control cases from the healthy population were matched for breast tissue density and age with the nearest neighbor algorithm[14] and served as training cases. The external test cohort (study cohort 1: 35 cancers, 35 matched controls) was used to assess the final performance of the resulting ANN model.

Study 2: Performance in a Typical Screening Cohort

For the second study, all eligible cases from January to September 2012 (125 cancer cases, 770 controls) were used for training, with cases from the months October to December as test cases to simulate a pseudoprospective setting without matching or selection of controls (study cohort 2: 18 cancer cases, 233 controls).

A human readout of the test data sets (study cohorts 1 and 2) served as reference standard.

Deep Learning Analysis

For the image analysis, we used a multipurpose image analysis software (ViDi Suite Version 2.0; ViDi Systems Inc, Villaz-Saint-Pierre, Switzerland). The software, in the following paragraph referred to as "ViDi" and afterwards generalized as "dANN," uses deep learning algorithms to detect and classify anomalies.[15,16] It is used for quality inspection purposes in the fabrication of solar panels, textiles, and various high-precision mechanical parts with complex shapes, such as watch parts and medical screws. It is currently not approved by any entity for diagnostic use in the clinical routine. Deep learning or deep neural networks are different from conventional "shallow" neural networks, where a large single layer of neurons is directly connected to the output layer, in that they contain 1 or more so-called hidden layers not directly connected to the output neurons, which enables them to solve much more complex problems.[15] All computations were performed on a GeForce GTX 960 graphics processor unit.

Training of the dANN

In all pathologic cases of the training data set, the anomalies were manually marked pixel-wise within ViDi by A.S.B, according to the description in the radiology report and annotations on the images and subsequently verified by A.B. To avoid overfitting when training the dANN on small data sets, the software offers extensive data augmentation options (perturbations), which include transformations in scale, rotation, luminance/contrast, aspect ratio, and shearing. The options we used are summarized in Supplementary Table 2, Supplemental Digital Content 2, http://links.lww.com/RLI/A311. ViDi furthermore uses proprietary image filtering and smart/adaptive sampling procedures to optimize training of neural networks on small data samples, as well as several training algorithms described in the literature.[17–19] ViDi uses a circular "virtual lens," which sweeps the image. The content of the circular region as well as an area roughly 5 times the size of the region is trained or evaluated by the underlying dANN for anomalies, although the surrounding area is weighted at a lower priority as illustrated in Supplementary Figure 1, Supplemental Digital Content 3, http://links.lww.com/RLI/A314. A total score from 0 (not pathologic) to 1 (pathologic) for the whole image is computed as well as a heat map overlay with suspicious anomalies highlighted. Training was performed with a 2:1 split

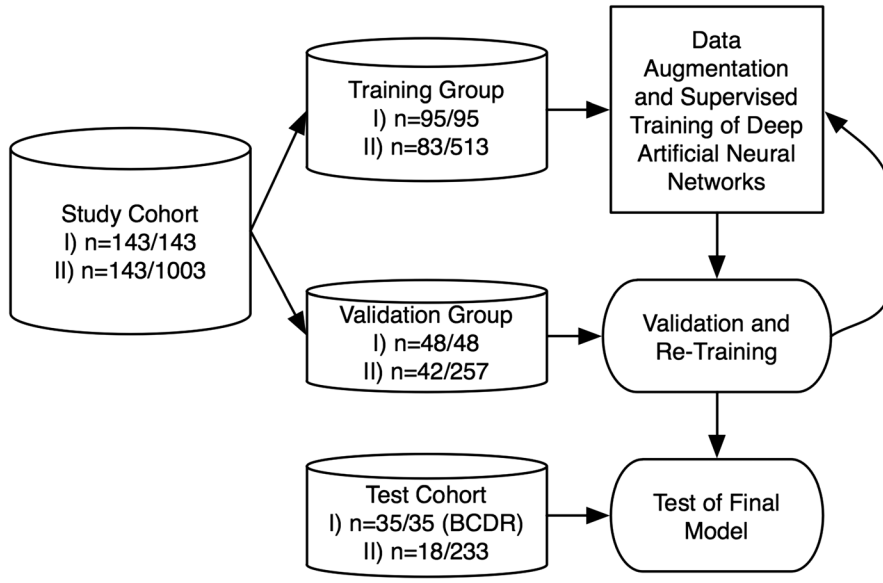

FIGURE 2. Schematic of the training/validation/test process used in this study.

between training and validation set (Fig. 2). The validation set is used to prevent overfitting of the dANN on the training data. Total training time as well as time to evaluate each test image with the trained dANN was measured.

Human Readout

The mammographies from the study cohorts were presented on a standard reporting workstation in random order to 3 radiologists (A.B, 7 years; T.F, 10 years; and M.M, 3 years of experience in breast imaging) who rated the images on a 5-point Likert-type scale for malignancy (corresponding to the Breast Imaging Reporting And Data System classification with 5 meaning >98% probability of breast cancer). There were 2 separate readout sessions for study 1 and 2, respectively. A.B performed the readout more than 3 months after having verified the annotations. Interreader agreement was assessed with the intraclass correlation coefficient (ICC)[20] and interpreted as follows: excellent agreement, >0.75; good agreement, 0.59–0.75; fair agreement, 0.40–0.58; and poor agreement, <0.40 after the suggestion of Cicchetti and Sparrow.[21]

Statistical Analysis

The statistical analysis was performed in R version 3.3.1 (R Foundation for Statistical Computing, Vienna, Austria). Continuous variables were expressed as mean and standard deviation and range, categorical variables as counts or percentages. Receiver operating characteristic (ROC) analysis was performed for the dANN and the human readers, overall and by breast density. Diagnostic accuracy was expressed as the area under the ROC curve (AUC) and compared with the nonparametric test by DeLong et al,[22] and the method proposed by Obuchowski[23] where applicable, which accounts for multiple views per patient. Sensitivity and specificity were calculated at the optimal cutoff (Youden index). A P value below 0.05 was considered indicative of significant differences, with Bonferroni correction for multiple comparisons where appropriate.

RESULTS

Study Cohort

Mean age of the internal study cohort was 59.6 ± 11.7 years (range, 35–88 years) in cancer cases, 59.1 ± 11.0 years (range, 35–83 years) in the matched controls of cohort 1, and 56.6 ± 9.3 years (range, 32–85 years) in the complete healthy cohort. Of the cancer cases, 68 patients were examined as part of their routine screening, 63 presented with a palpable nodule, 1 patient had bloody discharge, and in 10 patients, the cancer was incidentally detected in another imaging modality. In 1 patient, the indication was not reliably extractable from the records. The final histologic diagnoses were ductal in situ carcinoma in 25 lesions, invasive ductal carcinoma in 90, invasive lobular carcinoma in 13, mucinous carcinoma in 3, and borderline in 12 lesions, respectively. Median maximal lesion diameter was 15 mm (interquartile range, 10–25 mm).

Mean age of the external cohort was 56.3 ± 11.6 years (31–82 years) in cancer cases and 58.8 ± 13.1 years (31–88 years) in the matched controls. The number of patients per breast density category is shown in Table 1.

Detected Features

The heat maps showed that small mass lesions were often reliably identified if surrounding spiculae were present (Fig. 3) or in vicinity to grouped microcalcifications (Figs. 4, 5). Large masses and intramammary lymph nodes were also readily identified due to their higher density. Moreover, dense areas close to the surface were also rated as more suspicious if there was thickening and/or retraction of the overlaying cutis (Fig. 4), which, however, potentially resulted in overdiagnosis of dense glandular tissue close to the cutis (Fig. 5). Asymmetry could not be assessed, as the analysis is performed on a per-image basis.

Study 1: Performance of the ANN in a Symmetric Cohort With High Frequency of Breast Cancer

Training of the dANN was completed successfully for all images. Total training time was 59 minutes. Processing time per test image for generation of score and heat map overlay was 139.1 ± 45 milliseconds.

Diagnostic accuracy on the training data was AUC of 0.81 (95% confidence interval [CI], 0.78–0.84), and the performance of the trained dANN on the external test cohort (BCDR) was AUC of 0.79 (95% CI, 0.71–0.86), as displayed in Figure 6A, which however was not significantly different with $P = 0.63$. At the optimal cutoff point, sensitivity/specificity was 59.8/84.4% in the training data and 71.6/69.6% in cohort 1. Diagnostic accuracy was the highest in breasts with the lowest

TABLE 1. Breast Density Distribution of the Cohorts From Our Population

Density	A	B	C	D
Pathological cohort	21	59	42	21
Matched control cohort	22	60	41	20
Full control cohort	101	268	203	102

density (AUC_A = 0.98) and markably lower in more dense breasts (AUC_B = 0.79, AUC_C = 0.77, AUC_D = 0.74).

All readers completed the readout successfully under 60 minutes (reader 1, 37 minutes; reader 2, 45 minutes; reader 3, 43 minutes). Diagnostic accuracy was not significantly different among readers (AUC = 0.83, 0.91, and 0.94; P = 0.017, 0.06, and 0.42); however, 2 readers performed significantly better when compared with the dANN (P = 0.17, = 0.003, and < 0.0001). The sensitivity/specificity was 84.1/97.0% for reader 1, 82.6/74.6% for reader 2, and 97.1/88.1% for reader 3. The diagnostic accuracy for the different breast densities ranged between 0.68 and 1.0 and is summarized in Table 2; due to the largely overlapping CIs, further statistical testing was omitted. Interreader agreement was excellent (ICC = 0.85).

Study 2: Performance in a Typical Screening Cohort

dANN training was completed in 4:38 hours. Processing time per test image was 145.0 ± 50 milliseconds.

Training accuracy was AUC of 0.85 (95% CI, 0.82–0.87); performance of the trained model on the realistic screening population was AUC of 0.82 (95% CI, 0.75–0.89) with an optimal sensitivity/specificity of 73.7/72.0%. Diagnostic accuracy was again the highest in low-density breasts (AUC_A = 0.94) and substantially lower in more dense breasts (AUC_B = 0.83, AUC_C = 0.79, AUC_D = 0.79).

Readout times were 49 minutes (reader 1), 57 minutes (reader 2), and 130 minutes (reader 3). Diagnostic accuracy measured by AUC was not significantly different between the readers (AUC = 0.79, 0.77, and 0.87; P = 0.18, 0.32, and 0.83) or the ANN (P = 0.45, 0.56, and 0.62). However, all readers exhibited higher specificity but lower sensitivity when compared with the ANN, with a sensitivity/specificity of 60.0%/94.4% for reader 1, 60.0%/93.6% for reader 2, and 80.0%/90.2% for reader 3. This effect can also be appreciated in the ROC curves in Figure 6B. The performance by breast density of the readers

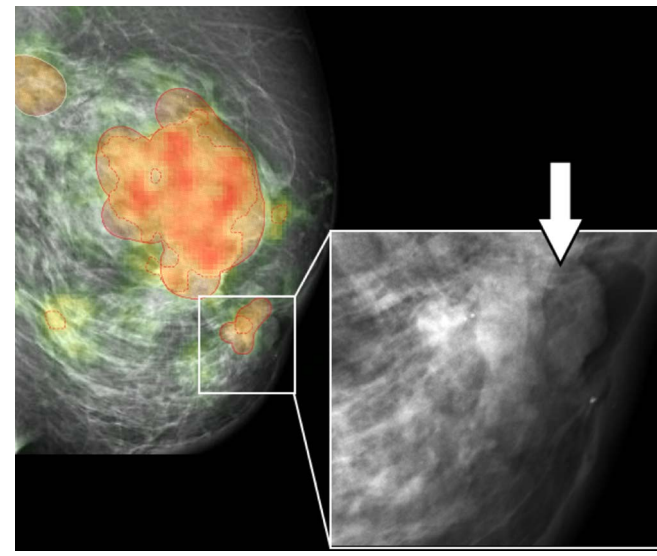

FIGURE 4. A 69-year-old patient with a large multicentric invasive ductal carcinoma. The smaller satellite lesion was correctly identified by the neural network (solid lines indicates manual outline; dashed lines, neural network detection), possibly due to the adjacent skin thickening or nipple retraction (arrow in magnification view). The score for this view was 0.85.

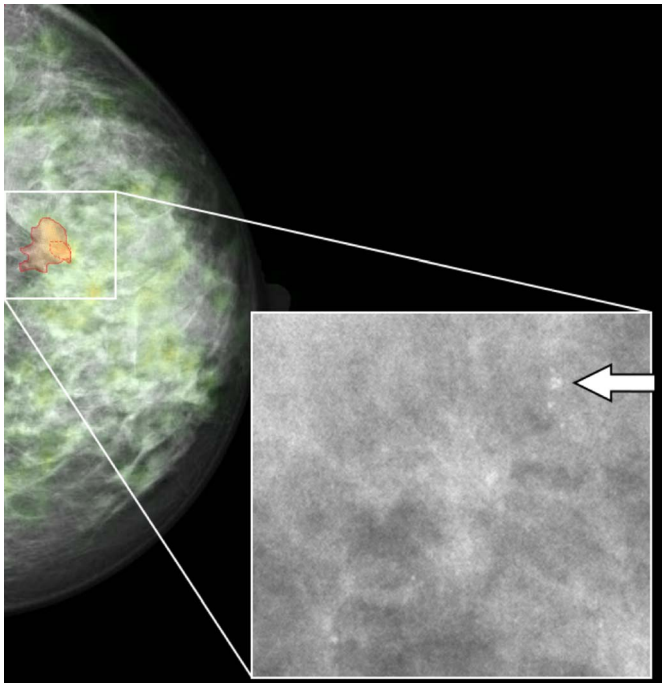

FIGURE 3. A 46-year-old patient with invasive ductal carcinoma. Because of the spiculated border and surrounding microcalcifications (arrow in magnification view), the neural network detected the lesion (solid line indicates manual outline; dashed line, neural network detection) despite the small size and generally dense glandular tissue. The red border is added in pictures above the threshold for suspicious breasts (score for this picture was 0.59).

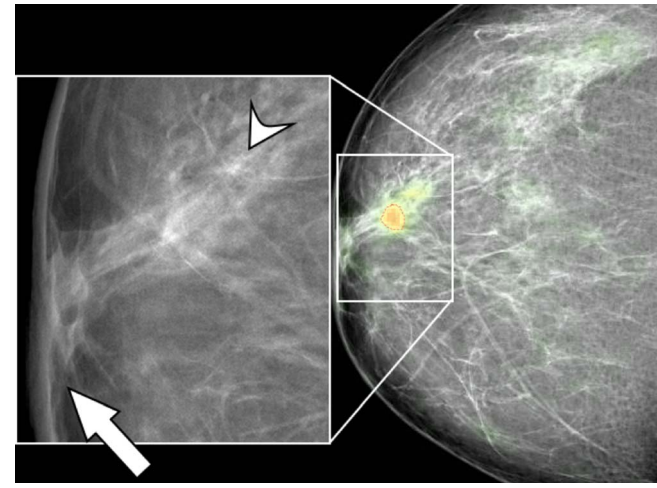

FIGURE 5. Example of a false-positive finding in a 44-year-old female from the external test cohort (BCDR-DN01). The high score of 0.68 was indicative of a suspicious breast finding. The focal glandular tissue was misclassified as suspicious due to its focal appearance, the adjacent microcalcifications (arrowhead), and possibly also the thickened appearance of the cutis (arrow). Image courtesy of Dr A.M. Guevara Lòpez, University of Porto, Portugal.

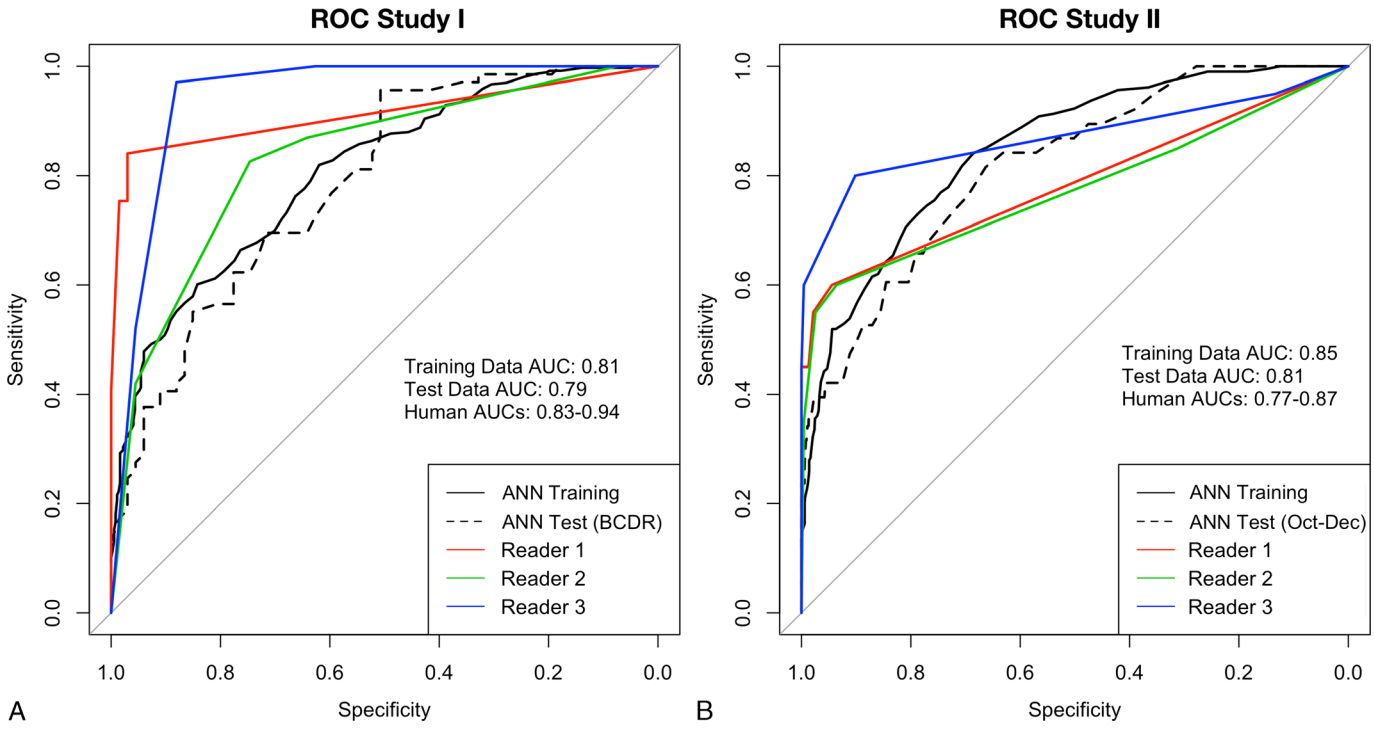

FIGURE 6. A, Receiver operating characteristic curve of the local study cohort and the BCDR test data set (black lines) as well as the human readers (red, blue, and green lines) with the corresponding AUCs. B, ROC curve of the whole 2012 study population (black) as well as the performance of the human readers (red, blue, and green lines) on the test cohort (October to December 2012).

is summarized in Table 3. The ICC was again excellent (0.81) indicating no gross disagreement between readers.

DISCUSSION

In this study, we were able to demonstrate that general-purpose dANN built for generic image analysis are able to identify pathologic patterns in mammography. The dANN software could be trained with a data set in the order of 140 pathological mammographies in comparison to matched healthy controls and provided similar diagnostic accuracy compared with experienced human readers.

As summarized in Table 2, the diagnostic accuracy achieved by our dANN is comparable to other recently published deep learning models.[9,24,25] In contrast to the present article, these studies make use of specialized dANNs for breast cancer detection. A notable exception is the study of Ertosun and Rubin,[24] who compare several different generic dANNs. They are also the only ones who make use of probabilistic maps, similar to the heat map overlay shown in the present study. However, they still used a rather large amount of data (2250 cases) to train their neural networks, as do most other machine learning studies who achieved similar or higher diagnostic accuracy.[26] This is problematic, because only very few high quality are publicly available and they usually contain much fewer cases.[11] Moreover, many algorithms require preprocessing by either manual or automatic segmentation to remove extracorporeal air or the pectoral muscle.[9,25,27] In the latter case, important diagnostic information (peripheral tumors, pathologic lymph nodes) may even be lost. Another limitation of previous studies is that the dANNs are only trained with rectangular tiles of the given images. One of the advantages of the software tested in the present study is the evaluation of a circular area while simultaneously taking into account the encompassing tissue. This is probably the reason that even small lesions are identified if suspicious features are present in the vicinity (Figs. 3, 4), although this also lead to lower specificity when compared with human readers (Figs. 5, 6B). Lastly, none of these studies offer a direct comparison to a radiologists' performance.

Computer-assisted detection (CADe) in mammography is nothing new, although the traditional CADe systems rely on other machine learning algorithms and hence need large amounts of reference data.[24] The underlying engine in the present study, however, is a novel more versatile form of artificial intelligence, which is applicable to a wide variety of problems as evident by its use in different industries. Although a comparison of specialized software would be interesting, this is beyond

TABLE 2. Diagnostic Performance of Previously Published Deep ANN Models in the Detection of Breast Cancer

Year	First Author	ANN Model	No. Training Cases	AUC
2015	Ertosun and Rubin[24]	GoogLeNet	2250	0.85
2015	Ertosun and Rubin[24]	AlexNet	2250	0.84
2015	Ertosun and Rubin[24]	VGG-Net 16	2250	0.82
2016	Present Study	ViDi Red	286	0.81
2016	Sun et al[9]	4 Convolutional ANN and 1 fully connected layer[9]	840	0.70
2016	Kallenberg et al[25]	Convolutional Sparse Autoencoder[25]	2244	0.61

TABLE 3. Diagnostic Performance by Breast Density

		Breast Density			
	Reader	**A**	**B**	**C**	**D**
Study 1	1	0.92 (0.83–1.0)	0.90 (0.74–1.0)	0.90 (0.83–0.98)	1.0 (NA)
	2	0.78 (0.63–0.94)	0.99 (0.97–1.0)	0.71 (0.57–0.84)	0.68 (0.30–1.0)
	3	0.92 (0.82–1.0)	0.98 (0.94–1.0)	0.98 (0.95–1.0)	0.88 (0.70–1.0)
	ANN	0.98 (0.96–1.0)	0.79 (0.74–0.84)	0.77 (0.70–0.84)	0.74 (0.63–0.84)
Study 2	1	0.89 (0.68–1.0)	0.91 (0.74–1.0)	0.64 (0.43–0.85)	0.62 (0.26–0.98)
	2	0.88 (0.63–1.0)	0.93 (0.80–1.0)	0.45 (0.19–0.72)	0.84 (0.53–1.0)
	3	1.0 (0.98–1.0)	0.75 (0.42–1.0)	0.89 (0.72–1.0)	0.79 (0.46–1.0)
	ANN	0.94 (0.87–1.0)	0.84 (0.74–0.96)	0.69 (0.57–0.82)	0.69 (0.47–0.92)

Values are presented as AUC (95% CI).

the scope of the current study. Moreover, the use of the latest, nonapproved technology may allow a realistic estimate of what the diagnostic radiologist may expect from similar tools in the coming years. In its current form, the software still tends to be too sensitive eg, towards some dense glandular formations obvious to the human reader; however, it shows greater sensitivity for more subtle lesions. A potential immediate usage may thus be the one of a "second look," as current CADe systems are used. The clear advantages are the speed and the ability of the neural network to be trained and adapted to the local patient population. Hence, it may be interesting for centers caring for high-risk patients.[28]

Our study has several limitations that need to be acknowledged. First, as a large university hospital in [the canton of Zurich] without a government-funded screening program, our patient population is different than in centers who perform mainly primary screening. A large quantity of patients are referrals as evident by the large number we had to exclude due to prior procedures or external mammograms. This in combination with the retrospective study design entails a high danger of selection bias. We thus emphasize the need for further studies in other populations. In addition, we are currently planning to validate our results in a prospective study.

Second, although we did not include any patients with postsurgical changes in our cohort, the BCDR control only included minimal patient history and may thus contain patients with subtle scary changes, which would introduce a bias.

Third, experienced radiologists still tended to outperform this ANN in the experimental data set with a 1:1 ratio of controls to cancers and had a higher specificity when evaluating potentially suspicious patterns. Because the implementation of the software used for this study is not optimized for mammography, it neither "understands" the concept of multiple views per patient nor laterality nor time evolution. The importance of these concepts is reflected in the evaluation criteria of risk stratification by the Breast Imaging Reporting And Data System lexicon, and first attempts do aid diagnosis by asymmetries have been made.[29] It may be that these new concepts will be good targets for recurrent ANNs, which have recently been successfully used to segment a magnetic resonance imaging of the brain, that is, a 3-dimensional data set with multiple sequences.[30] Furthermore, the inclusion of clinical and bioptic data may improve accuracy of the resulting model.[31] However, it may also be hypothesized that generic dANN will show a faster development in the future compared with dedicated radiological software due to the scale effects with other applications.

In conclusion, we showed that deep learning algorithms designed for generic image analysis can be trained to detect breast cancer on mammography data with high diagnostic accuracy (AUC = 0.82) comparable to experienced radiologists (AUC = 0.79–0.87). Further optimization of the neural network architecture, the chaining of several neural networks into a purposeful workflow, and prospective studies in the target populations will be needed to prove the clinical usefulness of deep learning in the detection and diagnostic workup of breast cancer.

ACKNOWLEDGMENTS

The authors are in no way affiliated with nor do they have any financial stakes in ViDi Systems Inc. An academic license for the software (ViDi Suite) was purchased for the normal price without any discount. ViDi Suite is currently not approved by the Food and Drug Administration or any other entity for diagnostic use. The authors thank Reto Wyss, PhD, from ViDi Systems Inc for his help and sharing his expertise in image analysis.

REFERENCES

1. Kuhl CK. The changing world of breast cancer: a radiologist's perspective. *Invest Radiol*. 2015;50:615–628.
2. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*. 2005;353:1773–1783.
3. Hubbard RA, Kerlikowske K, Flowers CI, et al. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study. *Ann Intern Med*. 2011;155:481–492.
4. Age Trial Management GroupJohns LE, Moss SM. False-positive results in the randomized controlled trial of mammographic screening from age 40 ("Age" trial). *Cancer Epidemiol Biomarkers Prev*. 2010;19:2758–2764.
5. Haas BM, Kalra V, Geisel J, et al. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. *Radiology*. 2013;269:694–700.
6. Peters S, Hellmich M, Stork A, et al. Comparison of the detection rate of simulated microcalcifications in full-field digital mammography, digital breast tomosynthesis, and synthetically reconstructed 2-dimensional images performed with 2 different digital x-ray mammography systems. *Invest Radiol*. 2016. [Epub ahead of print].
7. Salz T, Richman AR, Brewer NT. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psychooncology*. 2010;19:1026–1034.
8. Brodersen J, Siersma VD. Long-term psychosocial consequences of false-positive screening mammography. *Ann Fam Med*. 2013;11:106–115.
9. Sun W, Tseng T-LB, Zheng Bin, et al. A preliminary study on breast cancer risk analysis using deep neural network. In: *Breast Imaging*. Vol 9699. Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing; 2016:385–391.
10. Sickles EA, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS® Atlas, Breast imaging reporting and data system. *J Am Coll Radiol*. 2013:39–48.
11. Moura DC, Lopez MAG, Cunha P, et al. Benchmarking datasets for breast cancer computer-aided diagnosis (CADx). In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Vol 8258. Lecture Notes in Computer Science. Berlin, Germany: Springer Berlin; 2013:326–333.
12. Moura DC, Guevara López MA. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *Int J Comput Assist Radiol Surg*. 2013;8:561–574.

13. Ramos-Pollán R, Guevara-López MA, Suárez-Ortega C, et al. Discovering mammography-based machine learning classifiers for breast cancer diagnosis. *J Med Syst*. 2012;36:2259–2269.
14. Ho DE, Imai K, King G, et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*. 2007;15:199–236.
15. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
16. Bengio Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*. 2009;2:1–127.
17. LeCun YA, Bottou L, Orr GB, et al. Efficient BackProp. In: *Neural Networks: Tricks of the Trade*. Vol 7700. Lecture Notes in Computer Science. Berlin, Germany: Springer Berlin Heidelberg; 2012:9–48.
18. Hinton GE. To recognize shapes, first learn to generate images. *Prog Brain Res*. 2007;165:535–547.
19. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput*. 2006;18:1527–1554.
20. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428.
21. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic*. 1981;86:127–137.
22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.
23. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics*. 1997;53:567–578.
24. Ertosun MG, Rubin D. Probabilistic visual search for masses within mammography images using deep learning. In: *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference*. New York, NY: IEEE; 2015:S1310–S1315.
25. Kallenberg M, Petersen K, Nielsen M, et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Trans Med Imaging*. 2016;35:1322–1331.
26. Dhungel N, Carneiro G, Bradley AP. Automated mass detection in mammograms using cascaded deep learning and random forests. In: *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference*. New York, NY: IEEE; 2015:1–8.
27. Guo Y, Dong M, Yang Z, et al. A new method of detecting micro-calcification clusters in mammograms using contourlet transform and non-linking simplified PCNN. *Comput Methods Programs Biomed*. 2016;130:31–45.
28. Sardanelli F, Podo F, Santoro F, et al. Multicenter surveillance of women at high genetic breast cancer risk using mammography, ultrasonography, and contrast-enhanced magnetic resonance imaging (the high breast cancer risk Italian 1 study): final results. *Invest Radiol*. 2011;46:94–105.
29. Casti P, Mencattini A, Salmeri M, et al. Towards localization of malignant sites of asymmetry across bilateral mammograms. *Comput Methods Programs Biomed*. 2017;140:11–18.
30. Stollenga MF, Byeon W, Liwicki M, et al. Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In: *Advances in Neural Information Processing Systems*. La Jolla, CA: Neural Information Processing Systems (NIPS) Foundation; 2015:2998–3006.
31. Knuttel FM, van der Velden BH, Loo CE, et al. Prediction model for extensive ductal carcinoma in situ around early-stage invasive breast cancer. *Invest Radiol*. 2016;51:462–468.