Mammography Image Quality Assurance Using Deep Learning

Tobias Kretz, Klaus-Robert Müller, Tobias Schaeffter, and Clemens Elster

***Abstract*—*Objective:* According to the European Reference Organization for Quality Assured Breast Cancer Screening and Diagnostic Services (EUREF) image quality in mammography is assessed by recording and analyzing a set of images of the CDMAM phantom. The EUREF procedure applies an automated analysis combining image registration, signal detection and nonlinear fitting. We present a proof of concept for an end-to-end deep learning framework that assesses image quality on the basis of single images as an alternative. *Methods:* Virtual mammography is used to generate a database with known ground truth for training a regression convolutional neural net (CNN). Training is carried out by continuously extending the training data and applying transfer learning. *Results:* The trained net is shown to correctly predict the image quality of simulated and real images. Specifically, image quality predictions on the basis of single images are of similar quality as those obtained by applying the EUREF procedure with 16 images. Our results suggest that the trained CNN generalizes well. *Conclusion:* Mammography image quality assessment can benefit from the proposed deep learning approach. *Significance:* Deep learning avoids cumbersome pre-processing and allows mammography image quality to be estimated reliably using single images.**

***Index Terms*—Deep learning, image regression, mammography image quality assessment.**

I. Introduction

MAMMOGRAPHY screening using x-ray radiation is an important diagnostic tool and routinely applied for early detection of breast cancer [1], [2]. Since cancerous tissue is denser than healthy tissue, its x-ray attenuation is slightly higher, which results in a contrast difference in the image acquired. The radiation contrast depends also on the x-ray energy spectrum and the amount of scattered radiation [3]. The contrast difference is visible in the recorded image as a signal. Increasing the radiation dose increases the signal-to-noise ratio and thus the detectability of signals in the recorded images.

While higher detectability implies improved diagnostic quality, an increase in the radiation dose can cause health risks [4]. The choice of radiation dose is thus crucial, since it should guarantee sufficient diagnostic power while keeping the dose exposure as small as possible. In order to ensure that the dose level chosen is adequate, image quality assessment can be carried out in a way suggested in a guideline of the European Reference Organization for Quality Assured Breast Cancer Screening and Diagnostic Services (EUREF) [5]. By following that guideline, at least 16 images of the contrast-detail phantom for mammography (CDMAM) [5] are acquired and subsequently analyzed. In order to determine a device's ability to resolve small details, the CDMAM phantom consists of regular structures with prescribed diameters and various thicknesses to emulate different signal-to-noise ratio levels. For the analysis of the real images, conventional signal detection methods are applied, followed by a regression using a logistic function [5]. In order to apply these methods, the position of the signals in the recorded image needs to be known, which in turn requires the additional application of pre-processing procedures.

The EUREF Guideline analysis results in a contrast-detail curve [6] that characterizes the ability to detect small structures of the phantom in dependence on their size. More precisely, a contrast-detail curve consists of twelve points assigned to images of the CDMAM phantom [7]. Each point consists of a diameter and the minimum thickness needed such that a signal evoked from a disc with that diameter and thickness will be correctly identified by an automatic signal detection procedure with high probability. The rationale for this procedure is that the ability to visualize small structures and contrasts in a technical phantom can be linked to the detection of microcalcifications in clinical images [8], [9]. EUREF distributes a software program called CDMAM Analyzer that inputs a stack of at least 16 CDMAM images and outputs the corresponding contrast-detail curve; throughout the paper, this is referred to as the EUREF Guideline procedure.

Over the past decade, deep learning has significantly influenced the development of analysis methods in medical imaging [10]–[12]. In contrast to traditional methods that usually make use of a set of hand-crafted features, deep learning can be implemented end-to-end, thus allowing representative features to be learned during the training phase. This enables the

Manuscript received October 31, 2019; revised February 3, 2020; accepted March 20, 2020. Date of publication April 14, 2020; date of current version November 20, 2020. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) and in part by the Korea government under Grants 2017-0-00451 and 2017-0-01779. The work of K.-R. Müller was supported by the German Ministry for Education and Research (BMBF) under Grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025 A, and 01IS18037 A and by DFG (EXC 2046/1, Project-ID 390685689). *(Corresponding authors: Klaus-Robert Müller and Clemens Elster.)*

Tobias Kretz and Tobias Schaeffter are with the Physikalisch-Technische Bundesanstalt.

Klaus-Robert Müller is with the TU Berlin, 10623 Berlin, Germany, with the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, South Korea and also with the Max Planck Institute for Informatics, 66123 Saarbrücken (e-mail: klaus-robert.mueller@tu-berlin.de).

Clemens Elster is with the Physikalisch-Technische Bundesanstalt, 10587 Berlin, Germany (e-mail: clemens.elster@ptb.de).

Digital Object Identifier 10.1109/TBME.2020.2983539

algorithms to also find patterns in the data that are not accessible for humans.

Meanwhile, deep learning outperforms humans in the task of image classification [13]. Several studies explore the increasing impact of deep learning on medical imaging [14]–[16]. Deep learning is especially well suited for use in Computer Aided Detection and Diagnosis (CAD), which can be used to provide a second opinion for doctors and physicians in clinical health care [17], [18]. Convolutional neural nets [19] (CNNs) now represent the state of the art in image classification [13], [20]. Because of their block structure, CNNs have far fewer parameters than standard, fully connected feedforward nets of a similar layer size [21], which makes it easier to train them.

One challenge in deep learning is to understand the behavior of a trained net. To this end, several approaches for explainability have recently been proposed in the context of classification [22]. In [23], a method is discussed which computes local gradients to analyze the sensitivity of the prediction when changing the values of single pixels. Layerwise relevance propagation (LRP) [24] produces a score for each pixel in the image that reflects its relevance for the output of the neural net, and this approach has proven in many instances to be a sensitive tool for understanding the behavior of a neural net ([25]–[28]).

To the best of our knowledge, methods from deep learning have not been applied so far in the context of mammography image quality assessment. The goal of this paper is to explore the potential of deep learning for image quality assessment in mammography. To this end, virtual mammography has been implemented to construct a database of images of the CDMAM phantom with known ground truth [29]. A regression CNN (cf. [30]–[34]) was trained to model the contrast-detail curves. No cumbersome pre-processing such as image registration was used. Subsequently, a gradient technique was applied to understand the behavior of the trained CNN, and the trained CNN was tested on independent test images of the CDMAM phantom.

The paper is structured as follows: Section II illustrates the development of a data set for training and testing a deep neural net for mammography image quality analysis, introduces the employed CNN architecture and shows the methology used for interpreting the CNN's decision. A detailed comparison of the deep learning approach with the EUREF Guideline reference procedure based on simulated and real images of the CDMAM phantom is then presented in Section III and discussed in Section IV. Finally, some conclusions are drawn.

II. Method

This section provides some background of mammography quality assurance, describes the generation of a labeled database and its separation into training and independent test data, as well as the chosen CNN architecture. Furthermore, the training procedure based on transfer learning is detailed, along with the explainability method used to visualize the behavior of the trained neural net.

A. Background Mammography Quality Assurance

In following a European guideline [5], mammography image quality is expressed in terms of the contrast-detail curve derived by applying an automated procedure for high-resolution images of the CDMAM phantom (ca. 3600 × 2300 pixels). An example image of the CDMAM phantom is shown in Fig. 1. The real image shows pixels belonging to a margin, the phantom itself, as well as annotations and grid pixels on the phantom. The phantom consists of circular structures arranged in a regular grid, where the diameter and thickness of the structures vary across the grid cells. Each grid cell consists of two discs: one located in the center and the other located in one of the four corners. The contrast-detail curve expresses the minimum thickness in dependence on a given diameter of the discs that is needed to correctly localize the corner containing the second disc with high probability. A mammography device has passed the quality assurance test if its contrast-detail curve falls below a prescribed limit curve [5].

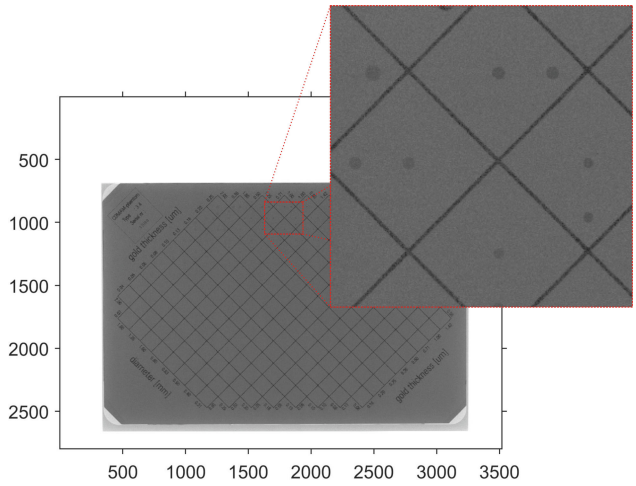


Fig. 1. Image of the CDMAM phantom. The image comprises margin (white), phantom (gray) and annotation/grid pixels (black). Circular structures with varying diameters and thicknesses are regularly arranged in a grid structure.

B. Data

We constructed a database that consists of simulated and real images of the CDMAM phantom. Application of the EUREF software program CDMAM Analyzer [5] was used to construct contrast-detail curves for all images in the database. The calculated contrast-detail curves were taken as ground truth. After having determined the ground truth, all images were downsampled for further processing by the CNN, cf. Fig. 2. Table I shows the sizes of the final training and test sets.

1) Simulated Images: An in-house tool for virtual mammography [29] was used to simulate images of the CDMAM phantom. The virtual mammography simulates a x-ray spectrum according to the target material, the voltage and the exposure time current product. A backward ray-casting procedure is applied to determine the path of the x-rays through the CDMAM phantom, and Lambert's law is used for the calculation of their absorption until they reach the detector. The characteristic curve of the detector is considered to be linear. In this way, a primary image is calculated which is further degraded by blurring, scattering and noise. Especially scattering and noise influence the radiation contrast of signals in the simulated image.

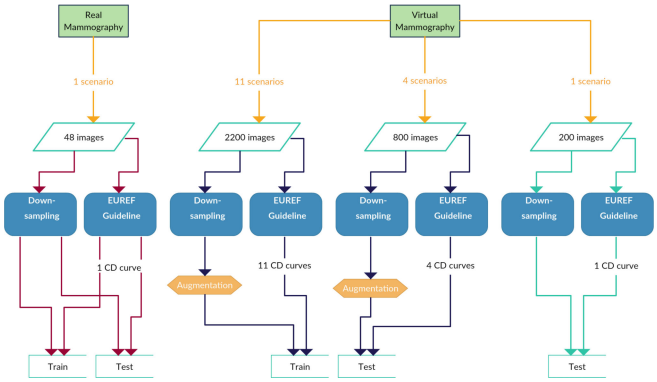


Fig. 2. Flowchart to illustrate the different data sets used to train and test a neural net for image quality assessment for simulated and real mammography images. Altogether, 17 different contrast-detail curves (CDs) are available.

TABLE I
SIZES OF THE FINAL TRAINING SET OF AUGMENTED SIMULATED AND REAL IMAGES AND THE FINAL TEST SET OF 4 PLUS 1 INDEPENDENT SIMULATION SCENARIOS AND REAL IMAGES THAT WERE NOT USED FOR THE TRAINING. FACTORS ILLUSTRATE THE MULTIPLICATION OF IMAGES THROUGH DATA AUGMENTATION

Set	Type	No. images	No. scenarios
Training	simulation	2200*6*4	11
	real	24*200	1
Test	simulation	800*6	4
	real	24	1
	simulation	200	1

TABLE II
SIMULATION PARAMETERS FOR THE 16 DIFFERENT SIMULATION SCENARIOS USED TO TRAIN AND TEST A NEURAL NET TO ESTIMATE A CONTRAST-DETAIL CURVE FROM A CDMAM PHANTOM IMAGE

Set	No.	Voltage $[kVp]$	Exposure $[mAs]$	SNR
Training	1	25	110	4.5
	2	25	110	4.9
	3	31	100	5.8
	4	25	110	5.9
	5	28	105	6.7
	6	31	100	6.7
	7	25	110	7.2
	8	27	100	7.3
	9	31	100	7.6
	10	25	100	8.5
	11	31	100	8.5
Test	1	28	115	4.9
	2	25	110	6.7
	3	25	110	9.7
	4	25	110	11.7
Generalization Test	1	23	105	3.1

Different SNRs were used in the virtual mammograms to emulate devices of varying quality with an exposure between 100–120 mAs and a tube voltage between 23–31 kVp, cf. Table II.

For each set of simulation parameters 200 images were simulated which differ in their random noise and in randomly applied shifts of the phantom. These random shifts follow a normal distribution with zero mean and standard deviation 1 px in x- and y-direction individually. We refer to these sets of 200 images simulated with one combination of tube voltage, exposure and SNR as simulation scenarios. The scenarios differ in the radiographic technique, and the corresponding simulation parameters are listed in Table II.

2) Real Images: 48 images of the CDMAM phantom recorded on a Siemens Mammomat Inspiration with a tube voltage of 30 kVp and an exposure of 110 mAs were included into the database. Fig. 1 shows one of the 48 real images.

3) Contrast-Detail Curves: Each image in the database is labeled by assigning to it a contrast-detail curve. The EUREF Guideline procedure is used to determine these contrast-detail curves. The EUREF Guideline procedure has to be applied to a group of images, and the uncertainty of the resulting contrast-detail curve depends essentially on the number of images in that group [29]. For the real images, all 48 images were used to calculate one contrast-detail curve, which was then assigned to all 48 images. For the simulated images, all 200 images belonging to the same set of simulation parameters were used to calculate a single contrast-detail curve, which was then assigned to each of the 200 images. Altogether, 16 different settings of parameters for simulating CDMAM images as listed in Table II were considered, leading altogether to the 17 different contrast-detail curves shown in Fig. 3a. In following [35], a log-log scale is used to present contrast-detail curves.

4) Downsampling: To achieve a reasonable performance when training the neural net, the dimensionality of the images was highly reduced. Random downsampling has been chosen for this purpose and all images were randomly downsampled to a size of 250×250 pixels. One reason for the choice of random downsampling is its simplicity in future applications as it does not require further preprocessing steps needed when using other dimension reduction methods such as, for example, PCA-based methods [36] or shearlet transforms [37]. Another possibility of reducing dimensionality is patch-wise sampling, where the original image is split into smaller patches. However, the contrast-detail curve summarizes local information from the whole image which can be challenging when applying patch-wise sampling, since each patch provides only a small portion of the whole image. The size of 250×250 pixels for the down-sampled images was chosen after comparing the performance of training and prediction for several different larger and smaller sizes. Fig. 4 shows examples of downsampled images.

5) Data Augmentation and Normalization: Simulated images of the CDMAM phantom used for training or testing were augmented by adding artificial margins and subsequently mirroring the images horizontally, vertically and horizontally & vertically. The motivation for adding margins is that real images are embedded in a constant background, acting as a margin. Six different margin sizes varying between 0 and 93 pixels were considered in the data set. The large margins are beyond typical sizes in applications and chosen to demonstrate the robustness of the approach.

For each image, the margin level was randomly set to a pixel value between 2600 and 10000 (to be compared with pixel values between 300 and 600 for the phantom area). A small random perturbation of 2% of the pixel intensity was added to each pixel of the margin. Fig. 4b shows an example of an augmented image with added margin after subsequent downsampling. To all

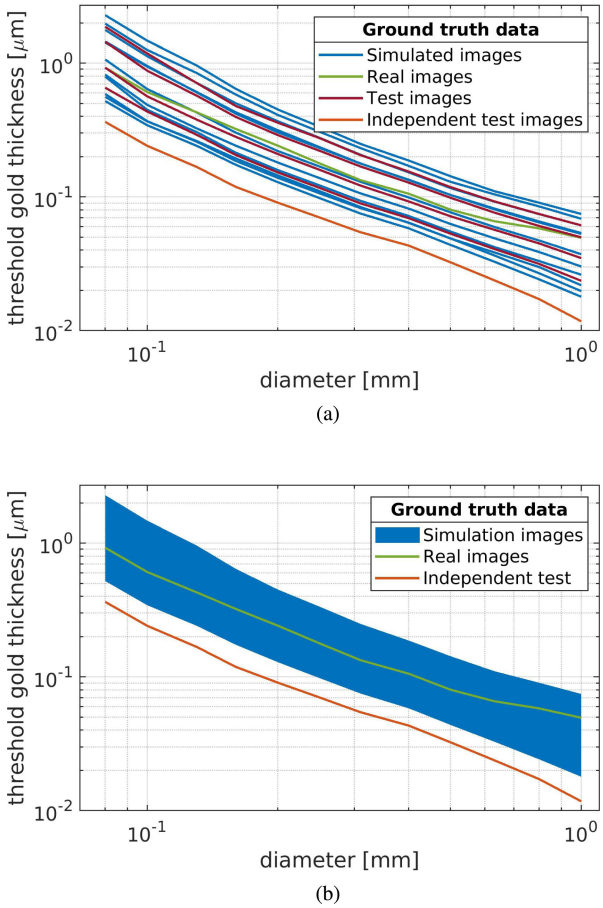


Fig. 3. Overview of the different ground truth contrast-detail curves. (a) Simulated images (blue) and real images (green) are used to train the neural net, cf. Table I. Independent simulated test data (red) are used for testing the trained net. Another independent test set (orange) is used to test the trained net's ability to generalize. Corresponding simulation parameters are listed in Table II. (b) Simplified visualization of the contrast-detail curves, where simulated training and testing data are summarized in a range (blue) while real images (green) and the independent test set (orange) are shown as individual curves.

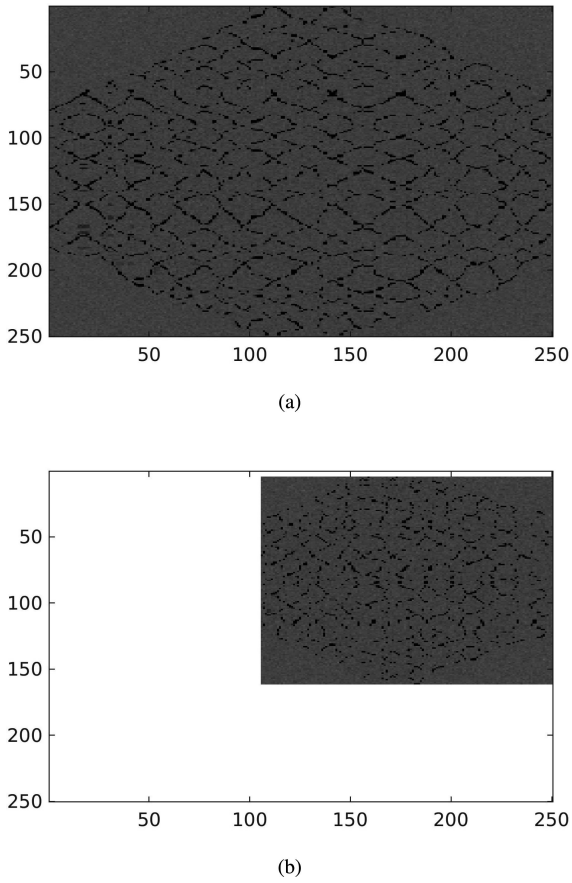


Fig. 4. Images of 250 × 250 pixels sampled from a simulated image of the CDMAM phantom with (b) and without (a) a margin of about 100 pixels (b).

augmented images the contrast-detail curve of the corresponding source image was assigned.

Image normalization is carried out by robustly estimating the pixel intensity of the phantom background and subtracting that value from the image. This is done automatically on a single image basis which ensures that the approach can be applied without any further knowledge about, e.g., a whole sample of images or physical parameters of devices, etc.

6) Training Data: The set of 48 real images was randomly split in two equally sized sets. One of them was used for training, and one for testing. All images belonging to the eleven simulation scenarios listed in Table II and shown as blue contrast-detail curves in Fig. 3a were taken for training, while the images belonging to the remaining simulation settings were used for testing, see also Fig. 3a. The training data was chosen to span a range of different quality levels with low and high exposure and SNR values as indicated in Fig. 3b. Note that the contrast-detail curve is shifted when a different combination of exposure and SNR is selected.

In order to balance the weight of the simulated and real images for training, the set of real images used for training was copied many times such that the number of its images was equal to the number of training images for each of the eleven different scenarios of simulated images, cf. Table I.

7) Test Data: The trained net was evaluated on test images by comparing the predicted contrast-detail curve for single images with the ground truth. As test data simulated images were used with parameter settings different from those used for simulated training images, i.e. simulated test data are independent from all training data (cf. Fig. 3a). Data for testing were simulated with either different SNR values or different exposure, or both. One additional test set was simulated with a significant lower SNR and medium exposure such that it does not fall into the range of the training data, see Fig. 3b. The remaining part of the real images not used for training was used for testing. Note that while each single real image has either been used for training or testing, training and testing is not independent regarding the real images in the sense that all of them were conducted by the same device.

C. Neural Net & Transfer Learning

A regression CNN was set up with 6.3×10^7 learnable parameters and architecture listed in Table III. The Matlab Deep

TABLE III
ARCHITECTURE OF THE CNN FOR CONTRAST-DETAIL CURVE ESTIMATION FROM CDMAM IMAGES

Layer	Type	Filter dimensions	Output dimensions
1	image input		$250 \times 250 \times 1$
2	convolutional	$3 \times 3 \times 1 \times 8$	$250 \times 250 \times 8$
3	batchnorm + relu		$250 \times 250 \times 8$
5	maxpool	2×2	$125 \times 125 \times 8$
6	convolutional	$3 \times 3 \times 8 \times 16$	$125 \times 125 \times 16$
7	batchnorm + relu		$125 \times 125 \times 16$
9	maxpool	2×2	$62 \times 62 \times 16$
10	convolutional	$3 \times 3 \times 16 \times 32$	$62 \times 62 \times 32$
11	batchnorm + relu		$62 \times 62 \times 32$
13	fully connected	123008×512	$1 \times 1 \times 512$
14	dropout		$1 \times 1 \times 512$
15	fully connected	512×512	$1 \times 1 \times 512$
16	fully connected	512×12	$1 \times 1 \times 12$
17	regression layer		1×12

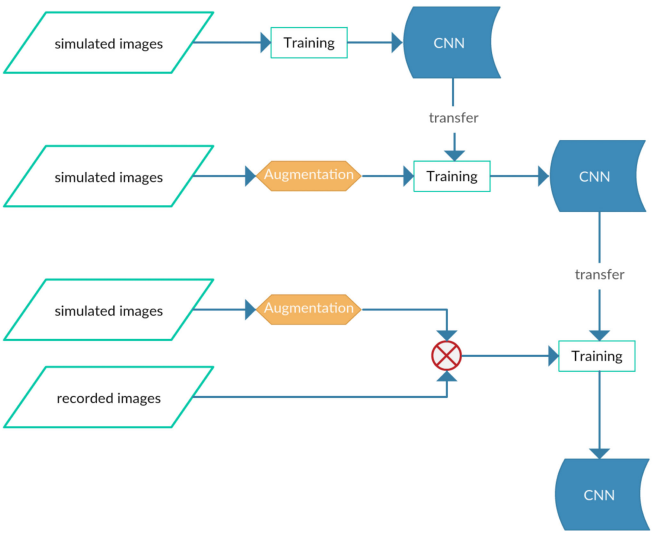


Fig. 5. Flowchart to illustrate the incremental learning strategy considering data extension. The training data was stepwise enlarged to include margins and further augmentations as well as real images in a later step.

Learning Toolbox [38] was used to implement the model. The mean squared loss was used as a cost function.

The CNN was trained in combination with an incremental learning strategy. In a first step the neural net was trained using only simulated training data that were not obtained through a data augmentation technique; all these simulated images are without margin. Furthermore, this step was initialized by first learning the net without the second fully connected layer, and finalized by transfer learning the second layer with dropout rate 0.2. In the next step, the training set was augmented with images containing small margins, while later, images with larger margins were included. In the final step then, the real images specified for training were also included into the training set. For each step of incremental learning, stochastic gradient descent with momentum was applied with an initial learning rate of $lr = 10^{-3}$, and a batch size of 128 images for 150 epochs. For every 30th epoch, the learning rate was dropped by 0.1. The incremental learning strategy is illustrated in Fig. 5.

D. Explainability

In order to determine which regions in the image are relevant for the trained net, we perturbed an input image and recorded the influence of the perturbation on the predicted contrast-detail curve. More precisely, for each of the points of the contrast-detail curve, the sensitivity of the trained net's prediction to a single pixel image perturbation was calculated. The combined sensitivity of the predicted contrast-detail curve was then determined by taking the root-mean-square sensitivity for the whole contrast-detail curve and all pixels, resulting in a gradient sensitivity image.

III. RESULTS

In this section, we provide a detailed comparison of the deep learning approach with the EUREF Guideline procedure based on simulated and real images of the CDMAM phantom. Recall that all simulated images used for testing were not used for training, and that the ground truth of the test images, i.e. their contrast-detail curves, are also different from the ground truths in the training data, cf. Fig. 3.

A. Prediction of Contrast-Detail Curves Using Test Data

Fig. 6a shows the predicted contrast-detail curves of the trained neural net for all simulated images in the test set. The ground truth of these test data differs from the ground truth of all training data, cf. Fig. 3, and is displayed as lines. Error bars indicate standard errors. The predictions, given as dots, follow the same intrinsic structure present in their ground truth, but also show some variability. The EUREF Guideline procedure requires at least 16 images of the CDMAM phantom. Fig. 6b shows in comparison averages of 16 single predictions of contrast-detail curves made by the trained neural net. Relative root mean squared errors averaged over all points and all contrast-detail curves in the test sets are 10% for the simulated images, and 5% for the real images.

Fig. 7a and Fig. 7b illustrate examples of predictions of contrast-detail curves made on the basis of single simulated images of the test data that all have the same ground truth. The Figures show the predicted threshold gold thickness values versus the corresponding gold thickness values obtained by the EUREF Guideline contrast-detail curve. The prediction of the CNN varies for all individual images and is distributed around the ground truth, which is indicated by the black line. Cases where the simulated test images contained a margin are distinguished. The results show that predictions of a similar quality are obtained regardless of whether the images contain an additional margin or not. It can be concluded that the net has successfully learned to cope with the situation of images having different portions of margin.

Fig. 8 shows analogous results for the real images in the test data, together with their common ground truth. A fairly good agreement can be observed. Recall that while the real images in the test set did not belong to the training data, the training data included (different) real images belonging to the same ground truth.

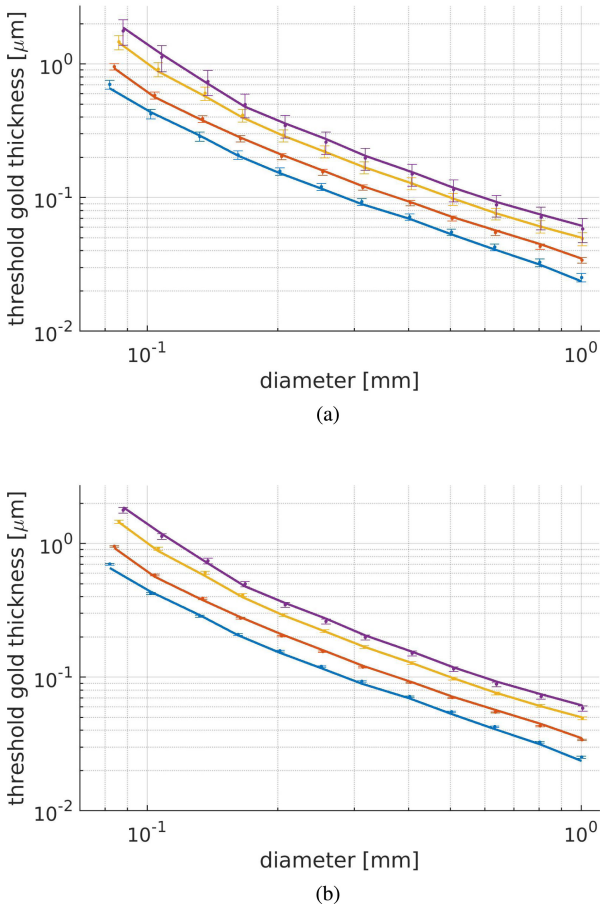


Fig. 6. Ground truth (line) and predictions (dots) of the trained neural net for the simulated test images (a) on the basis of single images and (b) averages of 16 single predictions from (a). Error bars display standard errors.

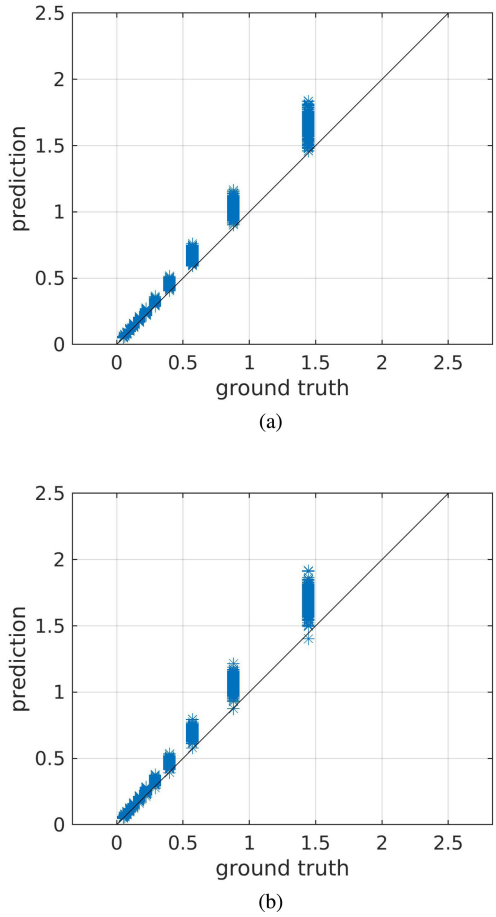


Fig. 7. Predicted threshold gold thickness (prediction) vs. the threshold gold thickness values obtained from the EUREF Guideline contrast-detail curve (ground truth) for one set of simulated images of the test data for (a): no margin and (b): with additional margin.

Ideally, for a fixed diameter the value of the contrast-detail curves for the different scenarios (i.e. the different image qualities) should be separated from each other. Fig. 9a illustrates the variability for a single diameter of the predictions made on the simulated test data, and Fig. 9b shows the variability for averages of 16 single predictions. While neighboring quality levels cannot be reliably distinguished from predictions made by the trained net on the basis of a single image, the estimation accuracy is large enough to distinguish between larger differences in the quality of a mammography device. For averages of 16 predicted contrast-detail curves, even neighboring quality levels can reliably be distinguished. We refer to Section III-D for a detailed assessment of the variability of the predicted contrast-detail curves in comparison with the uncertainty of the ground truth.

The coefficient of determination R^2 has been calculated for all cases in the test set in order to further assess the quality of the predicted contrast-detail curves. Fig. 10a and Fig. 10b display the distributions of these coefficients, the vast majority of which are near 1. For the test cases based on simulated data, the median R^2 of the 4800 cases equals 0.9859, while for the 24 test cases based on the real data, a value of 0.9955 was obtained.

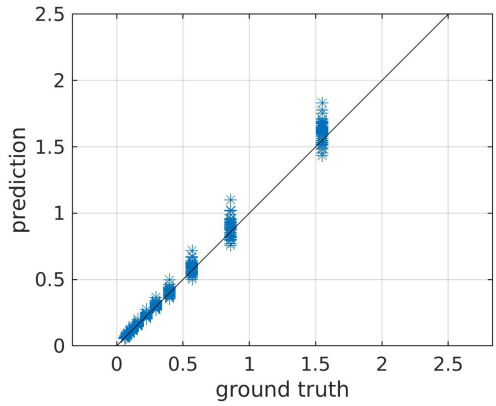


Fig. 8. Predicted threshold gold thickness (prediction) vs. the threshold gold thickness values obtained from the EUREF Guideline contrast-detail curve (ground truth) for the set of real images in the test data.

B. Generalization

In order to assess the trained net's ability to generalize, an additional scenario was tested for which the contrast-detail curve differs significantly from all cases in the training set, cf. the case on the right in Fig. 2 and the corresponding lower contrast-detail

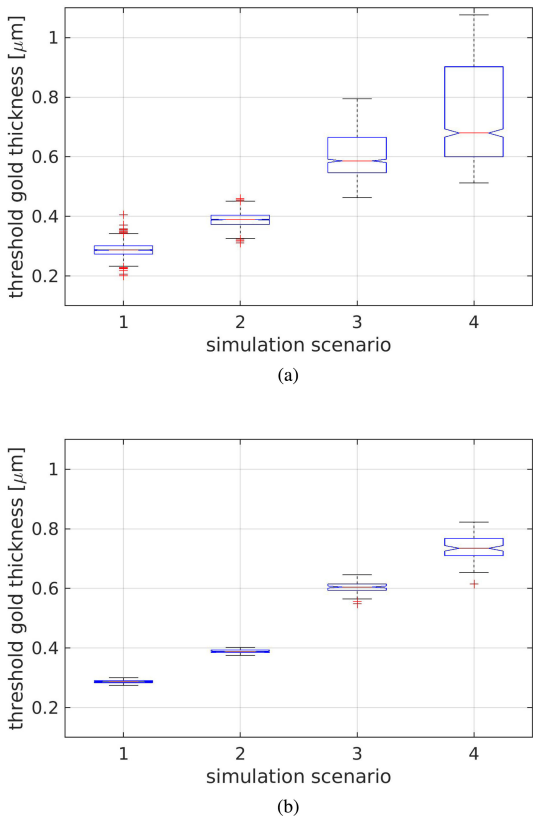


Fig. 9. Variation of predictions on the simulated test set for diameter 0.13 mm. (a) Single prediction of the contrast-detail curve and (b) averages of 16 single predictions of the trained net. The four different simulation scenarios represent the four different image qualities of the simulated test images, cf. Fig. 6.

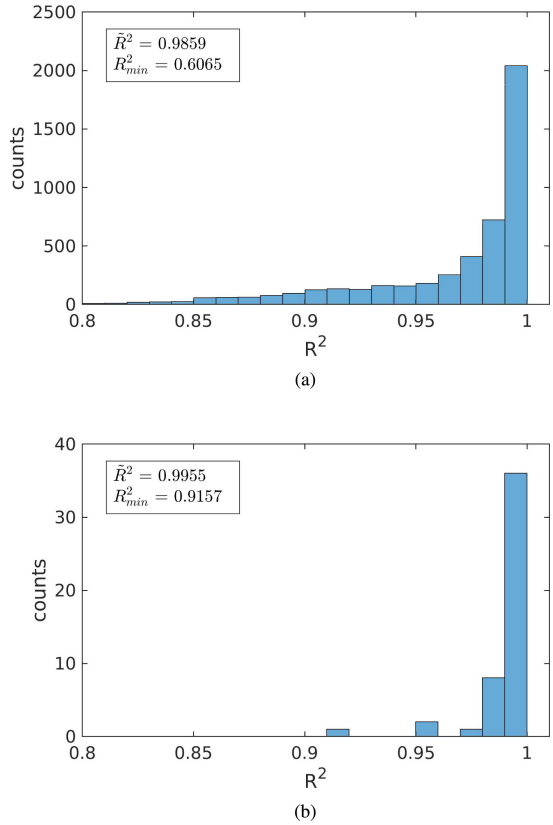


Fig. 10. Coefficients of determination R^2 for predictions of the trained net on the test set consisting of (a): 4800 simulated images and (b): 24 real images.

curve in Fig. 3. Fig. 11 shows the prediction of the trained net obtained from single images for this additional test case. The predictions are fairly accurate, suggesting that the trained net is able to generalize well.

C. Explainability

The trained net has learned to successfully predict contrast-detail curves from images with and without margins. In fact, the gradient sensitivity images (not shown) demonstrate that pixels belonging to the margins are irrelevant for the result of the trained net. The gradient sensitivity images also appear to indicate that those pixels are important which contain cells with pairs of diameters and thickness that are part of the contrast-detail curve.

D. Uncertainty

The variability of predictions made by the trained net was compared with that of the EUREF Guideline procedure for the 48 real images which are all of the same quality. By repeatedly drawing random subsets of 16 images from this set and by applying the EUREF Guideline procedure, the variability of this procedure (when using 16 images) was determined. Fig. 12 shows the corresponding results, which are compared with the results of the trained net based on single images. The results

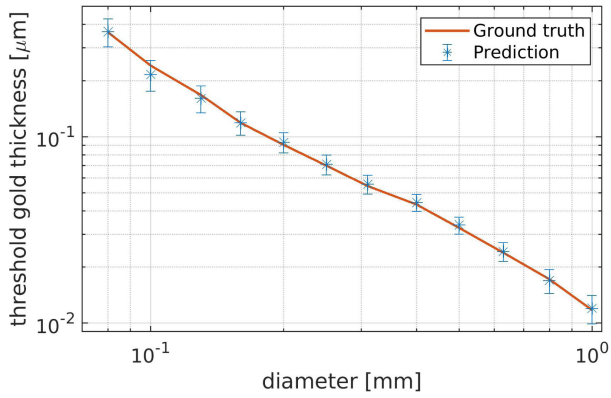


Fig. 11. Ground truth (red) and predictions by the trained net based on single images (blue). Error bars display standard errors.

demonstrate that the variability of predictions of the trained net for single images is similar to that of the EUREF Guideline procedure using 16 images. For the smallest diameter, the variability of the neural net prediction is even less than that of the EUREF Guideline procedure. The mean contrast-detail curves of both approaches agree well.

Recall that the EUREF Guideline procedure was used to define the ground truth by applying it to a large number of images, namely 200 images for the virtual mammograms and 48 real mammograms. However, since we observe a significant

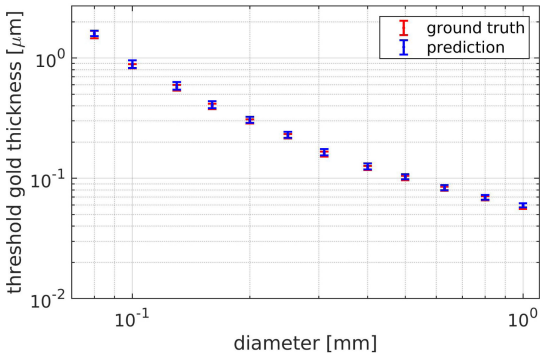
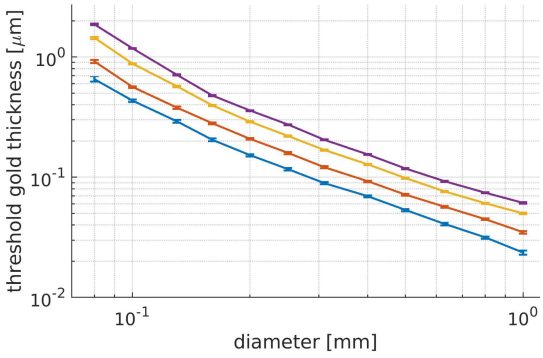
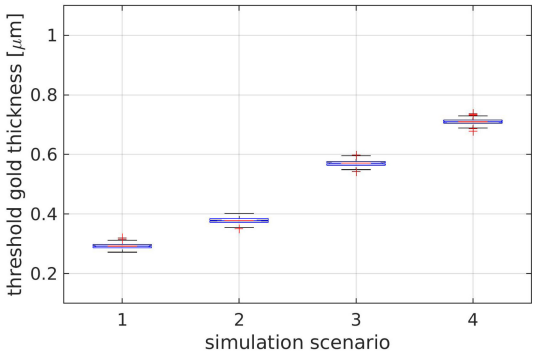


Fig. 12. Contrast-detail curves as obtained by repeated applications of the EUREF Guideline procedure for 16 randomly drawn images from the set of 48 real images (red) with common ground truth, and single predictions of the trained net of the 24 real test images. Error bars indicate standard errors.



(a)



(b)

Fig. 13. (a) Uncertainty of the four test ground truth curves, which has been determined on the basis of all 200 simulated images of each scenario. Error bars indicate standard errors. (b) Uncertainty of the different ground truth curves at a diameter of 0.13 mm, cf. Fig. 9.

variability for the results of the EUREF Guideline when using 16 images, the ground truth based on 48 (real images) or 200 (simulated images) will also contain an uncertainty. An estimate of that uncertainty can be obtained by scaling the standard deviation $\sigma(\rho)$ at each diameter ρ of the contrast-detail curve obtained by the EUREF Guideline procedure when repeatedly applying it to randomly drawn subsets of size 16 images; i.e., σ was multiplied by $(16/48)^{1/2}$ for the real images and by $(16/200)^{1/2}$ for the simulated images. Fig. 13a shows the resulting expected variation of the ground truth. By comparing this variability with the variability observed for the trained net (cf. Fig. 6), the variability of the latter is somewhat larger, yet Fig. 13a suggests that relevant uncertainty is still present in the ground truth. Note that Fig. 13a and b visualize the variability of the ground truth as determined by all 200 simulated images of the corresponding simulation scenario. Hence, they are smaller than the results displayed in Fig. 9a and b, which show the variation of predictions from single and 16 images, respectively. Fig. 12 demonstrates that our approach yields the same variability for the single image predictions as the EUREF Guideline procedure for 16 images.

Fig. 13a also reveals that the uncertainty in the ground truth increases with increasing threshold gold thickness. Fig. 7a and b show that an increasing spread between predictions of the neural net and ground truth also with increasing gold thickness. This increasing spread may be due to increased prediction errors, or increased ground truth errors, or both.

E. Downsampling

The high downsampling rate destroys the regular structure of the CDMAM phantom in the original image to some extent and a human observer would probably no longer be able to assess image quality. Our results demonstrate that a neural net can still assess image quality reliably from these highly downsampled images.

In order to further explore the impact of downsampling, we considered another case of simulated test images with a ground truth not used for training the net. In that case, however, the CDMAM phantom was altered by randomly changing the location of the second disc in each cell. Application of the trained net to a set of such (downsampled) simulated images led to predictions similar as those obtained by the original phantom. In another experiment we applied the trained net to simulated images in which the pixels were randomly permuted. In that case the trained net failed to make reasonable predictions.

IV. Discussion

Our results demonstrate that a trained CNN can be used to successfully analyze images of the CDMAM phantom in mammography image quality assurance. The accuracy of estimates of image quality made by the CNN from single images is similar to, or even better than that obtained by the current EUREF Guideline procedure using 16 images. The trained net was successfully applied to real images and generalized well to a scenario far from all scenarios used for training the net. These findings underline the potential reliability of deep learning for mammography quality assurance.

The CNN was trained on randomly downsampled images and was able to accurately predict the contrast-detail curves from images that were randomly downsampled. The downsampled images are probably not suitable for correct image quality assessment by a human observer, as many of the small structures got lost. While the good performance of the neural net on downsampled images is of some interest on its own, it can also

be relevant for the attempt to design simpler phantoms in the future. Image quality assessment refers to assessing a device's ability to resolve small structures rather than actually detecting those structures. Although the regular structure of the CDMAM image may no longer be visible in the downsampled images, the relevant information of image quality can still be retrieved by a trained neural net. Predictions of the trained net to simulated images obtained by randomly position the second gold discs in the cells were of similar quality, while randomly permuted pixels caused a failure of the prediction of contrast-detail curves by the neural net. This indicates that although downsampling leads to a loss of information, the structure is still relevant in the images and the trained net uses some structural information for the estimation of contrast-detail curves. Note that the trained net performs a regression rather than a detection task. It appears that the successful regression is achieved through a separation into detectable and non-detectable discs. In fact, the gradient sensitivity images suggest that those parts of the image which belong to cells having pairs of diameter and thickness belonging to the contrast-detail curve are particularly relevant. Training of the net would then amount to a detection of this region in an image, a task significantly simpler than that of detecting single cells. This interpretation would explain why the highly downsampled images are still sufficient for reliably estimating image quality assessment. Future research may further this conjecture, e.g. through randomly permuting the cells of the phantom or other means of distorting its underlying structure.

Another way of interpreting the surprisingly good performance of the trained net on highly downsampled images is that a regression task is considered, and that the original image contains highly redundant information for that. This aspect may be interesting also in other applications where machine learning is applied for regression on images, and where sparse sampling procedures are desired.

The EUREF Guideline procedure requires that images be segmented and phantoms aligned prior to applying its signal detection procedure. These pre-processing steps are not merely cumbersome but also introduce sources of uncertainty for the final result. The CNN, on the other hand, was trained on the basis of the (downsampled) raw data, and the net successfully learned to "pre-process," i.e., it automatically accounts for misalignment or the presence of margins.

Incremental learning can be applied by refining the net architecture [39], [40] as well as by changing the data [41]. The aim of incremental learning is to balance the already learned knowledge with a correct prediction of completely new data. While restarting the learning phase with randomly chosen net parameters will completely delete the net's knowledge, which is referred to as "catastrophic forgetting" in [42], [43], transfer learning utilizes the parameters learned thus far. One of our findings while training the CNN was that transfer learning was very helpful, specifically by a step-wise inclusion of simulated images containing margins of growing size. Training the net with all data from the scratch led to significantly worse results. This work provides primarily a proof of concept for the application of deep learning for mammography image quality assessment. Future research could address remaining limitations in the presented approach. For example, only few real images were used for training and testing. Furthermore, all real images belonged to the same ground truth, i.e. are images of the CDMAM phantom conducted by the same device. Since some of them were used for training, testing the trained net on the remaining real images did not allow for assessing the generalization on real images. Nevertheless, the encouraging results demonstrate that neural nets can be trained to accurately predict contrast-detail curves. To implement this in a quality assurance protocol, more real images should be conducted covering a large range of acquisition parameters and techniques, which is beyond the scope of this work.

A human based quality assessment is mainly focused on the detectability of small structures in an image of the CDMAM phantom. However, the applied downsampling appears to destroy much of this information. Future research may explore the impact of downsampling in more detail, e.g. by comparing results for different rates of downsampling or even attempting to learn the net on the original images.

Another interesting issue lies in the fact that the downsampled images are sufficient for a neural net to predict image quality, and that the trained net fails, for example, when applied to (downsampled) images with randomly rearranged images. It seems that random downsampling in some sense maintains the regular structure of the CDMAM phantom. Clarification of this issue may also be helpful in the design of cheaper technical phantoms.

To the best of our knowledge, deep learning has not yet been applied for image quality estimation in mammography quality assurance so far. For these reasons, we compared the proposed approach only with the current standard method provided by the EUREF Guideline procedure. However, alternative approaches from deep learning, and in particular other net architectures or transfer learning procedures, could be explored.

V. Conclusion

In mammography quality assurance, a contrast-detail curve is determined that quantifies a system's ability to visualize small structures. Contrast-detail curves are derived from multiple images of the CDMAM phantom. Following the European Reference Organization for Quality Assured Breast Cancer Screening and Diagnostic Services (EUREF), images of the CDMAM phantom are analyzed by an automated procedure that combines image registration, signal detection and nonlinear fitting.

We introduced a convolutional neural net to predict a contrast-detail curve from a single image of the CDMAM phantom. Virtual mammography was used to build a large image data set enriched with real images of the CDMAM phantom.

The trained CNN successfully predicted contrast-detail curves from single simulated and single real images of the CDMAM phantom. Furthermore, the trained CNN successfully learned to ignore the margins in the images, thus allowing images to be analyzed without cumbersome pre-processing.

The results show that the trained net can estimate contrast-detail curves from single images at least as precisely as when

applying the current EUREF Guideline procedure with 16 images.

These findings demonstrate the potential advantages of deep learning for mammography quality assurance. We conclude that mammography quality assurance can benefit from current techniques in deep learning.

Acknowledgment

The authors would like to thank S. Schopphoven from the German Reference Centre South-west for providing various images of the CDMAM phantom. They would also like to thank the reviewers and the Associate Editor for helpful comments and suggestions.

References

[1] L. Nyström *et al.*, "Long-term effects of mammography screening: Updated overview of the Swedish randomised trials," *Lancet*, vol. 359, no. 9310, pp. 909–919, 2002.
[2] A. L. Siu, "Screening for breast cancer: US preventive services task force recommendation statement," *Ann. Internal Medicine*, vol. 164, no. 4, pp. 279–296, 2016.
[3] ICRU, "ICRU Report No. 82: Mammography-assessment of image quality," *J. ICRU*, vol. 9, no. 3, pp. 1–104, 2009.
[4] D. J. Brenner *et al.*, "Cancer risks attributable to low doses of ionizing radiation: Assessing what we really know," *Proc. Nat. Acad. Sci.*, vol. 100, no. 24, pp. 13 761–13 766, 2003.
[5] I. Amendoeira *et al.*, *European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis*, 4th ed., Luxembourg: European Commission, 2006.
[6] N. Karssemeijer and M. A. O. Thijssen, "Determination of contrast-detail curves of mammography systems by automated image analysis," *Digit. Mammography*, vol. 96, pp. 155–160, 1996.
[7] N. Perry *et al.*, *European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis*. European Commission, 2013.
[8] L. M. Warren *et al.*, "The effect of image processing on the detection of cancers in digital mammography," *Amer. J. Roentgeno.*, vol. 203, no. 2, pp. 387–393, 2014.
[9] A. Mackenzie *et al.*, "The relationship between cancer detection in mammography and image quality measurements," *Phys. Medica*, vol. 32, no. 4, pp. 568–574, 2016.
[10] D. Mishra *et al.*, "Ultrasound image segmentation: A deeply supervised network with attention to boundaries," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1637–1648, Jun. 2019.
[11] H. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[12] P. Croce *et al.*, "Deep convolutional neural networks for feature-less automatic classification of independent components in multi-channel electrophysiological brain recordings," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 8, pp. 2372–2380, Aug. 2019.
[13] K. He *et al.*, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1026–1034.
[14] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
[15] A.-A. Nahid and Y. Kong, "Involvement of machine learning for breast cancer image classification: A survey," *Comput. Math. Methods Medicine*, vol. 2017, 2017.
[16] F. Klauschen *et al.*, "Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning," in *Seminars in Cancer Biology*, vol. 52. New York, NY, USA: Elsevier, 2018, pp. 151–157.
[17] N. I. Yassin *et al.*, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Computer Methods Programs Biomedicine*, vol. 156, pp. 25–45, 2018.
[18] T. Kooi *et al.*, "Large scale deep learning for computer aided detection of mammographic lesions," *Med. Image Anal.*, vol. 35, pp. 303–312, 2017.
[19] Y. LeCun *et al.*, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Dec. 1998.
[20] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1–9.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Inf. Process. Syst.*, vol. 25. F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. San Diego, CA: Curran Associates, Inc., 2012, pp. 1097–1105.
[22] D. Baehrens *et al.*, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, 2010.
[23] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.
[24] S. Bach *et al.*, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, 2015, Art. no. e0130140.
[25] I. Sturm *et al.*, "Interpretable deep neural networks for single-trial EEG classification," *J. Neurosci. Methods*, vol. 274, pp. 141–145, 2016.
[26] S. Lapuschkin *et al.*, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2912–20.
[27] S. Lapuschkin *et al.*, "Unmasking clever Hans predictors and assessing what machines really learn," *Nature Commun.*, vol. 10, pp. 1096–1103, 2019.
[28] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, 2018.
[29] T. Kretz *et al.*, "Determination of contrast-detail curves in mammography image quality assessment by a parametric model observer," *Physica Medica*, vol. 62, pp. 120–128, 2019.
[30] L. Kang *et al.*, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1733–1740.
[31] S. Bosse *et al.*, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.
[32] Y. Xie *et al.*, "Beyond classification: Structured regression for robust cell detection using convolutional neural network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 358–365.
[33] G. S. Babu, P. Zhao, and X.-L. Li, "Deep convolutional neural network based regression approach for estimation of remaining useful life," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2016, pp. 214–228.
[34] M. Liu *et al.*, "Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1195–1206, May 2019.
[35] J. A. Thomas *et al.*, "Contrast-detail phantom scoring methodology," *Med. Phys.*, vol. 32, no. 3, pp. 807–814, 2005.
[36] I. T. Jolliffe, "Principal components in regression analysis," in *Proc. Principal Compon. Anal.*, 1986, pp. 129–155.
[37] G. R. Easley, D. Labate, and F. Colonna, "Shearlet-based total variation diffusion for denoising," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 260–268, Feb. 2009.
[38] MATLAB, *Deep Learning Toolbox version 12.1 (R2019a)*. Natick, Massachusetts: The MathWorks, 2019.
[39] E. H. Wang and A. Kuh, "A smart algorithm for incremental learning," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 1992, vol. 3, pp. 121–126.
[40] B.-T. Zhang, "An incremental learning algorithm that optimizes network size and sample size in one trial," in *Proc. IEEE Int. Conf. Neural Netw.*, 1994, vol. 1, pp. 215–220.
[41] A. Engelbrecht and R. Brits, "A clustering approach to incremental learning for feedforward neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2001, vol. 3, pp. 2019–2024.
[42] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends Cognitive Sci.*, vol. 3, no. 4, pp. 128–135, 1999.
[43] J. Kirkpatrick *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Academy Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.