PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

End-to-end mammographic breast density quantification with deep learning: preliminary study on simulated mammograms

Sjoerd A. Tunissen, Andrea Motta, Franziska Mauter, Eloy García, Oliver Diaz, et al.

Sjoerd A. M. Tunissen, Andrea Motta, Franziska Mauter, Eloy García, Oliver Diaz, John M. Boone, Ioannis Sechopoulos, Marco Caballo, "End-to-end mammographic breast density quantification with deep learning: preliminary study on simulated mammograms," Proc. SPIE 12465, Medical Imaging 2023: Computer-Aided Diagnosis, 124650K (7 April 2023); doi: 10.1117/12.2654131

SPIE.

Event: SPIE Medical Imaging, 2023, San Diego, California, United States

End-to-End Mammographic Breast Density Quantification with Deep Learning: Preliminary Study on Simulated Mammograms

Sjoerd A. M. Tunissen[1], Andrea Motta[1], Franziska Mauter[1, 2], Eloy García[3], Oliver Diaz[3], John M Boone[4, 5], Ioannis Sechopoulos[1, 6, 7], Marco Caballo[1]

1 Department of Medical Imaging, Radboud University Medical Center, PO Box 9101, 6500 HB Nijmegen, The Netherlands.
2 Physikalisch-Technische Bundesanstalt, Bundesallee 100, 38116 Braunschweig, Germany
3 Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain.
4 Department of Radiology, University of California Davis, Sacramento, California, United States
5 Department of Biomedical Engineering, University of California Davis, Sacramento, California, United States
6 Dutch Expert Centre for Screening (LRCB), Nijmegen, The Netherlands.
7 Technical Medical Centre, University of Twente, Enschede, The Netherlands

ABSTRACT

High breast density (BD) is recognized as an independent risk factor for breast cancer development, in addition to negatively impacting the sensitivity of mammography. Although BD is normally assessed with the BI-RADS reporting system, this evaluation is qualitative and has been shown to vary considerably across readers. In this pilot study, we present a deep learning (DL) method to quantify BD from a standard two-view (cranio-caudal, and medio-lateral-oblique) mammography exam. With the aim of developing a method based on an objective ground truth, the DL model was trained and validated using 88 simulated mammograms from an equal number of distinct 3D digital breast phantoms for which BD is known. The phantoms had been previously generated through segmentation and simulated mechanical compression of patient dedicated breast CT images, allowing for the exact calculation of BD in each case. Different data augmentations were applied prior to simulation, to increase the dataset size, yielding a total of 528 cases. These were divided, randomly and on a patient level, into training (N=360), validation (N=60), and test sets (N=108). The DL model performance was tested by stratifying the breasts into four different density ranges: 1-15%, 15-25%, 25-60%, and >60%. The median absolute errors and interquartile ranges (IQR), in percentage points, were 3.3 (IQR: 3.5), 3.4 (IQR: 2.5), 3.5 (IQR: 3.9), and 14.8 (IQR: 8.4), respectively. Although preliminary, these results show the potential of the proposed approach for accurate BD quantification, which is based, as opposed to most previously proposed approaches, on an objective ground truth.

SUMMARY

In this work we present a deep learning (DL) based method to estimate breast density from simulated digital mammograms using the two standard views of a mammographic exam (cranio-caudal and medio-lateral oblique) as the main inputs to the DL model. For training and validating the DL model, ray-traced mammograms simulated from patient-based 3D digital breast phantoms (with known density) were used. The DL model was able to estimate breast density in our test set with an overall median absolute error of 3.6 percentage point, indicating the potential of the proposed approach.

Keywords: breast density; breast CT; digital mammography; deep learning

Medical Imaging 2023: Computer-Aided Diagnosis, edited by Khan M. Iftekharuddin, Weijie Chen, Proc. of SPIE Vol. 12465, 124650K · © 2023 SPIE
1605-7422 · doi: 10.1117/12.2654131

1. DESCRIPTION OF PURPOSE

Breast density is an important breast cancer risk factor [1]. For this reason, the estimation of breast density from digital mammograms is a highly active area of research [2], including the development of Artificial Intelligence (AI) models for this task. However, these models mainly characterize breast density based on the BI-RADS breast density categories, or based on ground truth obtained with annotations from human readers [3]–[8].

In this pilot study, we developed a deep learning model to quantify breast density from simulated mammograms obtained from patient-based phantoms with known density, to test if DL could potentially quantify breast density, and thus be of clinical value. Our model was tested, in terms of accuracy in density prediction, on two internal test sets independent from model fine-tuning and validation, test set.

2. METHODS

This section will discuss the breast phantoms used, the mammogram simulations, the preprocessing of the mammograms before putting them into the DL models, a DL model for determining the region of constant thickness (where the breast is in contact with the compression paddle), and the DL model for breast density estimation.

a. Patient-Based Breast Phantoms

Previously, a total of 88 digital breast phantoms were generated from as many patient images acquired with a dedicated breast CT system (Koning Corporation, Norcross, GA, USA). The images were acquired during an unrelated clinical trial aiming at the evaluation of breast computed tomography (CT) in a diagnostic setting. Each image was reconstructed using filtered back projection (Shepp-Logan kernel), and underwent automatic segmentation aimed at voxel-wise classification into four categories: air, adipose tissue, fibroglandular tissue, and skin [9]. These classified breast images were subsequently converted into finite element biomechanical models and underwent the simulation of mechanical compression using a previously developed computational method [10]. Compression was simulated and applied along the two standard directions acquired during a mammographic exam (cranio-caudal (CC), and medio-lateral oblique (MLO)). Each case was compressed to a different thickness, equivalent to the compressed breast thickness recorded in the image header of the CC- or MLO-view digital breast tomosynthesis (DBT) image of the corresponding patient breast.

As a result, 88 compressed breast phantoms were obtained for the CC and the MLO direction of compression, with isotropic voxel size equal to that of the acquired breast CT images (0.273 mm x 0.273 mm x 0.273 mm), with thickness equal to that obtained during the corresponding DBT examination. The compressed phantoms were augmented to increase the size of the dataset by increasing or decreasing the size of the breast by 10%, and by dilation (radius of 2 and 4 voxels) or erosion (radius of 1 voxel) of the fibroglandular tissue. These 5 augmentations resulted in a total of 528 compressed phantoms (6x88, per view), which were divided, randomly and on a patient level, into train (N = 360), validation (N = 60) and test sets (N = 108).

b. Mammogram Simulations

Mammographic images of all 528 phantoms were simulated in two main steps, ray-tracing and the calculation of a primary image. For the ray-tracing, a pre-developed and validated GPU-based Cone-Beam projector was used [11]. This projector takes the point source, detector geometry, and phantom as inputs and generates a thickness map of the implemented detector dimensions, for each of the four voxel classes (air, adipose tissue, fibroglandular tissue, and skin). Subsequently, the thickness maps for the different materials are used to simulate the air kerma (AK) projection using the discretized version of the polychromatic Lambert's law (Equation 1):

$$AK(x,y) = \sum_{e} e \cdot N_e \cdot \exp\left(-\sum_{m} \mu(e,m) \cdot T(m,x,y)\right) \cdot \left(\frac{\mu_{tr}}{\rho_{air}}\right)_e \quad (1)$$

where $AK(x,y)$ is the air kerma at each detector pixel (x,y), e is the current energy bin of the spectrum model, N_e is the photon fluence of the current energy bin, $\left(\frac{\mu_{tr}}{\rho_{air}}\right)_e$ is the mass-energy-transfer coefficient for the air for energy e to convert the photon energy fluence to air kerma, $\mu(e,m)$ is the attenuation coefficient of material m at energy e, and $T(m,x,y)$ is the thickness of material m for each detector pixel (x,y). To then obtain the primary image, $AK(x,y)$ was converted to digital units using linear scaling between the measured air kerma and the average background pixel value of the simulated mammography system. The spectrum model used was based on the work of Hernandez et al [12]. The simulations were performed using the geometry and acquisition settings of the Siemens Mammomat Inspiration (Forchheim, Germany) clinical system, with the tube voltage varying between 26 kV and 32 kV according to the breast thickness [13] and a detector pixel size of 0.085 mm. The simulations were then subsequently repeated, only for the test set, using the geometry and acquisition settings of a different mammographic system (Hologic Selenia Dimensions), with the tube voltage varying between 25 kV and 36 kV according to the breast thickness and a detector pixel size of 0.07 mm. For both simulated systems, the tube voltage was selected, for each phantom, according to those used by the automatic exposure control of the respective system.

c. Mammogram Preprocessing

All simulated mammograms were preprocessed before they were put into the DL model. First, the smallest rectangular region of interest (ROI) encompassing the entire breast was automatically selected for each simulated mammogram. The rest of the image was excluded to limit the background information in the mammograms. This operation resulted in each mammogram being cropped to a different dimension, according to the size of each projected breast. Therefore, all mammograms were subsequently resized to a dimension of 256x128 pixels (linear interpolation), and the new pixel size (ranging from 0.357 mm to 0.907 mm (CC) and 0.428 mm to 0.863 mm (MLO) for the horizontal direction and 0.228 mm to 1.236 mm (CC) and 0.232 mm to 1.284 mm (MLO) for the vertical direction) resulting from this resizing operation was calculated and saved. The mammogram was down sampled to these relatively small dimensions to make the framework faster. Since the density is a global breast characteristic, it should be predictable from lower resolution images adequately. Second, the remaining open field pixels in each mammogram were set to zero, and the breast pixels were normalized by the value of the pixel(s) assumed to contain only adipose tissue. This was automatically selected by choosing the pixel(s) with the highest value within the region of the mammogram with constant thickness (i.e., in full contact with the compression paddle and the breast support table), known exactly in simulation. This normalization was performed to standardize the pixel value range in each mammogram to a reference value, and thus correct for any potential bias in pixel values introduced by the x-ray spectrum being discretized into steps for given thickness ranges. Finally, a last normalization step was performed to ease the training of the developed DL model for density estimation. For this, the image values were inverted, normalized to the maximum pixel value present in the whole training set, and the range of pixel values was made broader by cubic scaling.

d. Automatic Detection of Adipose-only Pixel

To make our method fully automatic and able to work only with the simulated mammograms as input, two similar U-Nets (one for CC, and one for MLO) [14] were implemented to automatically segment the part of the mammogram with constant breast thickness, used to identify the first normalization factor (i.e., pixel containing only adipose tissue). The U-Nets were trained and fine-tuned using the simulated mammograms (input) and the respective binarized summed thickness maps (output) from the training and validation set, respectively.

The implemented U-Nets were 4 layers deep, with final sigmoid activation. Each layer in the down-sampling part consisted of two blocks of 2D convolution, batch normalization and ReLU followed by max pooling. In

the bottleneck, two blocks of 2D convolution, batch normalization and ReLU were used. Each block in the up-sampling part consisted of transposed convolution, concatenation with the corresponding downsampling block, and two blocks of convolution, batch normalization and ReLU. The U-Nets were trained for 25 epochs with a batch size of 8, with a learning rate of 0.0001 with a decay factor of 0.8 every 4 epochs. The Adam optimizer was used to minimize a pixel-wise binary cross entropy loss. All convolution kernels had size 3 and stride 1, all pooling layers had kernel size 2 and stride 2.

The U-Nets performance was quantified with the Dice similarity coefficient, with respect to the ground truth given by the summed thickness maps. After training the U-Nets, the results were eroded (radius of 3 pixels) before applying them to the test set to segment the region of the breast with constant thickness, from which the normalization factor was calculated as described in the previous section. The erosion was performed to ensure that the mask includes only fully compressed voxels, correcting for potential errors at the boundary that might have affected the identification of the adipose-only pixel.

e. Deep Learning Model for Density Estimation: Training, Validation, and Testing

The DL model developed to estimate breast density from the simulated mammograms is shown in Figure 1. The DL model consisted of two input streams (one for the CC view and one for the MLO view) concatenated at a later stage in the model. Each stream started with an average pooling block to reduce the input dimension, followed by five blocks of 2D convolutions with ReLU activations, each followed by a max pooling layer. The last four of these five blocks were implemented with batch normalization before each 2D convolution. After these blocks, another batch normalization was applied, and the output flattened. The stream ends with six dense layers with ReLU activations, with a dropout layer (0.2) inserted for regularization after the first block. All convolutions were performed with kernel size 3 and stride 1, all pooling was performed with kernel size 2 and stride 2.

Before the two streams were concatenated, three additional inputs were provided to each stream in the fully-connected part of the model, before the last two fully-connected layers: the pixel sizes, in both directions (x, y), of the mammogram after resizing, and the compressed breast thickness (scalar value, equal to the distance between compression paddle and support table). These extra input parameters were supplied to provide information about the mammogram resolution (pixel size), and anatomy and, implicitly, used spectrum (breast thickness). The last layer of the model contained one node with linear activation for continuous density predictions.

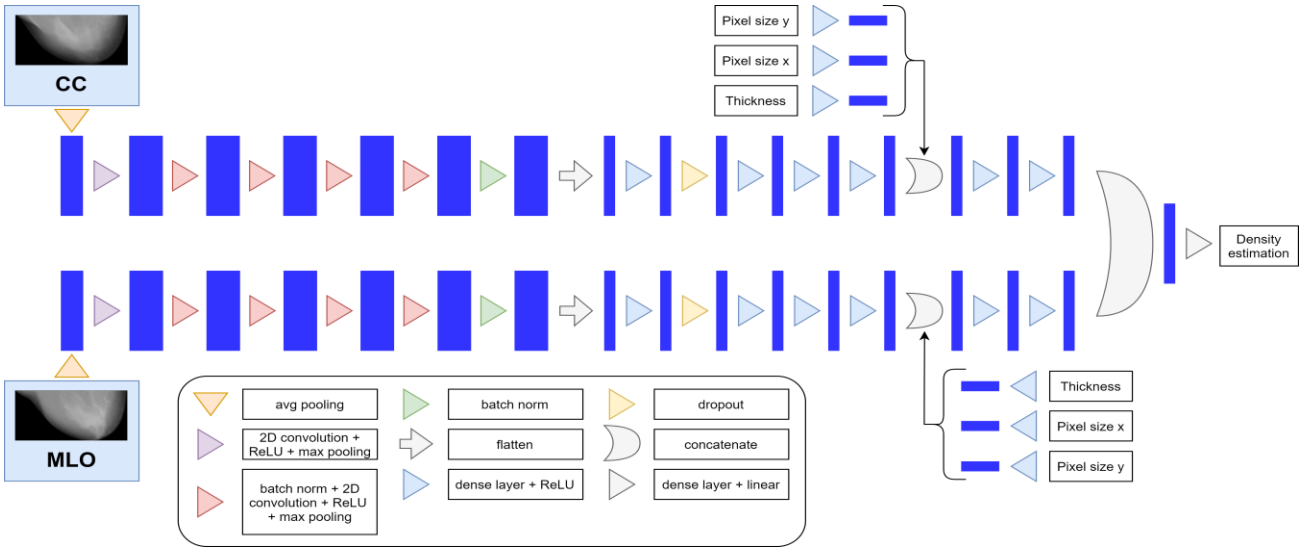

Figure 1: Schematic drawing of the DL model use for breast density estimation.

To train the model, the Adam optimizer was used with an L_2-loss. Learning rate decay was implemented, with an initial learning rate of 0.001, and a decay factor of 0.5 after every 20 epochs. The DL model was trained with a batch size of 32 for 150 epochs in total. Only the training mammograms obtained by replicating geometry and settings of the Siemens Mammomat system were used to train the model.

The results in density estimation on the test set mammograms were quantified using the median absolute error (mAE) and the corresponding interquartile range (IQR). These metrics were determined for four different density ranges: 1-15%, 15-25%, 25-60%, and >60%. Performance evaluation was performed on the mammograms kept for testing (and therefore not used for training or fine-tuning) simulated for both mammographic systems, without retraining the model (which was therefore only trained once for the Siemens Mammomat system).

3. RESULTS

The median breast density in the dataset of phantoms used in this study was 21.4%. The median breast thickness was 63.69 mm (range: 30.18-82.36 mm) for the CC cases and 59.48 mm (range: 26.91-81.74 mm) for the MLO cases.

The median Dice similarity of the two U-Nets for the segmentation of the region of the mammogram with constant breast thickness portion in the test set mammograms was 0.962 (IQR: 0.033) and 0.959 (IQR: 0.025) for CC and MLO, respectively.

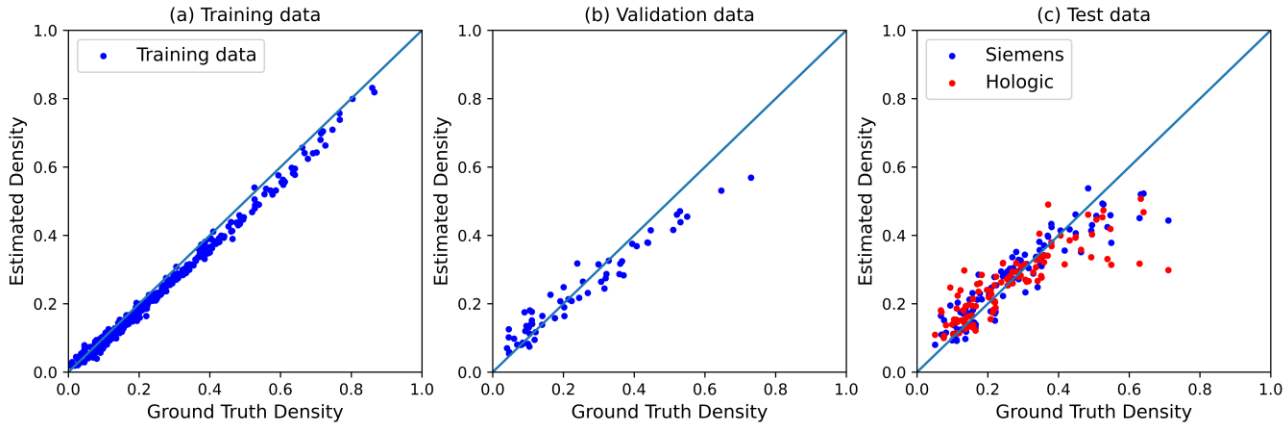

Figure 2: Estimated vs. actual density for the training, validation, and test set.

The estimated breast density for each sample in the training, validation, and test sets are plotted in Figure 2. For low breast densities, the same trend is present in all three sets. For densities higher than 50%, the model seems to suffer of a negative bias (Figure 2, panel c). This is explainable by the under-expression of cases with density higher than 50% in our dataset (10.0%), and the increased possibility of the absence of an adipose-only pixel for normalization.

Table 1: Median Absolute Error and interquartile range for the test set, for the four density ranges.

Density range (%)	**1-15**	**15-25**	**25-60**	**>60**
Median Absolute Error (IQR) for Siemens system (%-point)	3.3 (IQR: 3.5)	3.4 (IQR: 2.5)	3.5 (IQR: 3.9)	14.8 (IQR: 8.4)
Median Absolute Error (IQR) for Hologic system (%-point)	4.5 (IQR: 4.3)	3.5 (IQR: 3.6)	3.4 (IQR: 5.1)	24.2 (IQR: 17.6)

The test set mAE in density estimation was 3.6 percentage point (interquartile range: 3.4), and 3.7 (interquartile range: 5.3), respectively for the mammograms simulated with the Siemens and the Hologic system. Results for different density ranges, for both simulated systems, are reported in Table 1.

4. NEW OR BREAKTHROUGH WORK TO BE PRESENTED

To the best of our knowledge, an end-to-end framework for breast density estimation from simulated CC and MLO-view mammograms simultaneously, as developed in our study, has not been previously proposed.

5. CONCLUSIONS

In this work we presented a DL-based method to estimate breast density from simulated mammographic exams. The proposed method is fully automatic and can estimate breast density from simulated mammograms of different vendors accurately, without retraining. Unlike most previous studies [3]–[8], our approach was validated in terms of accuracy in breast density estimation against an objective ground truth, which was made possible thanks to using patient-based phantoms with known density obtained with breast CT. The difficulty of correctly estimating the density in cases with high breast density is probably due to two reasons. The first is the relatively few cases at high densities included in the dataset. The second is the higher possibility of the absence of an adipose-only pixel at higher breast densities. Of course, this study should be considered preliminary, due to limited dataset size, and the simplified simulations performed (only primary, non-scattered, x-rays and an ideal detector). However, due to the relatively large pixel size after resizing, the influence of noise and resolution loss should be low. Furthermore, the use of an anti-scatter grid in mammography should also limit the influence of scatter (or the lack thereof) on our results. Future work includes training and testing using a larger dataset, and the inclusion of the other factors contributing to the image generation in mammography simulations. It is desired to get this method to work on processed (i.e., "for presentation") simulated mammograms however, as these processes are usually black boxes and vendor specific, this will be a difficult task and vendor cooperation will probably be needed. Ultimately, evaluation on real patient data is needed to prove if this method is accurate enough to be useful in the clinic.

This work has not been previously submitted for presentation or publication elsewhere.

REFERENCES

[1] V. A. Mccormack, I. Dos, and S. Silva, "Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk: A Meta-analysis," 2006, doi: 10.1158/1055-9965.EPI-06-0034.

[2] E. F. Conant, B. L. Sprague, and D. Kontos, "Beyond BI-RADS density: A call for quantification in the breast imaging clinic," *Radiology*, vol. 286, no. 2, pp. 401–404, Feb. 2018, doi: 10.1148/RADIOL.2017170644.

[3] O. Haji Maghsoudi *et al.*, "Deep-LIBRA: An artificial-intelligence method for robust quantification of breast density with independent validation in breast cancer risk assessment," *Med. Image Anal.*, vol. 73, Oct. 2021, doi: 10.1016/J.MEDIA.2021.102138.

[4] A. Gastounioti *et al.*, "Evaluation of LIBRA software for fully automated mammographic density assessment in breast cancer risk prediction," *Radiology*, vol. 296, no. 1, pp. 24–31, Jul. 2020, doi: 10.1148/RADIOL.2020192509/SUPPL_FILE/RY192509SUPPF4.JPG.

[5] M. Kallenberg *et al.*, "Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1322–1331, May 2016, doi: 10.1109/TMI.2016.2532122.

[6] J. Lee and R. M. Nishikawa, "Automated mammographic breast density estimation using a fully convolutional network:," *Med. Phys.*, vol. 45, no. 3, pp. 1178–1190, Mar. 2018, doi: 10.1002/MP.12763.

[7] C. D. Lehman *et al.*, "Mammographic breast density assessment using deep learning: Clinical implementation," *Radiology*, vol. 290, no. 1, pp. 52–58, Jan. 2019, doi: 10.1148/RADIOL.2018180694/SUPPL_FILE/RY180694SUPPA1.PDF.

[8] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, "A deep learning method for classifying mammographic breast density categories," *Med. Phys.*, vol. 45, no. 1, pp. 314–321, Jan. 2018, doi: 10.1002/MP.12683.

[9] M. Caballo, C. Fedon, L. Brombal, R. Mann, R. Longo, and I. Sechopoulos, "Development of 3D patient-based super-resolution digital breast phantoms using machine learning," *Phys. Med. Biol.*, vol. 63, no. 22, p. 225017, Nov. 2018, doi: 10.1088/1361-6560/AAE78D.

[10] E. García *et al.*, "Realistic compressed breast phantoms for medical physics applications," *Proc. SPIE 11513*, vol. 11513, no. 22, pp. 30–37, May 2020, doi: 10.1117/12.2564273.

[11] N. Moriakov, J.-J. Sonke, and J. Teuwen, “LIRE: Learned Invertible Reconstruction for Cone Beam CT,” *arXiv*, 2022, [Online]. Available: http://arxiv.org/abs/2205.07358.
[12] A. M. Hernandez and J. M. Boone, “Tungsten anode spectral model using interpolating cubic splines: Unfiltered x-ray spectra from 20 kV to 640 kV,” *Med. Phys.*, vol. 41, no. 4, 2014, doi: 10.1118/1.4866216.
[13] M. Caballo *et al.*, “Patient-derived heterogeneous breast phantoms for advanced dosimetry in mammography and tomosynthesis,” *Med. Phys.*, 2022, doi: 10.1002/MP.15785.
[14] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *Navab, N., Hornegger, J., Wells, W., Frangi, A. Med. Image Comput. Comput. Interv. – MICCAI 2015. MICCAI 2015. Lect. Notes Comput. Sci.*, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28/COVER.