Machine Learning: Science and Technology

PAPER • OPEN ACCESS

Explainability for deep learning in mammography image quality assessment

To cite this article: N Amanova *et al* 2022 *Mach. Learn.: Sci. Technol.* **3** 025015

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
3 February 2022

REVISED
14 June 2022

ACCEPTED FOR PUBLICATION
17 June 2022

PUBLISHED
4 July 2022

PAPER

Explainability for deep learning in mammography image quality assessment

N Amanova*, J Martin and C Elster

Physikalisch-Technische Bundesanstalt, Abbestraße 2-12, Berlin, 10587, Germany

* Author to whom any correspondence should be addressed.

E-mail: narbota.amanova@ptb.de

Keywords: deep learning, explainability, mammography image quality assessment

Abstract

The application of deep learning has recently been proposed for the assessment of image quality in mammography. It was demonstrated in a proof-of-principle study that the proposed approach can be more efficient than currently applied automated conventional methods. However, in contrast to conventional methods, the deep learning approach has a black-box nature and, before it can be recommended for the routine use, it must be understood more thoroughly. For this purpose, we propose and apply a new explainability method: the oriented, modified integrated gradients (OMIG) method. The design of this method is inspired by the integrated gradientsmethod but adapted considerably to the use case at hand. To further enhance this method, an upsampling technique is developed that produces high-resolution explainability maps for the downsampled data used by the deep learning approach. Comparison with established explainability methods demonstrates that the proposed approach yields substantially more expressive and informative results for our specific use case. Application of the proposed explainability approach generally confirms the validity of the considered deep learning-based mammography image quality assessment (IQA) method. Specifically, it is demonstrated that the predicted image quality is based on a meaningful mapping that makes successful use of certain geometric structures of the images. In addition, the novel explainability method helps us to identify the parts of the employed phantom that have the largest impact on the predicted image quality, and to shed some light on cases in which the trained neural networks fail to work as expected. While tailored to assess a specific approach from deep learning for mammography IQA, the proposed explainability method could also become relevant in other, similar deep learning applications based on high-dimensional images.

1. Introduction

The evaluation of image quality, otherwise known as *image quality assessment* (*IQA*) [1], plays an important role in various image-processing applications. Especially in safety-relevant use cases such as medical applications, a reliable and robust assessment of image quality is of particular relevance. In mammography, the main goal of IQA is to ensure that the image shows the breast anatomy and lesions compatible with breast cancer. This main goal must be achieved with the lowest possible radiation dose. A variety of phantoms for IQA in mammography have been proposed [2]. The method we take as a reference here was recommended by European Guidelines [3, 4] and is based on the automated readout of the contrast-detail curves (CDCs) from a set of recorded images of a CDMAM[1] phantom [3]. A method recently proposed for the CDC determination involved using a deep neural network [5, 6]. This method, which requires only a single image and allows the calculation of uncertainties, is considered in this work. The approach is one of an increasing number of deep learning-based methods in the field of medical IQA [7–11]. Unfortunately, neural networks suffer from a lack of explainability due to their black-box nature [12], which is of particular relevance in

[1] Contrast-Detail Phantom for Mammography.

safety-critical areas such as medicine. The goal of this paper is to provide an explainability for the deep learning approach from [5, 6] for IQA.

While a variety of methods exists for the explainability of neural networks [13, 14], we found none of them to yield satisfactory results for the deep learning-based IQA in mammography. We therefore propose a new method specifically tailored for the use case, which we call *oriented, modified integrated gradients* (OMIG) method. The resulting explainability map supports users of a deep learning approach in mammography IQA in various ways:

- It makes it possible to understand which parts of the CDMAM phantom have a significant impact on the prediction.
- This, in turn, allows the trustworthiness of the employed neural network to be assessed.
- Ascertaining which parts of the phantom are the most informative for a deep learning-based IQA in mammography might inspire the design of new phantoms for this application.

Although the proposed explainability method is tailored to the particular task of IQA in mammography, we believe that its ideas could be beneficial also in other applications of deep learning. However, the exploring of the potential of the proposed approach for other cases is beyond the scope of this article. Nevertheless, we will still provide a short discussion on this topic.

The paper is structured as follows: section 2 briefly recapitulates a few aspects of IQA that are important for mammography, introduces the employed approach based on deep learning, and finally presents our explainability approach. This includes a technique for determining the important input features from the downsampled data, which will be referred to as 'upsampling' within this work. Results are then discussed in section 3. Finally, we discuss limitations, give an outlook and some conclusions.

2. Methods

2.1. A recap of IQA for mammography

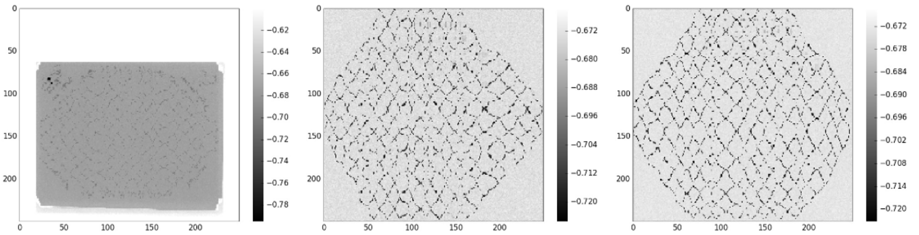
As indicated in the introduction, we here follow the European Guidelines that recommend the use of the CDMAM phantom for IQA in mammography. This phantom, which is shown in figure 8 in the appendix, is a polymethyl methacrylate-coated aluminum plate with gold disks of various diameters and thicknesses placed on it. The disks, that represent breast lesions such as microcalcifications [15], are inserted in a matrix (grid) of square cells. The CDMAM phantom includes disks with 16 diameters whose detection is highly depending on the mammography system together with the radiographic factors. A CDC is determined from images of the CDMAM phantom and describes the minimum threshold gold thickness for various diameters necessary for the corresponding disks to be detected.

We will here use the method of automated readout, as first introduced in [16] and recommended in [3], which provides the results for the 12 disks with diameters between 1 and 0.08 mm [17], since the smallest diameters (smaller than 0.1 mm) are harder to detect. The 12 values of the CDC are then rescaled via an experimentally found relation [3, 18] in order to be comparable to human readout. From the point of view of IQA each disk diameter is relevant although the values of the limiting curve are defined only for five diameters. Four of these five diameters are contained in the CDC range obtained by the automated readout.

Given images of the CDMAM phantom the CDC can be computed by the *CDMAM Analyser* software [19, 20] which requires at least 16 recordings of the phantom. Following the approach in [5, 6], we will here train a deep neural network (to be more specific, an ensemble of networks, cf section 2.2 below) on images labeled with the ground truth CDCs to predict CDCs using a single image. The CDCs determined by the *CDMAM Analyser* software serve as the ground truth. Image quality is considered better if the CDC has smaller values, because a smaller CDC characterizes a smaller minimum threshold gold thickness. We will make extensive use of this fact below.

2.2. Details on training and used data

We trained an ensemble of $M = 10$ neural networks (a so-called deep ensemble [21]) that had the same architecture as in [5], using 52 800 simulated images and 4800 real images (constructed from 48 recordings via augmentation [5, 6]). The neural networks had 12 output neurons that correspond to the 12 points of the CDC. For training we used the negative log-likelihood loss [21] assuming homoscedastic noise, whose variance is learned during the training. We used an Adam optimizer [22] with a learn rate of 10^{-5}. The training lasted 300 epochs with a batch size of 50 images. For regularization purposes, we used an $L2$ regularization of 10^{-4} and a dropout layer with a dropout rate of 0.2. Details of the network architecture can be found in figure 10 and table 2 in appendix C together with an illustration of the training procedure that uses incremental learning in figure 11. Using an ensemble instead of a single network allows the predictions



(a) Recorded (b) Scenario #1 (c) OOD

Figure 1. Example of the downsampled images from the test data set: a recorded image (a) and two simulated images from scenario #1 (b) and the OOD set (c), cf section 2.2 for details. Note that, for the two simulated images (b) and (c) we chose two simulated images without margin for the purpose of presentation. For simulated examples with a margin, as in the recorded image (a), see figure 9 in the appendix.

Table 1. The simulation parameters used by the software for the virtual mammography [29]. The software considers the tube voltage, current-exposure time product and the noise level which is defined as the standard deviation of the Gaussian white noise that is fed into a Gaussian filter which simulates the noise in mammography images.

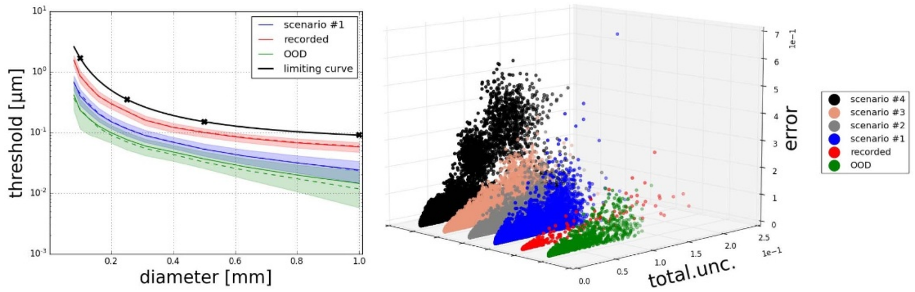
Scenario	Voltage (kVp)	Exposure (mAs)	Noise level
#1	28	115	9.55
#2	25	110	13.06
#3	25	110	18.91
#4	25	110	22.91
OOD	23	105	6.04

to be improved and makes them more robust [23–26]. Furthermore, uncertainties can be evaluated in this way, which is of importance for making deep learning usable in a medical context [27, 28]. We follow [5, 6] in downsampling the images in the training data to a resolution of 250 × 250, which was observed in those works to reduce the required computational resources and to improve the convergence of the neural network.

To evaluate the trained neural networks, we used data sets which consisted of 24 recorded images and 5000 simulated images [5, 6]. These test images were not part of the training data. While the recorded images were originated from the same device with the same anode/filter combination, tube voltage and current-exposure time product, the simulated images split into the groups whose simulation parameters are described in table 1. The training and the test data sets were normalized by the mean and standard deviation obtained from the pixel intensities of the whole images of the training data set. The four groups, of 1200 test images each, were simulated using the simulation parameters from table 1 and six different margin sizes. We refer to these groups as 'scenario #1–#4'. Additionally, a single set of 200 out-of-distribution test images was simulated with a choice of simulation parameters that fundamentally differ from the ones used for training and for scenario #1–#4. We will refer to this group as 'OOD'.

For further details on the data, see [5, 6]. To evaluate the trained neural network, the test images must be downsampled to a resolution of 250 × 250. Some downsampled test images are shown in figure 1, including a real (i.e. recorded) test image. Figure 1(b) shows a simulated image from scenario #1. Note that this particular example has no margin, in contrast to the recorded example in (a). The training and test sets contain simulated images with and without margins of various sizes, cf [5, 6] and figure 9 in the appendix for an example including a margin. Figure 1(c) shows an image from the OOD data set (again an image without a margin).

Figure 2 shows some results of the impact of our trained neural networks on the test data set. For the specific images from figure 1 the predicted CDCs together with their uncertainty multiplied by a factor of 1.96 are shown in figure 2(a). The plotted uncertainty shows the total uncertainty, whose square is the sum of the squared epistemic and aleatoric uncertainties, see [21, 30] for details. The factor 1.96 is chosen so that the corresponding interval estimate has a coverage probability of 95% assuming a Gaussian distribution. The corresponding ground truths for the three images, that were obtained by the *CDMAM Analyser* software, are depicted by the dashed lines. The black solid line connects the four values of the limiting curve described in section 2.1. For all data points in the test data set, the relationship between the prediction error and the predicted uncertainty is shown in figure 2(b). By far, the highest errors were recorded for scenario #4, which is set apart from the other groups in the test set by its high noise level, cf table 1. In addition, the expanded total uncertainty (total uncertainty times 1.96) covered the ground truth only in 61% of the cases. The



(a) predicted CDCs (b) error vs. uncertainty

Figure 2. Performance of the trained deep ensemble. Figure (a) shows the predicted CDCs (solid lines), together with their expanded total uncertainty, for the three images from figure 1. The black solid line represents the limiting curve. Figure (b) shows the error of the ensemble prediction vs. its total uncertainty for *all* data points in the test data set. The points were grouped and color-coded along a third dimension, according to the scenarios listed in section 2.2.

explainability method presented in this work allowed us to investigate this point further and to study *why* the predictions are worse for scenario #4. For all other groups in the test data set, (i.e. the recorded images, scenario #1–#3 and the OOD set), the error was substantially smaller and the observed coverage rates of the interval estimates were rather conservative—namely, 100% (recorded), 97.1% (sc. #1), 99.1% (sc. #2), 96.9% (sc. #3), and 98.7% (OOD).

2.3. Explainability

Explainability for neural networks is often done by analyzing the effect of the input features, which in our case are the image pixels, on the output [31] (i.e. the 12 points of the CDC). Such an explanation is often presented by a heatmap that highlights which features of the input image affect the output of the network to the greatest extent. While there are several generally applicable methods [13, 14], none of the established methods we explored yielded expressive results for the mammography example studied in this work. Moreover, the computations of some methods are too complex and time-consuming to be combined with the upsampling procedure, which we explain below and which is required in order to obtain informative, high-resolution results. We therefore propose a new method, OMIG, that has been tailored to the needs of the CDMAM evaluation.

Before presenting a detailed analysis, let us summarize several factors unique to deep learning-based IQA in mammography:

(a) Recall that a CDC expresses a higher image quality if its threshold values are lower. This gives the problem at hand a *natural direction* that we can exploit.
(b) We want to use a *deep ensemble* as this allows the uncertainty to be quantified, which is of great relevance in a medical case such as mammography.
(c) While the network was trained with *downsampled inputs* to ensure scalability and convergence, what is ultimately required is a heatmap that corresponds to the original high-resolution image, as this will allow finer patterns to become visible.

2.3.1. OMIG explainability

The motivation behind the explainability method introduced in this work is the following question:

How can we modify a given input image x to a new image x' in such a way that for x' the neural network is expected to predict a better image quality?

Here, the intention is that a quantitative answer to this question will give us an information about the importance of the features in x on the prediction. The goal is to construct a modified hypothetical image x', somehow close to the original image x. We then expect that the image x' yields a CDC that lies lower than the CDC obtained from the original image x.

Let us introduce some notations. We denote by $\mu_m(x)$ with $m = 0, \ldots, M$ the output of the M ensemble members (in our case $M = 10$) for the input image x. Recall, that we predict $K = 12$ points of the CDC, so that each $\mu_m(x)$ is in fact K-dimensional. Let us denote the (again K-dimensional) prediction of the deep ensemble as $\tilde{\mu}(x) = \frac{1}{M}\sum_{m=1}^{M}\mu_m(x)$ and the K components of $\tilde{\mu}(x)$ by $\tilde{\mu}^k(x)$. We now summarize the K predicted points of $\tilde{\mu}(x)$ into a scalar, by taking the 'midpoint' of the CDC, that is $\frac{1}{K}\sum_{k=1}^{K}\tilde{\mu}^k(x)$. We will assume this midpoint as a scalar measure for the position of the CDC. More precisely, we will say that the CDC is lowered when this is true only for the midpoint. It turns out that this, somehow simplified, interpretation is completely sufficient for our purposes. This reduces the question above to the task of obtaining from x an x' such that $\frac{1}{K}\sum_{k=1}^{K}\tilde{\mu}^k(x') < \frac{1}{K}\sum_{k=1}^{K}\tilde{\mu}^k(x)$. To construct x' we proceed as follows: given x, compute $\tilde{\mu}(x)$ and compute the gradient of the midpoint

$$R(x) = \frac{\partial\left(K^{-1}\sum_{k=1}^{K}\tilde{\mu}^k(x)\right)}{\partial x}. \tag{1}$$

Then, use $R(x)$ to successively update x to create a modification x' by moving x in the direction of the negative gradient. More precisely, set $x_0 = x$ and, for some positive $c > 0$ and integer L set $\varepsilon = \frac{c}{L}$, compute

$$x_{l+1} = x_l - \varepsilon \cdot R(x_l) \tag{2}$$

for $l = 0,\ldots,L-1$. Finally, set

$$x' = x_L.$$

The motivation behind choosing $\varepsilon \propto \frac{1}{L}$ is that for large L the discrete steps from (2) form a continuous path whose length depends on c and whose endpoint is given by x'. We then define the OMIG explainability for x as the difference of the end and the starting point of this path

$$\Delta x = x' - x. \tag{3}$$

In the experiments, we use $L = 50$ and $c = 1$ which we found to perform well for the considered use case. L should be large enough so that the according step size $\varepsilon = \frac{c}{L}$ will expectedly lead to a lowering of the midpoint in the update (2). The parameter c determines the length of the path leading to x'. Choosing a large c (and scaling L accordingly) will push x' toward a local minimum, while choosing a small c will reduce the OMIG explainability to the explainability of sensitivity analysis. As we want x' which is close to x but still allow for a modification $\Delta x = x' - x$ that contains sufficient information, c should be chosen neither too large nor too small. For the results in section 3 below we included a comparison for various choices of c in the appendix, in figure 14.

The visual performance of Δx can be substantially enhanced by removing 'outlier components' of the gradient $R(x_l)$ for $l \geqslant 1$ before performing the iteration step (2). To do this, we compute for each of the components $R(x_l)_i$ of the gradient $R(x_l)$ the following Z-score:

$$Z = \frac{R(x_l)_i - \mu\left(R(x_l)\right)}{\sigma\left(R(x_l)\right)},$$

where $\mu\left(R(x_l)\right)$ and $\sigma\left(R(x_l)\right)$ represent the mean and the standard deviation of the components of $R(x_l)$ respectively. A component $R(x_l)_i$ is then set to zero during the lth iteration if the absolute value of its Z-score is greater than a predefined threshold t. In this work, we used $t = 3$, which excludes, on average, about 1%–2% of the components in our case. The full algorithm for computing the OMIG explainability is summarized in algorithm 1.

Note that the dimension of Δx coincides with that of x, thus, we can plot it as an image (a 'heatmap'). From the construction of the OMIG explainability, we have the following interpretation of the sign of the pixels of Δx: given an image x and its OMIG explainability Δx, we know that,

- if, for pixels with a *positive* value in Δx, we *increase* the intensity of the corresponding pixel in x according to this value *and*
- if, for pixels with a *negative* value in Δx, we *decrease* the intensity of the corresponding pixel in x according to this value,

we expect to obtain an image with a *better predicted image quality*, i.e. we expect the CDC to decrease. In general, we expect image quality to improve if the following is done:

- pixels of the modified Δx that have positive values indicate pixels of the original x that must be *more visible* and
- pixels of the modified Δx that have negative values indicate pixels of the original x that must be *less visible*.

This information will be helpful in interpreting the results in section 3 below. Note, however, that the design of the proposed OMIG procedure means we only expect the predicted CDC to decrease if we modify all pixels in x according to their value in Δx *at the same time*; thus, one should be careful not to interpret local structures in Δx without taking the entire image into account. Moreover, our interpretation is restricted to the midpoint of the CDC only, rather than using the whole curve.

Algorithm 1. OMIG explainability.

```
Input: input image $x$, ensemble of neural networks, parameters $\varepsilon > 0$, $t > 0$ and an integer $L > 0$
$x_0 = x$;
Obtain initial gradient $R(x_0)$ with (1);
for each $l$ in $0, \ldots, L-1$ do
    $x_{l+1} = x_l - \varepsilon \cdot R(x_l)$;
    Estimate $R(x_{l+1})$ with (1);
    for every component $R(x_{l+1})_i$ of the vector $R(x_{l+1})$ do
        $Z \leftarrow \frac{R(x_{l+1})_i - \mu(R(x_{l+1}))}{\sigma(R(x_{l+1}))}$;
        if $|Z| > t$ then
            $R(x_{l+1})_i \leftarrow 0$;
        end if
    end for
end for
$x' = x_L$;
Output: OMIG explainability $\Delta x = x' - x$
```

Algorithm 2. Upsampling of an (OMIG) explainability map

```
Input: original (not downsampled) image $x_{\text{original}}$
for each $n$ in $1, \ldots, N$ do
    Downsample $x_{\text{original}}$ to $x_{\text{downsampled}}$;
    Save chosen pixels in $\mathfrak{I} = \{(\iota_{x,i}, \iota_{y,j})\}_{i,j=1,\ldots,250}$;
    Estimate $\Delta x$, as in Algorithm 1, for $x_{\text{downsampled}}$;
    Create $\delta$ filled with zeros, with $\dim(\delta) = \dim(x_{\text{original}})$;
    for each $i, j$ in $1, \ldots, 250$ do
        $\delta_{\iota_{x,i},\iota_{y,j}} \leftarrow \Delta x_{i,j}$
    end for
    if $n = 1$ then
        $(\Delta x)_{\text{upsampled}} \leftarrow \frac{1}{N}\delta$;
    else
        $(\Delta x)_{\text{upsampled}} \leftarrow (\Delta x)_{\text{upsampled}} + \frac{1}{N}\delta$;
    end if
end for
Output: $(\Delta x)_{\text{upsampled}}$
```

In section 3 below we will use for comparison two other methods from the literature that are somewhat conceptually close to the presented method, namely occlusion sensitivity [32] and sensitivity analysis [33]. For occlusion sensitivity the explainability value (in respect to the CDC midpoint) for an input x at a pixel i is given by $\frac{1}{K}\sum_{k=1}^{K} \tilde{\mu}^k(x_{[i]}) - \frac{1}{K}\sum_{k=1}^{K} \tilde{\mu}^k(x)$, where $x_{[i]}$ is constructed from x via setting its value at i to 0 and where we summarized, similar as for OMIG, the CDC by its midpoint. Occlusion sensitivity measures, by construction, the impact of single pixels. Sensitivity analysis, on the other hand, shows the gradient of $-\frac{1}{K}\sum_{k=1}^{K} \tilde{\mu}^k(x)$ with respect to x, i.e. with the notation from (1), $-R(x)$. Sensitivity analysis therefore shows the effect of changing all pixels at once. Note that due to (2) we have for OMIG $\Delta x = -\sum_{l=0}^{L-1} \varepsilon R(x_l) = -\sum_{l=0}^{L-1} \frac{c}{L} R(x_l)$, so that the OMIG method can be seen as a way to enrich the information of the sensitivity analysis gradient $R(x)$ by 'integrating' it along a path. In addition to the methods presented in figure 3, we studied layerwise-relevant-propagation [34] but found it, as in [6], unsatisfactory for our use case.

2.3.2. Upsampling explainability maps

As explained in section 2.2, the data used to train the neural networks are downsampled to a resolution of 250×250 by randomly selected pixels in the original images while keeping track of their original order [6]. As a result of this procedure, 99% of the original image is omitted. While this leads to a better convergence of the training and requires less resources for computation, it raises the problem that the trained neural network can only process downsampled images. Specifically, any OMIG explainability Δx as in (3) will have the dimensions of the downsampled input. For the sake of interpretability, it would be more useful to have an explainability map in the resolution of the original, not downsampled, image instead, as this would allow us to study the impact of finer structures such as individual disks. To this end, we here propose an *upsampling* of the explainability map.

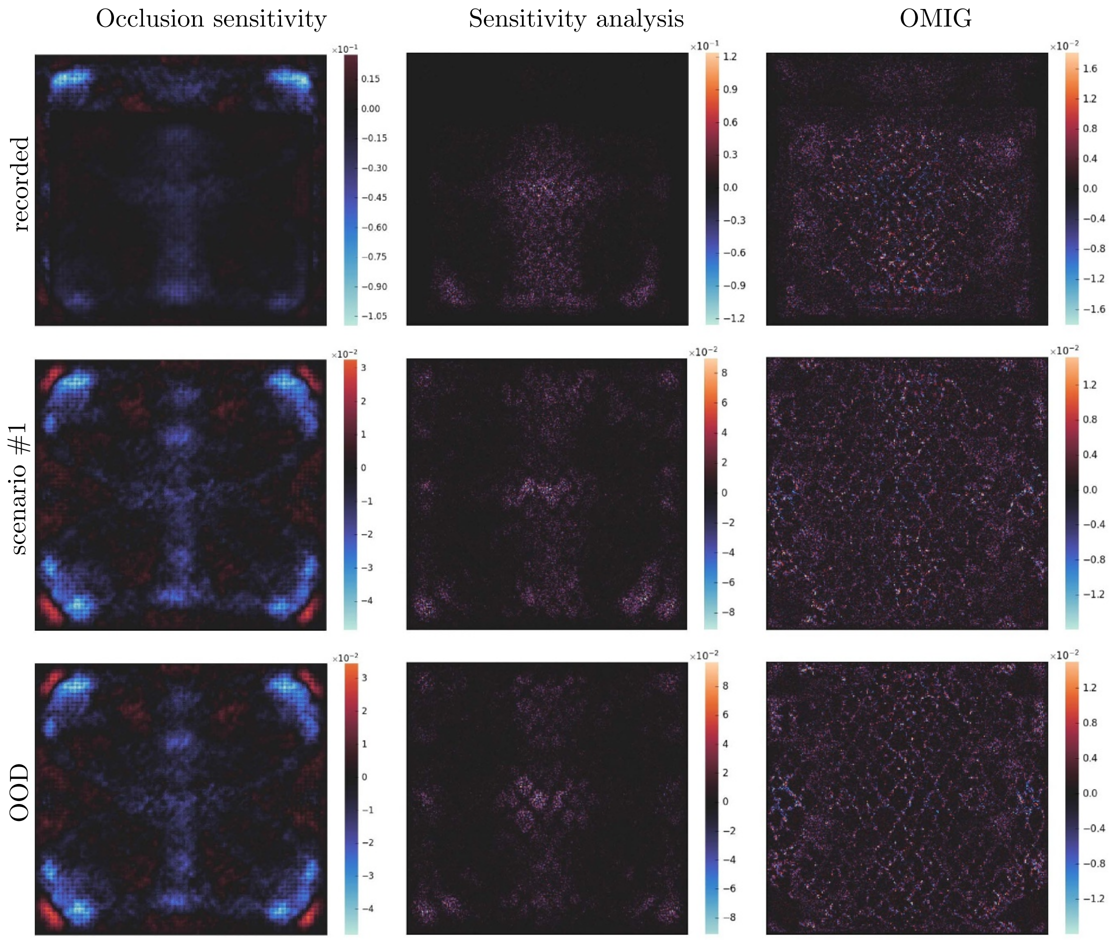


Figure 3. *Not* upsampled heatmaps of various explainability methods for the images from figure 1. The different rows show the three different images from figure 1. From top to bottom: the recorded image from figure 1(a), the scenario #1 image from figure 1(b) and the OOD image from figure 1(c). The different columns show different explainability methods. From left to the right: occlusion sensitivity, sensitivity analysis, OMIG, the one proposed in this work.

The upsampling procedure is described in detail in algorithm 2. Here, the image is downsampled several times and the results of the OMIG explainability are collected in a single explainability map of the size of the original, not downsampled, image. In each iteration, the 250 × 250 randomly chosen pixels $\mathfrak{I}$ are saved. After computing the explainability of map Δx as in (3), the values of Δx are then assigned to a high-resolution δ, with the same dimensions as the original, not downsampled, image, according to the pixels stored in $\mathfrak{I}$. All pixels of δ that are not stored in $\mathfrak{I}$ are filled with zeros. This procedure is repeated N times and the obtained δ's are averaged to yield an upsampled explainability map $(\Delta x)_{\text{upsampled}}$ that has the same dimensions as the original, not downsampled, image.

3. Results

The right column of figure 3 shows the resulting explainability maps obtained by the proposed OMIG method. The rows of figure 3 represent the explainability maps for three different images: the recorded image from figure 1(a) in the top row, the simulated image from scenario #1 from figure 1(b) in the middle row and the simulated OOD image from figure 1(c) in the lower row. As seen in the OMIG explainability maps, the OMIG method highlights a large area of the input image and successfully detects margins. Most importantly, however, when compared to the sensitivity analysis whose results are presented in the middle column and the occlusion sensitivity from the left column, the OMIG method is the only method presented that actually highlights expressive patterns. The CDMAM phantom grid is clearly visible on the OMIG heatmaps, but due to the downsampling procedure it is impossible to determine if any other phantom features besides the grid are highlighted.

It is evident that each of the four methods studied produces distinct results. Occlusion sensitivity and sensitivity analysis put a particular emphasis on the middle region. However, the highlighted region is blurry and does not highlight any more specific structures. Moreover, for occlusion sensitivity, this region is

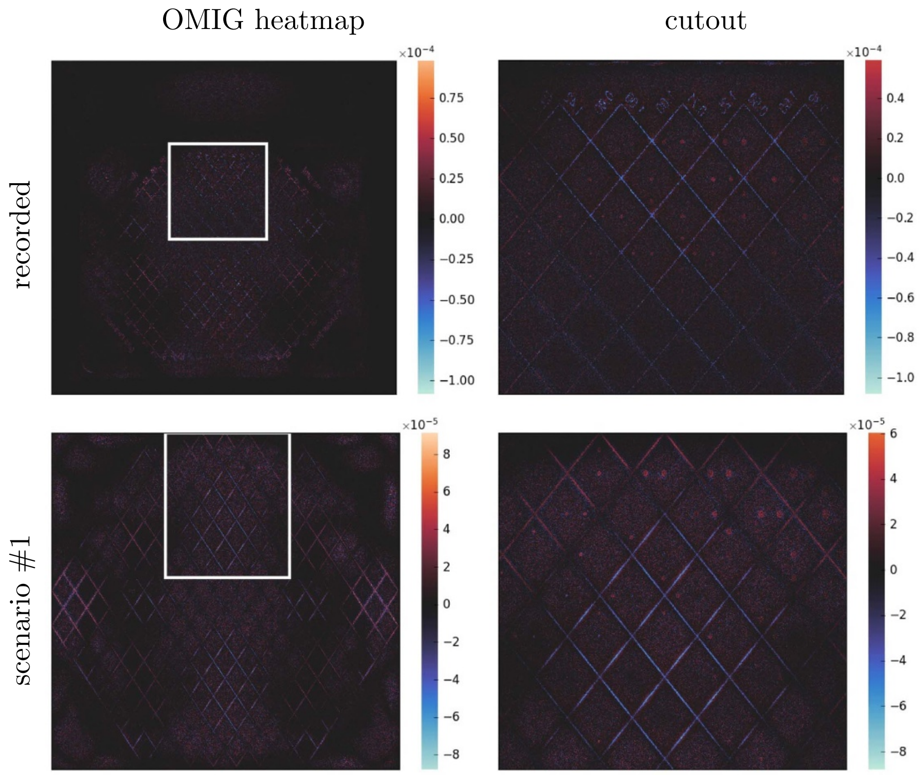


Figure 4. Upsampled OMIG explainability maps for the recorded image from figure 1(a) and the simulated, scenario #1 image from figure 1(b). The white rectangles in the left column are shown enlarged in the right column.

invariant of the position of the actual phantom, as can be seen from its performance in the recorded image with a margin (figure 1(a)), in the image from scenario #1 (figure 1(b)) and the image from the OOD set (figure 1(c)), both of which had no margin.

Due to the fact that the heatmaps in figure 3 were obtained from highly downsampled image data, as explained in section 2.2, the explainability maps on the downsampled images are not particularly useful for analyzing the impact of finer structures in the phantom such as individual disks. For this reason, we now employ the upsampling procedure from algorithm 2 to construct high-resolution explainability maps. This involves evaluating the OMIG algorithm from section 2.3.2 N-times, where we chose $N = 5000$. While an evaluation of the OMIG explainability for a single downsampled image takes a few seconds only, executing a loop as in algorithm 2 takes around 4.5 h. In principle, this process is parallelizable and can thus be sped up. For occlusion sensitivity, the upsampling consumes too many computational resources to be of any practical use. In figure 12 in the appendix, we included the upsampled heatmaps for the sensitivity analysis method: here, in contrast to the OMIG method, no finer details become visible through upsampling. We will therefore restrict ourselves to the OMIG method from this point on.

The upsampled explainability maps for the recorded image from figure 1(a) and the simulated scenario #1 image from figure 1(b) are presented in figure 4. The left column of figure 4 shows the entire, upsampled heatmap for these two images (top: recorded image, bottom: simulated scenario #1 image). In the upper part of both heatmaps, individual disks are clearly visible. These parts are shown enlarged in the second column of figure 4. The corresponding area is marked by a white rectangle in the left column. As explained in section 2, the presented OMIG heatmaps were produced with $c = 1$ and $L = 50$.

The maps in figure 4 indicate that the neural networks trained for mammography IQA learned a ‘meaningful’ mapping, that is, a mapping that uses the geometrical structures of the phantom and does not, or does not only, rely on structure-agnostic aspects such as the noise level or on a ‘Clever Hans’ effect [35]. The color chosen for the pixels in the explainability map in figure 4 is red for pixels i with $\Delta x_i > 0$, blue for those with $\Delta x_i < 0$ and black for $\Delta x_i = 0$. For the interpretation of these colors, let us recall the logic behind the OMIG method: if we are given an image x and if, for each pixel i that is

- red (i.e. $\Delta x_i > 0$), the intensity of x_i will be *increased* and if, for each pixel that is
- blue (i.e. $\Delta x_i < 0$), the intensity of x_i will be *decreased*,

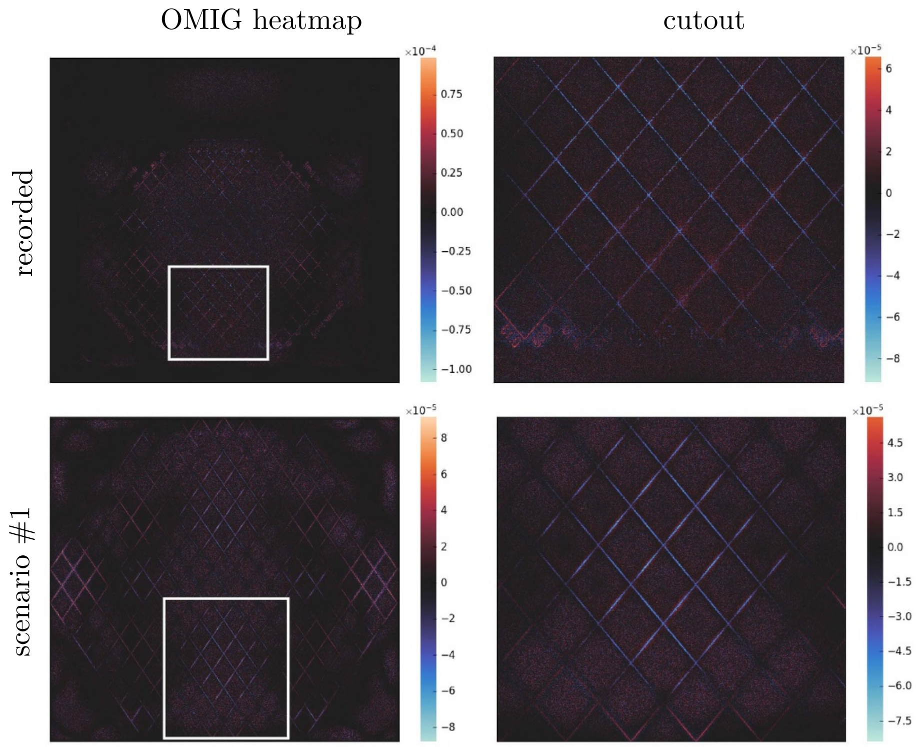


Figure 5. Different cutouts of the same OMIG explainability maps as in figure 4. The white rectangles in the left column are shown enlarged in the right column.

then we expect the deep ensemble to predict a *better image quality*, i.e. a lower CDC. Figure 13 in the appendix shows for the images considered in figures 4 and 6 below, the effect of the corresponding pixel modification on the predicted CDC. This illustrates that the OMIG heatmap does indeed highlight those features that are expected to influence the predicted CDC to move downwards.

Looking at the cutouts in figure 4, we observe that almost all of the disks in the cutouts are red. Broadely speaking, this means that increasing the intensity of the disks is expected to increase the predicted image quality. This is to be expected, as the detectability of single disks is what the CDC is designed to quantify. However, note that only a part of the disks and also other structures influence the prediction. Figure 5 shows further cutouts of the explainability maps from figure 4. Apparently, the disks in these cutout region do not really contribute to the prediction of the neural network.

The difference in the highlighting of disks between the cutouts in figures 4 and 5 is easy to understand. The cutout in figure 4 shows an area of the phantom with the disks of larger diameters, whereas the region depicted in the cutout in figure 5 contains the disks of smaller diameters (cf figure 8 in the appendix). The disks with larger diameters are more likely to be chosen during the downsampling procedure and therefore, on average, have a higher impact on the prediction of the trained neural network. This shows that the algorithmic choice to downsample images before training the networks has a noticeable effect on their ‘functional principle’.

The pattern behind the explainability values for the grid is more complex. For the examples used in figures 3 and 4, we observe a vertical strip of the grid, colored in blue, that contains the area where disks are highlighted in red. A possible explanation for this pattern is as follows: in figure 1, the lines of the grid become randomly corrupted due to the downsampling. These random shifts could make it harder to detect disks in front of their noisy background. Thus, in the areas where the disks are typically detected, such as the upper part of the images in figure 4, it may be possible to improve the detection performance by lowering the intensity of the surrounding grid while increasing the disk intensity: this could explain the blue-red coloring we see in the cutout. The images from the training data set are randomly mirrored horizontally and vertically, but are not rotated. Since the explainability map depends not only on the input image but also on the trained parameters, this might lead to a certain symmetry, as reflected in the blue coloring of the grid expanding to the bottom.

Note also that, for the recorded image in the first row of figure 3, we see that both the grid and the writing on the CDMAM phantom that surrounds the grid are highlighted. The explainability coloring of this writing follows a similar scheme and could thus be explained along the lines of the hypothesis mentioned above.

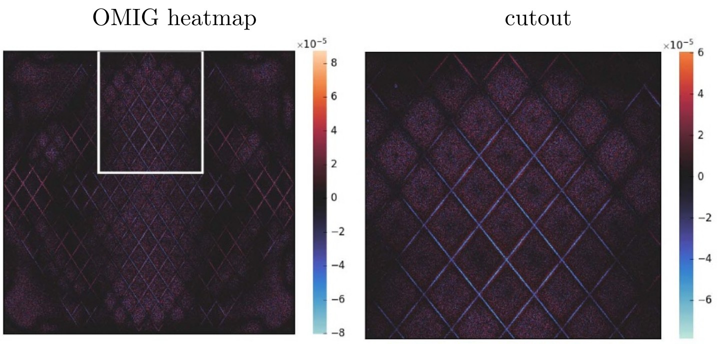


Figure 6. The upsampled OMIG explainability map for an image from the simulated scenario #4. The white rectangle in the left column is shown enlarged in the right column.

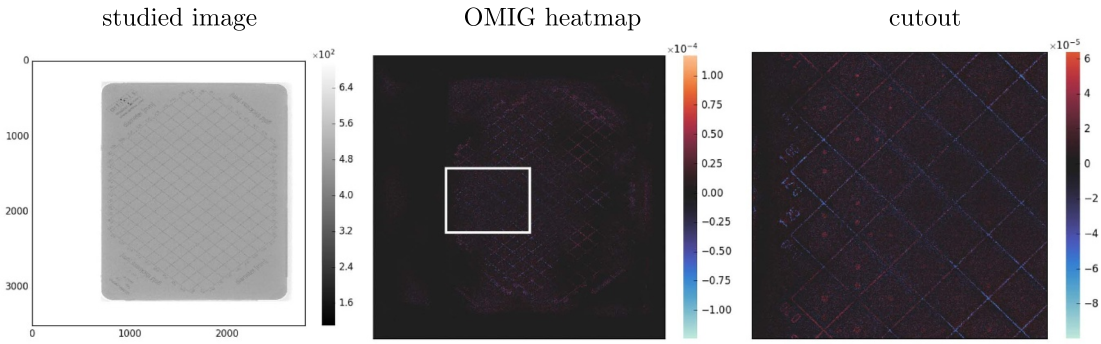


Figure 7. OMIG explainability (middle column) and a cutout (right column) for a rotated modification of the recorded image from figure 1, shown in the left column.

This shows that these features also have an influence on the prediction and could suggest an improvement of the network performance if such artifacts were included in the simulation.

We can thus summarize that the prediction of the trained neural network is influenced by the disks that have a large enough diameter to 'survive' the downsampling, but, in addition, also by the grid and the writing on its sides. These findings exemplify that a data-driven model does typically not follow the mechanics that might be most natural to us [12] and show the importance of explainability methods, such as OMIG, that tell us what features such a model actually relies on.

An explainability map allows an understanding not only of how a neural network creates a certain prediction but also why it fails to work as expected in certain cases. We saw in section 2.2 that simulation scenario #4, which has a particularly high noise level, could be considered such a case, since it leads to substantially increased errors. These errors are not properly covered by the predicted uncertainty of the deep ensemble, but can be better understood by means of the OMIG explainability. Figure 6 shows the OMIG explainability heatmap and a cutout similar to those in figure 3 but for an image from scenario #4. We observe that, while the grid seems to have an influence on the prediction, that is similar to the influence of the grid shown in figure 4, the disks do not appear in the explainability map. This indicates that the high noise level hinders the neural networks in detecting the disks of the phantom; thus, the networks can only base their prediction on the grid, which leads to a substantially worse performance, as seen in figure 2.

As a final example, we study the effect of a corrupted input image. The left column of figure 7 shows the recorded image from figure 1(a), but rotated by 90°. The middle column of figure 7 contains the corresponding OMIG explainability map. The right column then shows a cutout (an enlarged white rectangle in the middle row). For the heatmap, we observe that the grid coloring largely resembles that of the unrotated image in figure 3. This is reasonable because, for the training of the neural networks, no rotated images were considered; thus, the networks 'assume' the grid is oriented as in figure 1. Note that, even in the rotated image, some disks are highlighted; this is shown in the cutout region in the right column. Even though this region is now located in an area different to that of figure 3, we still observe the familiar red-colored disks surrounded by a blue-colored grid. In other words, by increasing the intensity of the disks and decreasing the intensity of the surrounding grid, we expect to increase the predicted image quality. This indicates that at least some of the networks use the shape of geometrical structures for their prediction, which suggests a certain robustness of the learned mapping.

Let us summarize our observations:

- The introduced OMIG explainability method reveals that the learned neural networks use geometrical patterns in the input image, such as the disks of the CDMAM phantom. This indicates that the deep ensemble seems to have learned a meaningful mapping, although it uses different information than a human observer or the CDMAM Analyser software. The disks which have the greatest impact are those that have the highest chance of surviving the downsampling.
- However, the highlighted patterns not only include the disks but also the grid itself and the writing on its sides. For the grid, OMIG reveals a rather complex pattern behind the prediction; this pattern could be the subject of future work.
- The patterns still appear in corrupted (i.e. rotated) images, which suggests a certain robustness of the learned neural networks.
- For cases in which the neural networks fail to work as expected, such as simulation scenario #4, the OMIG explainability helps to explain this poor performance.

These observations provide a rationale for the use of the OMIG method for a deep learning-based IQA in mammography, as it reduces the deep learning black-box nature and allows the limits of the learned networks to be studied, thereby revealing opportunities to improve the approach and enhance its trustworthiness.

4. Limitations and outlook

In section 3, we observed that the OMIG explainability maps mark certain disks as being of particular relevance for the prediction of the CDCs. This reveals two insights: first, as explained above, we can deduce that the employed downsampling procedure has a direct consequence on the learned mapping. Second, it follows that the neural network has learned a mapping, that while still using geometric structures in the phantom, is not identical to the one a human observer or the CDMAM Analyser software uses. The neural network uses only a part of the disks and not necessarily those that are most relevant for a non-deep-learning-based prediction. In addition, the explainability maps in figures 4, 6 and 7 show that other structures such as the grid and the writing also have an impact on the prediction of the CDCs made by the neural networks. In particular, we see that not all disks are necessary and that other structures can be used as well to correctly predict the image quality if one is willing to use a data-driven model. A more in-depth study of such findings may be helpful for designing alternatives to the CDMAM phantom such as anthropomorphic breast phantoms [36–38].

The method in this work builds on the integrated gradients (IG) approach from [39] but modifies it, by removing the need for a baseline and exploiting instead the natural direction toward an expected better image quality. This leads to a rather complex path connecting the input image x and its updated version x', in contrast to the path considered in the IG approach, which is merely a straight line connecting x and the baseline. While the 'IG path' will therefore always end with the baseline, the 'OMIG path' will end with different endpoints x' that are dependent on x and the chosen hyperparameters c and L. In this work, we used a large L; this was reasonable because, in the limit $L \to \infty$, the updates from (2) were expected to describe a continuous path in the input space. However, no such unique reasonable choice is possible for the proportionality factor c. The comparison of various c is presented in figure 14 in appendix E. We observed, empirically, that $c = 1$ works quite well for our use case; however, other factors may be more effective for other applications. The question of how to choose this parameter may, along with a study of the performance of the OMIG method for different tasks, be an interesting topic for future research.

From the explanation above, one can see that the OMIG approach is, in principle, applicable to any problem (e.g. in IQA but also in a broader context) that possesses a natural 'direction'. Given such a direction, one can choose a value for R that is similar to that in (1) in order to point in (the inverse of) this direction and then update x to create an x' according to (2). Finally, we can define the OMIG explainability via (3), which gives information on how x must be modified to shift the prediction toward the natural direction. For other IQA problems, this direction might again be expected to describe an improvement of the image quality. In a classification problem, on the other hand, such a direction might be chosen to make a certain class more likely. An investigation of these questions is reserved for future work.

5. Conclusion

We introduced a new explainability method (OMIG) for a deep learning-based IQA in mammography. Compared to the other explainability methods, OMIG yields more expressive results for this use case. To deal with the issue of the downsampled data, we proposed a simple upsampling algorithm that provided us with high-resolutional explainability maps based on which we were able to study the impact of fine structures such as individual disks. These maps show that the trained neural networks use the disks of the CDMAM phantom and indicate that the networks have learned a meaningful relationship. The explainability maps also reveal which of these disks and which other structures in the phantom have an impact on the prediction. This allows not only the mechanisms behind the neural network's prediction to be analysed but also cases in which the networks fail to perform as expected to be explained. The presented OMIG approach could also be useful for other problems that possess a natural direction.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

This work has been carried out within the PTB Project ML4MedIm. The authors would like to thank colleagues from the M4AIM Project for fruitful discussions and the referees for their helpful comments and suggestions.

Appendix A. CDMAM phantom

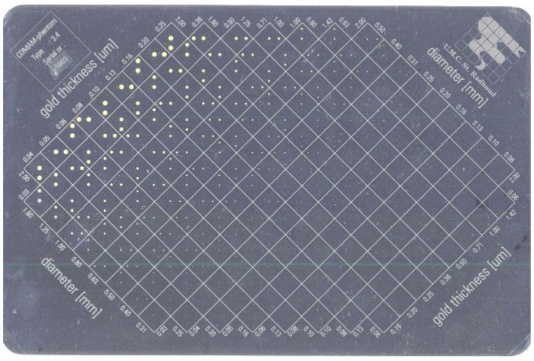


Figure 8. The CDMAM phantom consists of 205 square cells separated by a grid with each cell having two gold disks: one located at the cell center and the other randomly located at one of the corners [40]. We refer to [16, 18] for more detailed information on the contrast–detail curve estimation from the CDMAM phantom.

Appendix B. Simulated recordings

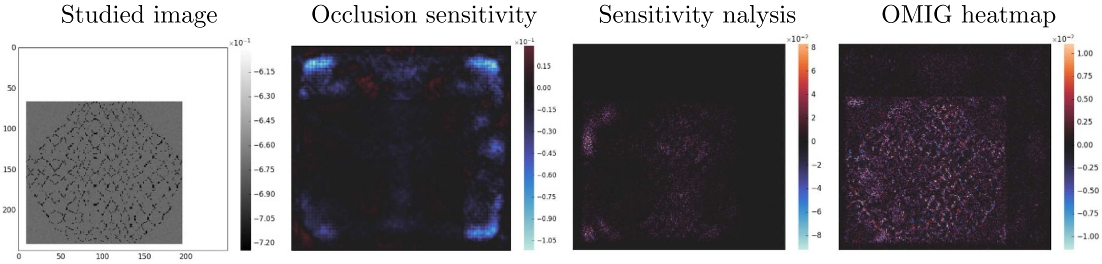

Studied image Occlusion sensitivity Sensitivity nalysis OMIG heatmap

Figure 9. Example of the downsampled image from the simulated test data set with margins (left column). *Not* upsampled heatmaps of various explainability methods. The different columns show different explainability methods. From second left to the right: occlusion sensitivity, sensitivity analysis and OMIG.

Appendix C. Network architecture

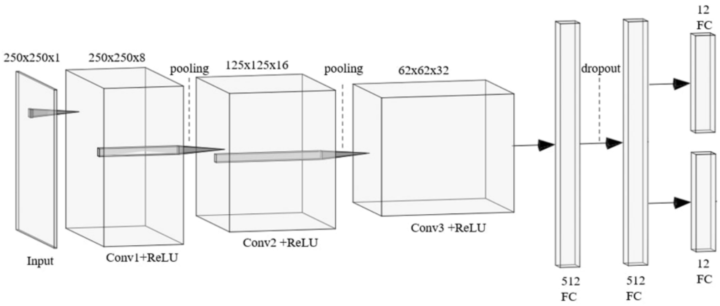


Figure 10. An illustration of the neural network architecture, created using [41].

Table 2. The detailed neural network architecture for the CDCs prediction.

Layer	Filter dimension	Output Dimension
Image input		$250 \times 250 \times 1$
Convolutional	3×3	$250 \times 250 \times 8$
Batch normalization + ReLU		$250 \times 250 \times 8$
Max pooling	2×2	$125 \times 125 \times 8$
Convolutional	3×3	$125 \times 125 \times 16$
Batch normalization + ReLU		$125 \times 125 \times 16$
Max pooling	2×2	$62 \times 62 \times 16$
Convolutional	3×3	$62 \times 62 \times 32$
Batch normalization + ReLU		$62 \times 62 \times 32$
Fully connected	$123\,008 \times 512$	$1 \times 1 \times 512$
dropout		
Fully connected	512×512	$1 \times 1 \times 512$
μ	512×12	$1 \times 1 \times 12$
σ	512×12	$1 \times 1 \times 12$

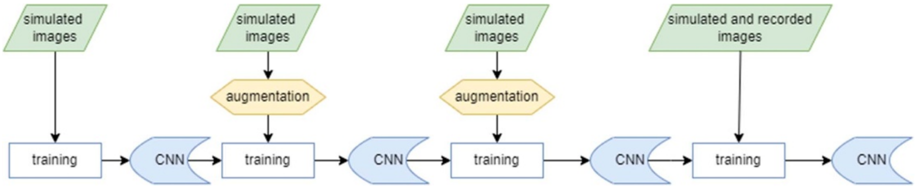


Figure 11. Illustration of the incremental learning strategy used for training process [5] created using [42].

Appendix D. Traing details

The training followed an incremental learning strategy based on the strategy from [5] and is depicted in figure 11 below. In the first training step, only the simulated images with no margins were used. The smaller margins were added in the second training step, while the largest margins were applied in the third step. In the last training step, the recorded and the simulated images of the CDMAM phantom were included.

Appendix E. Explainability maps

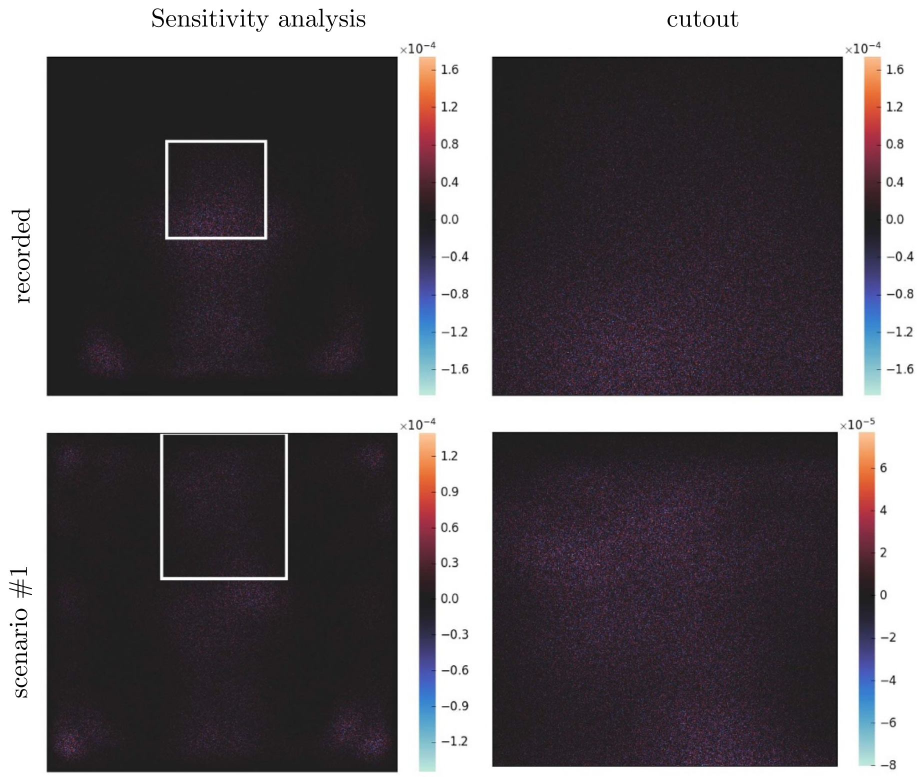


Figure 12. Upsampled sensitivity analysis explainability maps of the recorded image from figure 1(a) and the simulated, scenario #1 image from figure 1(b). The white rectangles in the left column are shown enlarged in the right column.

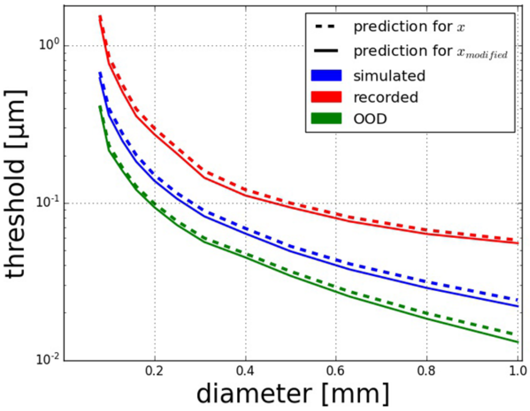


Figure 13. The dashed lines demonstrate the predicted CDCs for the original downsampled input x, while the solid lines show the predicted CDCs for the modified input x_{modified}. The modification was performed by adding the upsampled explainability map $(\Delta x)_{\mathrm{upsampled}}$ to the original not downsampled image x_{original}. The resulting not downsampled image $x_{\mathrm{original}} + (\Delta x)_{\mathrm{upsampled}}$ was then downsampled to x_{modified} in order to predict the CDCs.

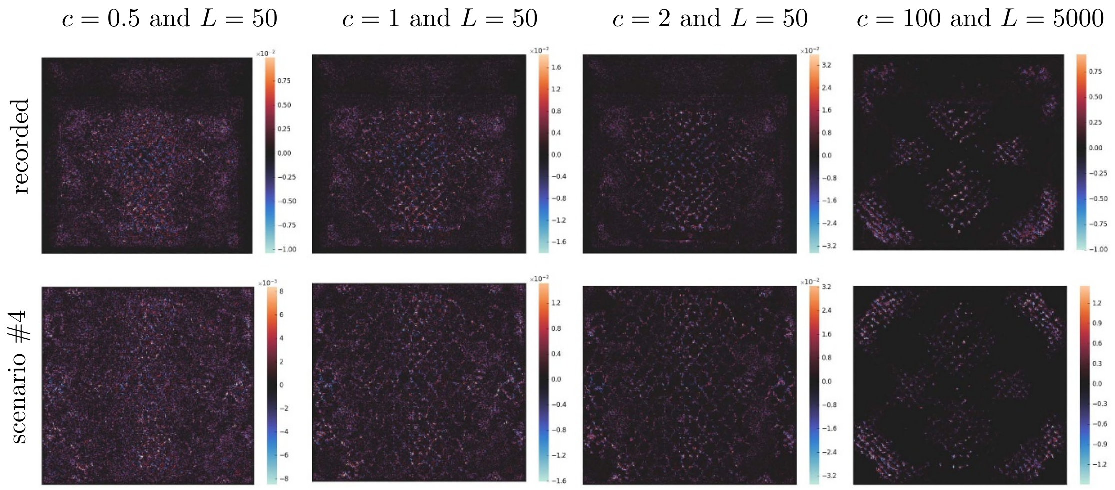


Figure 14. Comparison of the downsampled OMIG explainability maps for various choices of *c* and *L*. The rows show different images from figure 1: the recorded image from figure 1(a) at the top and the scenario #1 image from figure 1(b) at the bottom. The different columns show different values of *c*.

ORCID iDs

N Amanova https://orcid.org/0000-0002-4193-8906
J Martin https://orcid.org/0000-0001-5066-7661
C Elster https://orcid.org/0000-0003-0113-3713

References

[1] Ding K, Ma K, Wang S and Simoncelli E P 2020 Image quality assessment: unifying structure and texture similarity *IEEE Trans. Pattern Anal. Mach. Intell.* **44** 2567–81

[2] Barca P, Lamastra R, Aringhieri G, Tucciariello R M, Traino A and Fantacci M E 2019 Comprehensive assessment of image quality in synthetic and digital mammography: a quantitative comparison *Australas. Phys. Eng. Sci. Med.* **42** 1141–52

[3] European Commission, EUREF, EBCN and EUSOMA 2013 *European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis. Fourth Edition*. ed N Perry, M Broeders, C de Wolf, S Törnberg, R Holland, L von Karsa and E Puthaar (Luxembourg: Office for Official Publications of the European Union)

[4] European Commission, EUREF, EBCN and EUSOMA 2006 *European Quidelines for Quality Assurance in Breast Cancer Screening and Diagnosis. Fourth Edition. The Supplements of the European Guidelines* ed N Perry, M Broeders, C de Wolf, S Törnberg, R Holland, L von Karsa and E Puthaar (Luxembourg: Office for Official Publications of the European Communities)

[5] Kretz T, Mueller K-R, Schaeffter T and Elster C 2020 Mammography image quality assurance using deep learning *IEEE Trans. Biomed. Eng.* **67** 3317–26

[6] Kretz T 2020 Development of model observers for quantitative assessment of mammography image quality *Doctoral Thesis* Technische Universität Berlin, Berlin

[7] Piccini D, Demesmaeker R, Heerfordt J, Yerly J, Di Sopra L, Masci P G, Schwitter J, Van De Ville D, Richiardi J, Kober T *et al* 2020 Deep learning to automate reference-free image quality assessment of whole-heart MR images *Radiol. Artif. Intell.* **2** e190123

[8] Chan E J J, Najjar R P, Tang Z and Milea D 2021 Deep learning for retinal image quality assessment of optic nerve head disorders *Asia-Pac. J. Ophthalmol.* **10** 282–8

[9] Ranschaert E R, Morozov S and Algra P R 2019 *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks* (Berlin: Springer)

[10] Zhang S, Wang Y, Jiang J, Dong J, Yi W and Hou W 2021 CNN-based medical ultrasound image quality assessment *Complexity* **2021** 9938367

[11] Jiang T, Hu X-J, Yao X-H, Tu L-P, Huang J-B, Ma X-X, Cui J, Wu Q-F and Xu J-T 2021 Tongue image quality assessment based on a deep convolutional neural network *BMC Med. Inform. Decis. Mak.* **21** 147

[12] Burkart N and Huber M F 2021 A survey on the explainability of supervised machine learning *J. Artif. Intell. Res.* **70** 245–317

[13] Samek W, Montavon G, Vedaldi A, Hansen L K and Müller K-R 2019 *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* vol 11700 (Cham: Springer Nature)

[14] Molnar C 2022 *Interpretable machine learning* (Morrisville, NC: Lulu.com) p 320

[15] de las Heras Gala H, Schöfer F, Tiller B, del Río M C, Zwettler G and Semturs F 2015 A new method for dosimetry and image quality assurance in mammography and breast tomosynthesis (including abstracts 2358373 and 2492856)

[16] Karssemeijer N and Thijssen M 1996 Determination of contrast-detail curves of mammography systems by automated image analysis *Digit. Mammogr.* **96** 155–60

[17] Young K, Alsager A, Oduko J, Bosmans H, Verbrugge B, Geertse T, Engen R and Nccpm A 2008 Evaluation of software for reading images of the CDMAM test object to assess digital mammography systems *Proc. SPIE* **6913** 69131C

[18] Young K C, Cook J J and Oduko J M 2006 Automated and human determination of threshold contrast for digital mammography systems *Int. Workshop on Digital Mammography* (Springer) vol 4046 pp 266–72

[19] Young K C, Cook J J, Oduko J M and Bosmans H 2006 Comparison of software and human observers in reading images of the CDMAM test object to assess digital mammography systems *Proc. SPIE* **6142** 614206

[20] Young K, Brookes E, Hudson W and Halling-Brown M 2011 CDMAM analyser: software and instruction manual for automated determination of threshold contrast (Guildford: National Co-ordinating Centre for the Physics of Mammography)

[21] Lakshminarayanan B, Pritzel A and Blundell C 2016 Simple and scalable predictive uncertainty estimation using deep ensembles (arXiv:1612.01474)

[22] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)

[23] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J V, Lakshminarayanan B and Snoek J 2019 Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift (arXiv:1906.02530)

[24] Caldeira J and Nord B 2020 Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms *Mach. Learn.: Sci. Technol.* **2** 015002

[25] Arnez F, Espinoza H, Radermacher A and Terrier F 2020 A comparison of uncertainty estimation approaches in deep learning components for autonomous vehicle applications (arXiv:2006.15172)

[26] Hoffmann L and Elster C 2020 Deep neural networks for computational optical form measurements *J. Sens. Sens. Syst.* **9** 301–7

[27] Kompa B, Snoek J and Beam A L 2021 Second opinion needed: communicating uncertainty in medical machine learning *npj Digit. Med.* **4** 1–6

[28] Alizadehsani R *et al* 2021 Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991–2020) *Ann. Oper. Res.* 1–42

[29] Kretz T, Anton M, Schaeffter T and Elster C 2019 Determination of contrast-detail curves in mammography image quality assessment by a parametric model observer *Phys. Med.* **62** 120–8

[30] Kendall A and Gal Y 2017 What uncertainties do we need in Bayesian deep learning for computer vision? (arXiv:1703.04977)

[31] Okajima Y and Sadamasa K 2019 Deep neural networks constrained by decision rules *Proc. AAAI Conf. on Artificial Intelligence* vol 33 pp 2496–505

[32] Zeiler M D and Fergus R 2014 Visualizing and understanding convolutional networks *European Conf. on Computer Vision* (Springer) pp 818–33

[33] Montavon G, Samek W and Müller K-R 2018 Methods for interpreting and understanding deep neural networks *Digit. Signal Process.* **73** 1–15

[34] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W and Suarez O D 2015 On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation *PLoS One* **10** e0130140

[35] Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W and Müller K-R 2019 Unmasking Clever Hans predictors and assessing what machines really learn *Nat. Commun.* **10** 1096

[36] Ikejimba L C, Graff C G, Rosenthal S, Badal A, Ghammraoui B, Lo J Y and Glick S J 2017 A novel physical anthropomorphic breast phantom for 2D and 3D x-ray imaging *Med. Phys.* **44** 407–16

[37] Balta C, Bouwman R W, Sechopoulos I, Broeders M J M, Karssemeijer N, van Engen R E and Veldkamp W J H 2018 A model observer study using acquired mammographic images of an anthropomorphic breast phantom *Med. Phys.* **45** 655–65

[38] Balta C, Bouwman R W, Sechopoulos I, Broeders M, Karssemeijer N, van Engen R and Veldkamp W 2019 Can a channelized Hotelling observer assess image quality in acquired mammographic images of an anthropomorphic breast phantom including image processing? *Med. Phys.* **46** 714–25

[39] Sundararajan M, Taly A and Yan Q 2017 Axiomatic attribution for deep networks (arXiv:1703.01365 [cs.LG])

[40] Thomas J A, Chakrabarti K, Kaczmarek R and Romanyukha A 2005 Contrast-detail phantom scoring methodology *Med. Phys.* **32** 807–14

[41] LeNail A 2019 NN-SVG: publication-ready neural network architecture schematics *J. Open Source Softw.* **4** 747

[42] JGraph 2021 *diagrams.net, draw.io (Version: 15.5.2)* (available at: https://www.diagrams.net/) (accessed 14 October 2021)