Perez-Benito, FJ.; Signol, F.; Perez-Cortes, J.; Fuster Bagetto, A.; Pollan, M.; Pérez-Gómez, B.; Salas-Trejo, D.... (2020). A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation. Computer Methods and Programs in Biomedicine. 195:123-132. https://doi.org/10.1016/j.cmpb.2020.105668

# A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation

Francisco Javier Pérez-Benito[a,*], François Signol[a], Juan-Carlos Perez-Cortes[a], Alejandro Fuster-Baggetto[a], Marina Pollán[b,c], Beatriz Pérez-Gómez[b,c], Dolores Salas-Trejo[d,e], María Casals[d,e], Inmaculada Martínez[d,e], Rafael LLobet[a]

[a] *Instituto Tecnológico de la Informática, Universitat Politècnica de València, Camino de Vera, s/n, 46022 València, Spain*
[b] *National Center for Epidemiology, Carlos III Institute of Health, Monforte de lemos, 5, 28029 Madrid, Spain*
[c] *Consortium for Biomedical Research in Epidemiology and Public Health (CIBER en Epidemiología y Salud Pública - CIBERESP), Carlos III Institute of Health, Monforte de Lemos, 5, 28029 Madrid, Spain*
[d] *Valencian Breast Cancer Screening Program, General Directorate of Public Health, València, Spain*
[e] *Centro Superior de Investigación en Salud Pública CSISP, FISABIO, València, Spain*

## Abstract

**Background and Objective:** Breast cancer is the most frequent cancer in women. The Spanish healthcare network established population-based screening programs in all Autonomous Communities, where mammograms of asymptomatic women are taken with early diagnosis purposes. Breast density

*Abbreviations:*

*Corresponding author

*Email addresses:* fjperez@iti.es (Francisco Javier Pérez-Benito), fsignol@iti.es (François Signol), jcperez@iti.upv.es (Juan-Carlos Perez-Cortes), afuster@iti.es (Alejandro Fuster-Baggetto), mpollan@isciii.es (Marina Pollán), bperez@isciii.es (Beatriz Pérez-Gómez), salas_dol@gva.es (Dolores Salas-Trejo), casals_mar@gva.es (María Casals), martinez_inm@gva.es (Inmaculada Martínez), rllobet@iti.upv.es (Rafael LLobet)

assessed from digital mammograms is a biomarker known to be related to a higher risk to develop breast cancer.

It is thus crucial to provide a reliable method to measure breast density from mammograms. Furthermore the complete automation of this segmentation process is becoming fundamental as the amount of mammograms increases every day. Important challenges are related with the differences in images from different devices and the lack of an objective gold standard.

This paper presents a fully automated framework based on deep learning to estimate the breast density. The framework covers breast detection, pectoral muscle exclusion, and fibroglandular tissue segmentation.

**Methods:** A multi-center study, composed of 1785 women whose "for presentation" mammograms were segmented by two experienced radiologists. A total of 4992 of the 6680 mammograms were used as training corpus and the remaining (1688) formed the test corpus. This paper presents a histogram normalization step that smoothed the difference between acquisition, a regression architecture that learned segmentation parameters as intrinsic image features and a loss function based on the DICE score.

**Results:** The results obtained indicate that the level of concordance (DICE score) reached by the two radiologists (0.77) was also achieved by the automated framework when it was compared to the closest breast segmentation from the radiologists. For the acquired with the highest quality device, the DICE score per acquisition device reached 0.84, while the concordance between radiologists was 0.76.

**Conclusions:** An automatic breast density estimator based on deep learning exhibits similar performance when compared with two experienced

radiologists. It suggests that this system could be used to support radiologists to ease its work.
*Keywords:*
Breast density, Entirely Convolutional Neural Network (ECNN), Deep Learning, Dense tissue segmentation, Mammography

---

## 1. Background

Mammographic screening is a highly standardized procedure for breast cancer early detection programs, and the acquired mammograms are interpreted by specialized radiologists who batch read up to 50 mammographies per hour [1]. Full Field Digital Mammography (FFDM) is still one of the preferred methods for breast cancer screening programs. Technology innovations provide better imaging features that promote earlier diagnosis of breast cancer.

Percent Density (PD) which measures the percentage of fibroglandular tissue over the total breast, is known to be a marker of breast cancer development risk [2, 3]. The American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) has also reported a breast classification, based on density, shape, and granularity of the dense tissue [4], suggesting that not only the total amount but also its distribution matters [5, 6]. Besides, one of the principal problems in PD assessment is the inter and intra-observer variability [7–10].

In this sense, an automated tool exhibiting a high agreement with several radiologists could serve as one of the first steps in standardizing the read of breast density. Authors of [11] emphasize a human-like automatic tool could

be used as fully independent second reader of screening mammograms, where double reading is standard. A second human reader would only arbitrate discrepancies between the first human reader and the system, halving the workload for any screening program where double reading is standard.

Coupled with this are the tremendous opportunities and challenges for research which are brought by healthcare systems [12], in particular, breast screening programs. To manage and model this huge amount of data, the paradigm of Deep Learning (DL) has emerged. The abstraction ability of DL [13] has demonstrated promising results from speech recognition [14, 15], reconstructing brain circuits [16, 17] or predicting the effects of DNA mutations [18, 19] to medical imaging tasks [20, 21].

One of the most widespread paradigms used in computer vision problems solved via DL take advantage of Convolutional Neural Networks (CNN) [22]. It is based on the extraction of features that are of higher-order as the images go through more layers. CNNs are nowadays the state-of-the-art for many recognition and detection tasks [23–25].

The current work presents a fully automated framework for dense tissue segmentation. It includes breast detection, pectoral muscle exclusion and dense tissue segmentation. Among the contributions of this work, we can highlight (1) a preprocessing algorithm dealing with the variability of mammograms acquired from different devices in the training stage, (2) a new regression architecture Entirely CNN (ECNN), whose output are two parameters used as intrinsic segmentation features, improves classical CNN network (3) a loss function which maximizes the DICE score [26] by continuously rebuilding a probabilistic dense tissue mask, and finally, (4) the ability

to manually modify the segmentation using the DMScan software [27, 28].

## 2. Methods

### *2.1. Dataset and participants*

A multi-center study covered women from 11 hospitals of the *Comunitat Valenciana* which belong to the Spanish breast cancer screening network. The prior design of the study was a 1:1 case-control to find factors influencing the development of breast cancer. In this sense, a representation of the whole PD spectrum is assured.

The current study contains a total of 1785 women with ages from 45 to 70. For each patient who developed cancer, if available, the contralateral mammogram was taken from the screening visit previous to diagnostic, otherwise, the contralateral mammogram to the one diagnosed with cancer from the most recent screening visit was selected. Finally, if no previous mammogram existed, then the contralateral mammogram at the diagnostic time was extracted. Since in Spain "raw" mammograms are not routinely stored, all the mammograms are of the type "for presentation".

In 10 of the 11 facilities, the cranio-caudal (CC) and medio lateral-oblique (MLO) views were recruited for each woman, meanwhile, the other facility only collected the CC view. A brief summary of data from the different mammography facilities can be found in Table 1.

| Id | Unit | Mammography device | Number of women | Number of mammograms (Number of reads) |
|---|---|---|---|---|
| 01 | Castellón | FUJIFILM | 191 | 382(764) |
| 02 | Fuente de San Luis | FUJIFILM | 190 | 380(760) |
| 04 | Alcoi | IMS s.r.l. / Giotto IRE (*) | 66 | 132(264) |
| 05 | Xàtiva | FUJIFILM | 159 | 318(636) |
| 07 | Requena | HOLOGIC / Giotto IRE (*) | 28 | 56(112) |
| 10 | Elda | SIEMENS / Giotto IRE (*) | 311 | 622(1244) |
| 11 | Elche | FUJIFILM | 278 | 556(1112) |
| 13 | Orihuela | FUJIFILM | 117 | 234(468) |
| 18 | Denia | IMS s.r.l. / Giotto IRE(*) | 38 | 76(152) |
| 20 | Serrería | (**) | 177 | 354(708) |
| 99 | Burjassot | Senographe 2000D | 230 | 230(460) |
| | | **Total** | 1785 | 3340(6680) |

Table 1: Screening units, their mammography devices and the number of women and mammograms per device. (*) Implies the use of a new device [Gioto IRE] since 2015. (**) The device is not known.

Mammograms were analyzed by two experienced radiologists using DM-Scan [27, 28]. This software provides assisted semiautomatic tools to segment the breast and the fibroglandular tissue and to exclude undesired regions such as pectoral muscle or armpit.

### *2.2. Breast segmentation framework*

The segmentation pipeline is composed of a first step covering breast detection and pectoral muscle exclusion, a second step to normalize the histogram variability between acquisition devices, and then, the dense tissue parametric segmentation is carried out using a deep learning model that was trained using an ad-hoc loss function. Details on each of the aforementioned steps are given below.

*2.2.1. Background and breast detection*

We have used a heuristic, iterative algorithm based on connected components to obtain the gray level threshold that distinguishes breast from background. Even though there exist some issues concerning the use of connected components labeling on binary images [29], homogeneous breast shape makes this kind of algorithms suitable to be used for breast segmentation and exhibits perfect breast detection.

The first step of our approach is to assess the histogram of the image. Based on the premise that the most frequent pixel value has to belong to the background, a range of possible breast thresholds is defined.

Then, this range of thresholds is covered until only two homogeneous components are detected. The first step is to assure that the breast is left-oriented and to binarize the image using the first possible threshold, then apply the connected component labeling method. We chose the Scan plus Array-based Union-Find (SAUF) algorithm [30]. Finally, if only two components are obtained, the threshold is set if not, it is continued covering the range of thresholds.

*2.2.2. Armpit and pectoral muscle exclusion*

Several approaches have been proposed in the literature for armpit and pectoral muscle recognition and exclusion. The authors of [31] proposed a method based on homogeneous contours; the work presented in [32] proposed a combination of image processing, genetic algorithm, morphological selection, and polynomial curve fitting. The approach explained in [33] combines fractional differential enhancement methods with iterative thresholding algorithms meanwhile the authors of [34] propose the use of the outputs of three

existing algorithms (region growing, thresholding and $k$-means clustering) as the input of a machine learning-based computer-aided decision system.

The common key observed in all the aforementioned studies is the knowledge that pectoral muscle appears in a triangle of one of the top corners of the image. Based on this premise, we have defined a robust procedure to exclude pectoral muscles founded on negative gradient changes.

After assuring the image is left-oriented, we applied a Gaussian filter and a 50-pixel moving average to smooth edges and remove spurious isolated brightness pixels. As the muscle is a well contrasted border, it tends to be the last remaining after the smoothing process. We iteratively built a polygon that encloses the exclusion area by selecting the pixel with the lowest gradient every 50 rows until the column of the selected pixel was enough close to the left image border. Finally, the vertex that closed the polygon was the first pixel from the top left corner.

*2.2.3. Normalizing variability between acquisition devices*

The pixel size, grey-scale bit resolution, signal to noise ratio or detective quantum efficiency are important concepts related to image quality [35]. The different mammogram acquisition devices show a huge variability in the quality of mammograms. The first experiments carried out produced different performance results depending on the mammography facility. These results influenced the variability assessment among different devices and how it can negatively impact the training of a machine learning model. We evaluated the differences among the histograms of mammograms over the different mammography facilities by applying the framework proposed by Sáez et al. [36, 37] at image level and checking that well-differentiated mammography

facility-clusters appeared as can be seen in Figure 1a, where the images from medical centers using different devices were extracted.

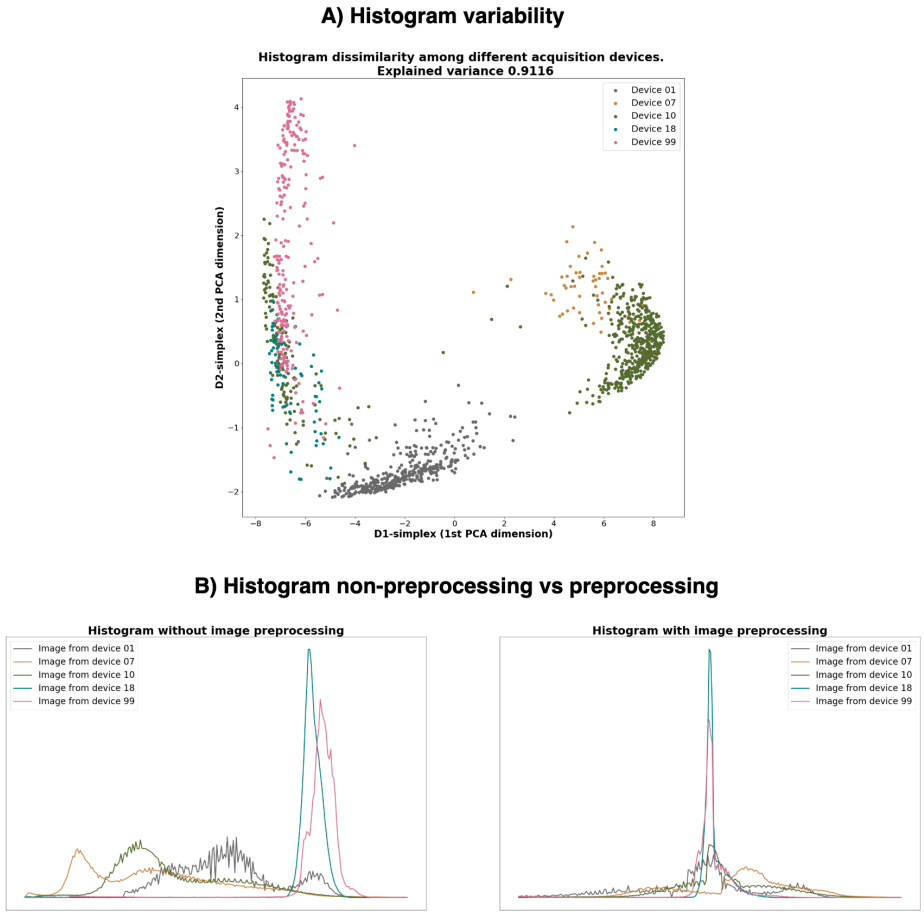


Figure 1: a. Differences among the histograms of the mammograms of the facilities with different acquisition devices. Well-differentiated clusters demonstrated the dissimilarity between acquisition devices. b. Example of histogram transformation using one mammogram from each of the different mammography facilities.

Mammogram features like resolution or signal to noise ratio depend on the electronic components of acquisition devices and produce a specific signature

visible on the image histogram. In this work, we propose a way to standardize them, which leads to better performance when a model using the images of the whole set of the mammography facilities is trained, avoiding the need of a specific model for each acquisition device.

The preprocessing steps proposed are the following, and the comparison of two histograms from two different acquisition devices can be found in Figure 1b):

1. Normalize the pixel values of the image between $[0, 1]$.
2. Shift histogram to set the minimum breast tissue pixel to 0.
3. Normalize again the pixel values between $[0, 1]$.
4. Standardize the breast pixel values to a normal distribution $Z \sim N(0, 1)$.
5. Adjust the pixel values so that the mode is 0.
6. Under the assumption that most typical percent density values are below 30% (above 70th percentile) and values under the 30th percentile only belong to fatty tissue, apply a linear stretching from percentile 30 to $-1$ and from percentile 70 to 1.
7. Apply once more a normalization to ensure inputs for the Deep Neural Network are between $[0, 1]$.

#### *2.2.4. Dense tissue segmentation with Entirely Convolutional Neural Network (ECNN)*

Recent works address dense tissue segmentation from different points of view. Authors of [38] used a fractal inspired approach and a multiresolution stack representation to extract 3D histogram features, which were used to apply *k-means* [39] to classify each pixel as fatty, semi-fatty, semi-dense or dense.

Another interesting approach is that proposed in [20], in which an unsupervised step to extract features, based on a sparse autoencoder, is followed by a supervised classifier which tried to classify each pixel as pectoral muscle, fatty or dense tissue. Close to this approach is the one of [40] that uses 4 fully convolutional networks, two to segment breast tissue on CC and MLO views and the other two to segment the dense tissue on those same views.

Since an accurate and objective *gold standard* does not exist for the segmentation task, the ground-truth of the model to be trained is the segmentation provided by two experienced radiologists who used a semi-atuomatic segmentation tool. Usually, these tools are based on the selection of two thresholds $th_B$ and $th_F$ to segment, respectively, the breast and the fibroglandular tissue. In our study we have used DMScan, a semi-automatic tool that provides a more accurate segmentation using a third parameter $\alpha$ explained below. Therefore, this tool interactively rebuilds a dense tissue mask using the values of three parameters.

- The breast region threshold ($th_B$). Pixels with values higher than $th_B$ are considered to belong to the breast.

- The brightness corrector $\alpha$. The X-ray attenuation depends on the thickness of the breast. The thicker the tissue irradiated, the greater the attenuation and, consequently, the brighter the image [27]. The first parameter is related to a brightness correction coefficient $k_{ij}$ by which each pixel is multiplied. The user-defined parameter $\alpha \in [0, 1]$ updates the $k_{ij}$ according to Equation 1 where $d_{ij}$ is the horizontal distance of the pixel $(i, j)$ to the image border or the pectoral muscle.

It compensates the variation of thickness along the breast.

$$k_{ij} = \alpha + 2(1 - \alpha)d_{ij} \tag{1}$$

- The fibroglandular tissue threshold ($th_F$). Pixels with values higher than $th_F$ are considered to belong to the dense tissue.

We propose an architecture in which convolutions were employed to extract the features needed to replicate the DMScan segmentation as image-intrinsic features: $\alpha$ and $th_F$. A similar architecture could be applicable to meet the requirements of other semi-automatic threshold-based tools. From now on, we will refer to this architecture as Entirely Convolutional Neural Network (ECNN). It was designed to work with $256 \times 256$ px sized images. The proposed architecture and its convolutional layers configuration are shown in Figure 2.

Besides, the activation function for the layers was the *Leaky Rectified Linear Unit (ReLU)*, with exception of the last layer which was set to *sigmoid* function to ensure output was $[0, 1]$-bounded. The activation functions are presented in Equation 2.

$$ReLU(x) = \begin{cases} x & if x > 0 \\ 0.2x & otherwise \end{cases} \tag{2}$$

$$sigmoid(x) = \frac{1}{1+e^{-x}}$$

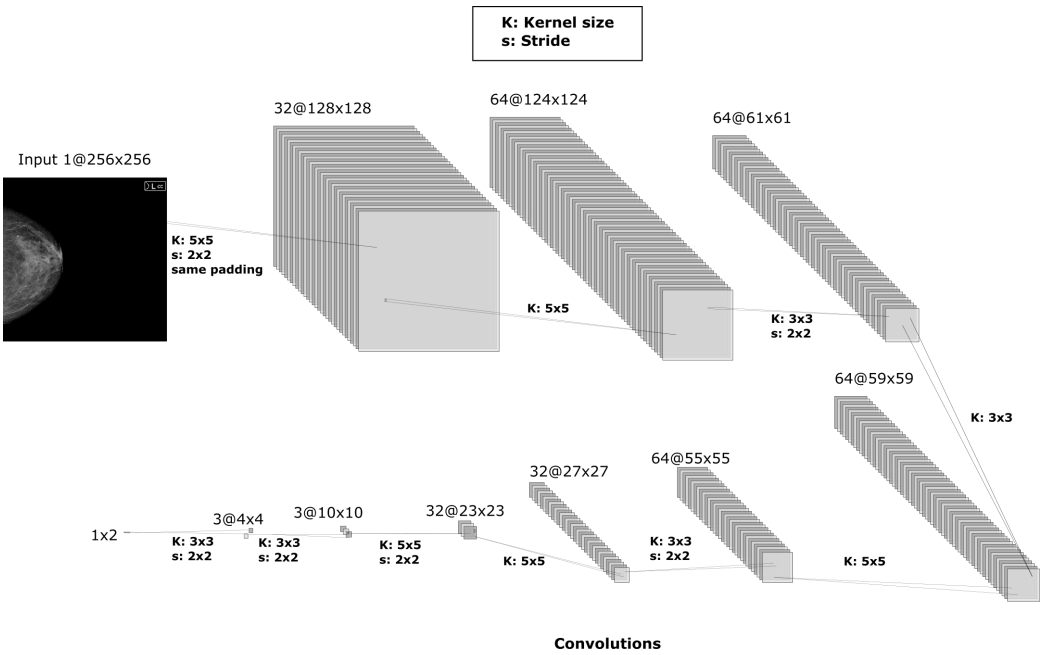


Figure 2: **Entirely Convolutional Neural Network (ECNN) architecture.** The kernel and the strides size for each layer are shown, padding was added to the first convolution to preserve information on the borders. Only convolutions are used to extract the features ($\alpha$ and $th_F$) needed to segment the dense tissue.

### *2.2.5. Continuous parameter-based DICE loss function*

To measure the performance of our model, we chose the widespread used Sørensen-Dice Similarity Coefficient [26] which measures how much two masks $M_1$ and $M_2$ overlap according to equation 3.

$$DICE(M_1, M_2) = \frac{2|M_1 \cap M_2|}{|M_1| + |M_2|} \tag{3}$$

The use of mean squared error is not monotonically related to the DICE score, leading to an erratic convergence on the learning stage. Furthermore, DICE is the function we want to maximize as it measures the agreement

between binary masks. Maximizing DICE is equivalent to minimizing $1-$ DICE. Given two masks $M_1$ and $M_2$, a DICE of $\frac{2}{3} = 0.66$ means that the number of pixels belonging to $M_1$ and $M_2$ is equal to the number of pixels that only belong to one of them. A DICE score of 0.8 implies that the number of pixels belonging to only one of the masks half the number of pixels that belong to both masks.

This was the reason to develop our metric based on DICE to be used as a loss function in the training stage. The underlying key is to build a map of probabilities in which each element represents the probability of the corresponding pixel belonging to dense tissue and, then, apply the DICE score between estimated mask and the dense tissue mask provided by the radiologists (ground truth). The metric can be represented according to Equation 4:

$$
\begin{aligned}
&\mathbb{R}^{[0,1]}_{256\times256} \times \mathbb{R}^{[0,1]} \xrightarrow{fil} \mathbb{R}^{[0,1]}_{256\times256} \times \mathbb{R}^{[0,1]} \xrightarrow{logistic} \mathbb{R}^{[0,1]}_{256\times256} \times \mathbb{R}^{\{0,1\}}_{256\times256} \xrightarrow{loss} \mathbb{R}^{[0,1]} \\
&fil\Big((m_{ij}), \hat{\alpha}\Big) \longmapsto \Big( [\hat{\alpha} + 2(1-\hat{\alpha})d_{ij}]\, m_{ij} \Big) \\
&\qquad\qquad logistic\Big((m_{ij}), \hat{th_F}\Big) \longmapsto \left( \frac{1}{e^{-(40[m_{ij} - \hat{th_F}])}} \right) \\
&\qquad\qquad\qquad\qquad loss\Big((m_{ij}), (n_{ij})\Big) \longmapsto 2\frac{\sum m_{ij} n_{ij}}{\sum m_{ij} + \sum n_{ij}}
\end{aligned}
\tag{4}
$$

Where $m_{ij} \in \mathbb{R}^{[0,1]}_{256\times256}$ is the mammography resized to $256 \times 256$ and $n_{ij} \in \mathbb{R}^{\{0,1\}}_{256\times256}$ is the dense tissue mask provided by an specialist. It is worth to mention that in $fil(.)$, $d_{ij}$ is the one defined in Section 2.2.4. The logistic function $logistic(.)$ was applied instead of a *step function* to maintain the continuity, and 40 was used as a slope factor to assure a quick transition

between 0 and 1.

Finally, the loss function, which from now on will be referred to as Continuous based Parameters DICE loss score (CPDICE) is defined according to Equation 5:

$$CPDICE\Big((m_ij), \hat{\alpha}, \hat{th_F}, (n_ij)\Big) = 1 - 2\frac{\sum \left(1 + e^{-40([\hat{\alpha}+2(1-\hat{\alpha})d_{ij}]m_{ij} - \hat{th_F})}\right)^{-1} n_{ij}}{\sum \left(1 + e^{-40([\hat{\alpha}+2(1-\hat{\alpha})d_{ij}]m_{ij} - \hat{th_F})}\right)^{-1} + \sum n_{ij}} \quad (5)$$

The corpus, consisting of a total of 3340 mammograms and segmented using DMScan by two radiologists (6680 reads), was randomly stratified taking 75% (4992 segmentations) as training set, from which 10% of the segmentations were extracted with validation purposes (*validation set*), and the remaining 25% (1688 segmentations) as test set. Both mammogram reads of the same image were always included in the same set. The maximum number of epochs was fixed to 500, the optimizer for the training stage was the Adam algorithm [41], and finally, the learning rate was set to 0.001.

*2.2.6. Dense tissue segmentation example*

Three examples of ECNN segmentation of test images using the steps previously described can be found in Figure 3. The segmentation is compared to those proposed by the two radiologists. The mammograms were recruited using different acquisition devices. The last example shows the emergence of the abdomen that is still not covered by our pipeline and may negatively influence performance results.

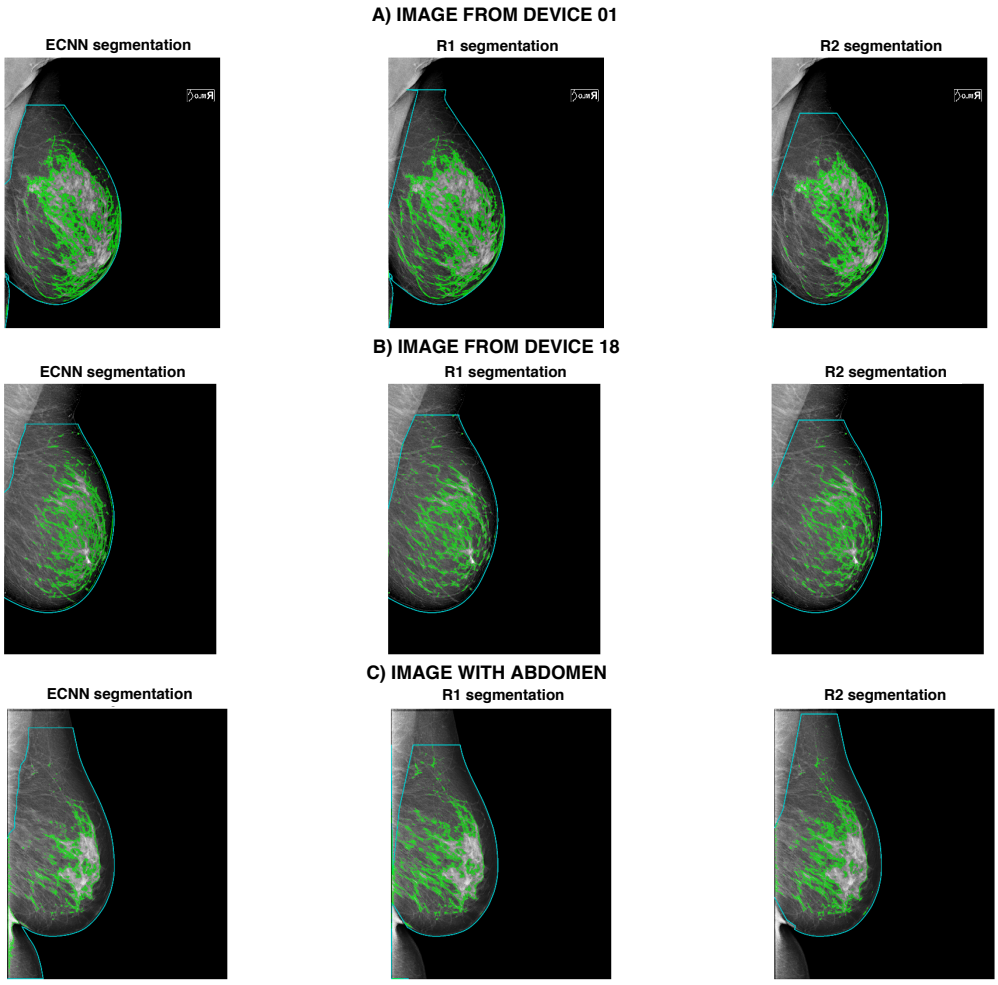


Figure 3: **ECNN segmentation compared to radiologists segmentations on different devices.** a. Segmentation of a mammogram acquired using the device of mammography facility 01. b. Segmentation of a mammogram acquired using the device of mammography facility 18. c. A mammogram from mammography facility 11 where abdomen tissue is found. Medio-lateral oblique mammograms were selected so the exclusion of the pectoral muscle could be seen, however, the abdomen is not excluded.

## 3. Results

As previously mentioned, our model was configured to be trained at most 500 epochs. The lowest loss error obtained was around epoch 400 and the final selected model was then obtained after this number of training iterations.

The lack of a real gold-standard, along with the inter-reader variability [11, 42] motivated us to train our ECNN using segmentations of more than one radiologist as explained before. This decision was made because we did not want a model behaving like a specific specialist, but we wanted a model that could obtain a level of agreement with any of the specialists comparable to the agreement among them. It is important to note that the segmentation of each radiologist is considered as an independent element. In this sense, if the model gets a perfect segmentation for a mammogram compared to a specific radiologist (R1 for instance), the segmentation of the same mammogram gives a difference concerning the other radiologist (R2) of exactly the difference between R1 and R2. This implies the existence of an unavoidable intrinsic error which has an impact on the performance of the model. It is also worth to mention that radiologists segmentations were labeled using DMScan, which provides an interactive tool to exclude the armpit and pectoral muscle. As can be seen in Figure 3, the approach implemented in the current study does not manage, for example, the presence of the abdomen tissue at the bottom of the image. This may also lead to an additional increase of the errors reported in this study.

### *3.1. ECNN as an alternative architecture to standard CNN*

As previously mentioned, one of the requirements of the present study is to learn the same parameters that the radiologist has access to. The use of approaches where each pixel or each local region could be freely assigned as dense or not dense was discarded due to the interest in comparing our results with those obtained using widely used threshold-based semi-automatic tools.

Then, to measure the performance of the proposed architecture -ECNN- we trained a fully connected convolutional neural network (CNN) to estimate the desired parameters. A typical architecture for similar tasks [43] composed of a convolutional part followed by a three dense layers (see Table 2 for architecture details) provided the intended parameter estimation. It was trained using the CPDICE as a loss function with a learning rate of 0.001.

| **Layer number** | **Type layer** | **Filters/Neurons** | **Kernel size** | **Strides** | **Padding** | **Activation function** |
|---|---|---|---|---|---|---|
| 1 | Convolutional | 32 | $3 \times 3$ | $1 \times 1$ | *same* | Leaky ReLu |
| 2 | Convolutional | 64 | $3 \times 3$ | $1 \times 1$ | *valid* | Leaky ReLu |
| 3 | Maxpooling | - | $2 \times 2$ | $2 \times 2$ | *valid* | - |
| 4 | Convolutional | 64 | $3 \times 3$ | $1 \times 1$ | *valid* | Leaky ReLu |
| 5 | Convolutional | 64 | $3 \times 3$ | $1 \times 1$ | *valid* | Leaky ReLu |
| 6 | Maxpooling | - | $2 \times 2$ | $2 \times 2$ | *valid* | - |
| 7 | Dense | 512 | - | - | - | Leaky ReLu |
| 8 | Dense | 512 | - | - | - | Leaky ReLu |
| 9 | Dense | 2 | - | - | - | Sigmoid |

Table 2: The details of CNN layers implementation. The first six layers extract image features (convolution stage) and the last three layers play the role of the regressor.

The results per mammography facility compared to those obtained with the ECNN are presented in Table 3.

| | ECNN | | CNN | | R1 vs R2 | |
|---|---|---|---|---|---|---|
| **mammography facility** | **DICE** | **CI** | **DICE** | **CI** | **DICE** | **CI** |
| **01** | **0.81** | [0.78, 0.84] | 0.79 | [0.76, 0.83] | 0.79 | [0.76, 0.83] |
| **02** | **0.83** | [0.79, 0.86] | 0.79 | [0.75, 0.83] | 0.79 | [0.76, 0.82] |
| **04** | 0.57 | [0.50, 0.65] | 0.60 | [0.53, 0.68] | **0.75** | [0.69, 0.81] |
| **05** | **0.84** | [0.81, 0.87] | 0.83 | [0.80, 0.86] | 0.65 | [0.61, 0.68] |
| **07** | 0.85 | [0.77, 0.94] | 0.81 | [0.69, 0.92] | **0.88** | [0.81, 0.96] |
| **10** | 0.68 | [0.65, 0.72] | 0.71 | [0.67, 0.75] | **0.77** | [0.75, 0.80] |
| **11** | **0.87** | [0.85, 0.88] | 0.83 | [0.81, 0.85] | 0.82 | [0.80, 0.84] |
| **13** | **0.86** | [0.83, 0.89] | 0.83 | [0.80, 0.87] | 0.78 | [0.75, 0.82] |
| **18** | 0.51 | [0.40, 0.64] | 0.56 | [0.46, 0.66] | **0.74** | [0.68, 0.79] |
| **20** | 0.61 | [0.55, 0.67] | 0.62 | [0.57, 0.67] | **0.78** | [0.75, 0.81] |
| **99** | 0.78 | [0.73, 0.83] | 0.75 | [0.69, 0.81] | **0.79** | [0.76, 0.82] |
| **Total** | **0.77** | [0.75, 0.78] | 0.76 | [0.74, 0.77] | 0.77 | [0.75, 0.78] |

Table 3: ECNN results compared to conventional convolutional architecture. CI refers to 95% confidence interval. ECNN outperforms in many of the devices the agreement between R1 and R2. CNN got better scores on some mammography facilities in which the quality of the mammogram is lower. The DICE scores for the DL models represent the DICE scores to the closer radiologist segmentation.

The conventional convolutional architecture only got significantly better results on mammography facilities 04 and 18. These mammography facilities correspond to the device with the lowest gray-level resolution. The DICE scores in these facilities show also poor agreement between radiologists. Although the best performance of ECNN compared to CNN only can be considered statistically significant for device 11, this approach provided, at least, a similar performance, and it is also faster, more interpretable, and has a lower computational load.

*3.2. ECNN improvement in function with training epochs*

Figure 4 shows the model assessment of test images at different epochs (10, 50, 100, 200, 220, 400 and 460) to make clear the achieved balance at different mammography facilities. Averaged-score of validation set also reported its best punctuation at epoch 400 when the validation set score monitored during the training stage.

According to these results, there exist mammography facilities in which the proposed model performance is significantly worse than the obtained in others. It is related to the acquisition device, the quality of acquired images, and probably the unbalanced number of images among different devices.

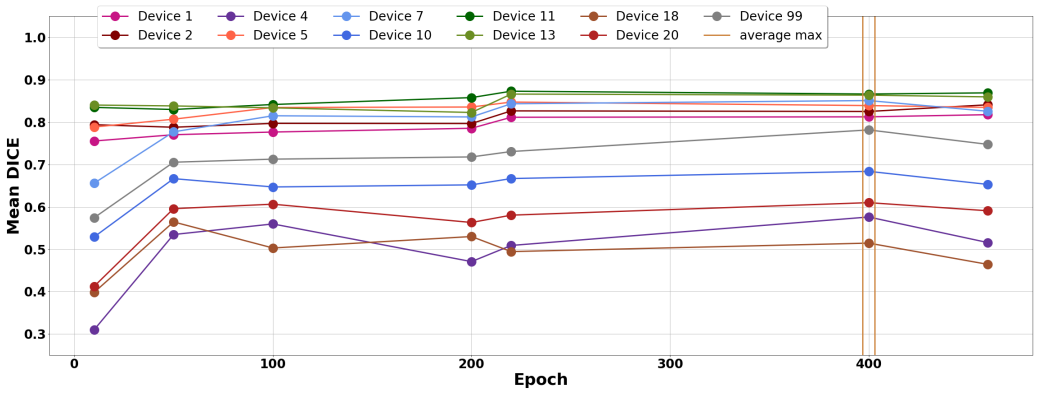


Figure 4: **DICE score per mammography facility at different epochs in the test set.** The first epochs already get acceptable results for images in which the quality is high. As training iterations increase, accuracy increases in these devices and the model is also able to improve its accuracy for the facilities in which their acquisition device image quality is worse. Finally, epoch 400 gets the best averaged score and the model is selected at this point.

It should be noted that devices of mammography facilities 1, 2, 5, 11, and 13 come from the same manufacturer and the sum of images in these mam-

mography facilities exceeds by far images coming from other manufacturers. It may influence the good performance at early epochs on images of these mammography facilities. The model seems to improve its results on images from other devices when the local maxima are near to be reached in these mammography facilities which share the same device (the most represented in the corpus).

### *3.3. ECNN segmentation compared with two radiologists*

A brief comparison of the obtained DICE scores can be found in Table 4.

These results demonstrate a good agreement level of ECNN with segmentations provided by experienced radiologists. As can be seen in Table 1, the mammography facilities with a FUJIFILM device (mammography facilities 01, 02, 05, 11, and 13) are those that present better results in Table 4. Those mammography facilities presenting lower levels of agreement for the ECNN are also the least populated. This situation makes us suspect that training the model using a balanced number of images per device could increase the reported scores. This probable increment in the performance would be always bounded by the lower gray-level resolution observed in these devices. It also leads to a lower agreement between specialists, with exception of the mammography facility 05 (FUJIFILM acquisition device) where DICE between radiologists is surprisingly low.

| Medical facility | test size | ECNN vs closer | R1 vs R2 | # ECNN closer to R1 than R2 | # ECNN closer to R2 than R1 | # ECNN closer to R1 or R2 |
|---|---|---|---|---|---|---|
| **01** | 96 | 0.81 | 0.79 | 52 | 35 | 58 |
| **02** | 96 | 0.83 | 0.79 | 51 | 43 | 63 |
| **04** | 34 | 0.58 | 0.75 | 7 | 3 | 8 |
| **05** | 80 | 0.84 | 0.65 | 64 | 63 | 76 |
| **07** | 14 | 0.85 | 0.88 | 3 | 5 | 6 |
| **10** | 156 | 0.68 | 0.77 | 42 | 57 | 65 |
| **11** | 140 | 0.87 | 0.82 | 63 | 85 | 100 |
| **13** | 60 | 0.86 | 0.78 | 30 | 43 | 49 |
| **18** | 20 | 0.51 | 0.74 | 2 | 4 | 6 |
| **20** | 90 | 0.61 | 0.78 | 15 | 19 | 27 |
| **99** | 58 | 0.78 | 0.79 | 19 | 25 | 35 |
| **Total** | 844 | 0.77 | 0.77 | 348 | 382 | 493 |

Table 4: ECNN segmentation DICE scores in function with acquisition devices. Test size column is the number of mammograms available in the test set for each mammography facility. The third column refers to DICE score when ECNN is considered as other radiologist. Fourth column is the DICE score between radiologists. The last three columns show the number of segmentations in which ECNN-R1 are closer than R1-R2, ECNN-R2 are closer than R1-R2 and ECNN-[R1 or R2] is closer than R1-R2.

ECNN outperforms in many devices when compared to the agreement between radiologists and still obtains better results in some devices when it is considered as an specialist. It highlights that almost 60% of ECNN segmentation masks (493 out of 844) are closer to one of the radiologists than the radiologists to each other. This percentage is increased in facilities with FUJIFILM devices. This suggests that ECNN could be considered as an independent reader, but a validation considering the segmentations from other radiologists is needed.

### *3.4. Histogram normalization importance*

Figure 5 shows how image preprocessing increases the performance of our ECNN.

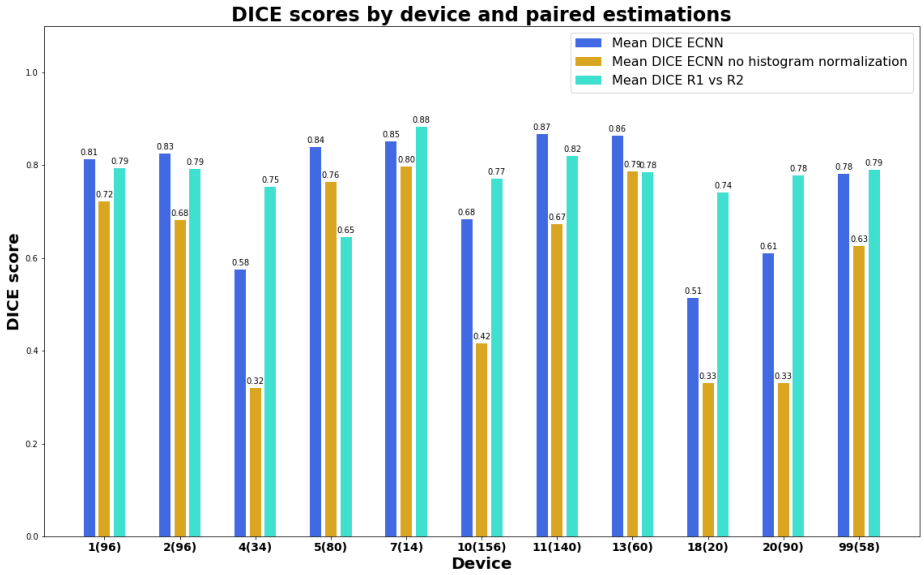


Figure 5: **Comparison of ECNN segmentation using and not using a preprocessing step.** It is observed that results using the proposed histogram normalization outperforms those obtained without any preprocess

The substantial increment in the performance of our model, when a preprocessing step is carried out, captures how variability among acquisition devices impacts in the mammogram analysis. These results support the need for standardization of gray-level values from different sources before modeling problems using mammograms.

### *3.5. Specific segmentation model per acquisition device*

Having images from different devices could act as a confounder for the models, so the next step was to check if the performance of percent density

estimation improved when a specific model is trained for each mammography facility. In this sense, two models using the train images only from one mammography facility were trained. One of the models was trained using mammograms from the mammography facility 01 and the other using those from the mammography facility 18. The performance results over the same samples (test corpus from devices 01 and 18) are shown in Table 5. They suggest that using a generic model does not imply a substantial loss of performance compared to a specific model.

| **Medical Center** | **test size** | **ECNN vs closer** | **R1 vs R2** | **# ECNN closer to R1 than R2** | **# ECNN closer to R2 than R1** | **# ECNN closer to R1 or R2** |
|---|---|---|---|---|---|---|
| **01** | 96 | 0.82(0.81) | 0.79 | 41(52) | 44(35) | 59(58) |
| **18** | 20 | 0.58(0.51) | 0.74 | 4(2) | 2(4) | 5(6) |

Table 5: Specialized models segmentation DICE scores in function with acquisition devices. Test size column is the number of mammograms available in the test set for each mammography facility. The third column refers to DICE score when ECNN is considered as other radiologist. Fourth column is the DICE score between radiologists. The last three columns show the number of segmentations in which ECNN-R1 are closer than R1-R2, ECNN-R2 are closer than R1-R2 and ECNN-[R1 or R2] is closer than R1-R2. Values in parentheses are the results for the global model.

The specialized model for mammography facility 18 obtained better results when compared to the global model but, still, poor concordance is maintained probably due to the lack of training images and/or the poor quality of them.

## 4. Discussion

According to [11, 44, 45], one of the important tasks for computer-aided diagnosis systems is to provide an accurate and reproducible assessment of mammographic breast density. We consider that our multi-center study demonstrates a good performance of breast density assessment using ECNN, and constitutes a first step in the standardization of how mammographic breast density is assessed. Globally, the score obtained by the proposed framework is comparable, in terms of concordance, to the score obtained by two radiologists.

Typical convolution usage covers pixel-level classification tasks, using convolutional autoencoder architectures [46, 47], or pattern recognition based classification tasks, using fully connected convolutional neural networks [48, 49], or Deep Residual Learning for BI-RADS breast density categories classification [50]. Since our output was continuous, approaches intended to pixel-level classification were discarded. A fully convolutional neural network to estimate the threshold segmentation-based parameters (CNN) was overcome by the architecture in which the desired paramenters are directly extracted as features of the image (ECNN). The performance of the ECNN is better than the obtained by CNN, however this architecture obtain significant better performance for two over the eleven facilities (04 and 18). These mammography facilities have the same acquisition device model and it is also the less represented one in the sample. We expect that increasing the number of images from devices of this model may improve the segmentation results. It is also worth to mention that automatic segmentation applied to the most represented device (FUJIFILM in facilities 01, 02, 05, 11, and 13) were closer

to one of the radiologists than each radiologist to the other 73% times (346 out of 472), implying a significant DICE score improvement, outperforming the radiologists concordance.

The main contributions of the present paper can be summarized as:

1. An intuitive preprocess protocol standardizes the histograms of breasts by centering the mode and stretching the tails of the histograms. It allows to extend the range in which the fibroglandular threshold is found. This step reduced the impact of using different acquisition devices.
2. A convolution-based architecture trained to learn the two desired parameters used by radiologists to segment the image. The results provided by this approach obtained slightly better results compared to state-of-the-art algorithms with lower computing workload.
3. An ad hoc, continuous, and differentiable loss function which rebuilds the intended mask from the estimated parameters and assesses the DICE score against the "training ground truth".
4. The approach followed makes easy that a radiologists perform a fine-tuning of the results by interactively modifying the segmentation parameters using a tool such as DMScan.

### *4.1. Limitations and future research*

While the parameter based approach was justified to make it compatible with threshold-based semi-automatic tools, exploring other, supervised or unsupervised, mask-based approaches is planned. Supervised mask based approaches could deal with the suboptimal results obtained in some devices and unsupervised approaches would let us complement the models using large databases without the need of human effort.

A second limitation is the pectoral muscle exclusion algorithm. The solution adopted in the present work, although robust, could be improved by taking into account other approaches mentioned in Section 2.2.2.

Finally, the use of "for presentation" mammograms instead of "raw" images may be the reason for some of the differences among acquisition devices. It is also desirable to check if "Raw" mammograms would avoid the preprocessing step.

## 5. Conclusion

Nowadays, with the explosion of complex models that can identify features and patterns which are undetectable to the human eye, having a large amount of labeled mammograms is highly necessary for basic and clinical research. In this sense, the availability of a tool that provides automatic segmentation of dense tissue on processed digital mammographies with a high level of concordance with the segmentation of experienced radiologists is desirable.

The work presented in this paper provides an automatic framework based on deep learning which detects the breast, excludes the pectoral muscle, and finally performs a dense tissue segmentation. Our approach is based on the estimation of two segmentation parameters which are learned as image level features. A preprocess step alleviates the influence of the variability among mammograms from different sources and improved the algorithm performance.

The concordance scores (DICE) of the proposed framework are close to the agreement achieved between two radiologists in a multi-center (and multi-

device) study. Images from those devices with the highest gray-level resolution provide concordance results even better than those raised by two experienced specialists, suggesting that our model could be used as a fully independent reader. As a final contribution, if the radiologist does not agree with the segmentation proposal, it may easily fine-tuned using a software tool, DMScan, built in our laboratory and freely available for research purposes.

**Acknowledgements**

The authors of this work like to thank to Guillermo García Colomina, Carlos Barata Ferrando and Empar Giner Ferrando for their support in recruitment and data collection.

**Funding**

This work was partially funded by Generalitat Valenciana through I+D IVACE (Valencian Institute of Business Competitiviness) and GVA (European Regional Development Fund) supports under the project IMAMCN/2019/1, and by Carlos III Institute of Health under the project DTS15/00080.

**Ethics approval and consent to participate**

This study was approved by the Research Ethics Committee of the Universitat Politècnica de València (project name: "DM-Scan Herramienta de lectura de densidad mamográfica como fenotipo marcador de riesgo de cáncer de mama") and consent was obtained from study participants at the time of screening.

## References


[1] C. K. Kuhl, The changing world of breast cancer: a radiologist's perspective, Invest. Radiol. 50 (9) (2015) 615 (2015).

[2] N. F. Boyd, J. M. Rommens, K. Vogt, V. Lee, J. L. Hopper, M. J. Yaffe, A. D. Paterson, Mammographic breast density as an intermediate phenotype for breast cancer, The Lancet Oncology 6 (10) (2005) 798–808 (2005).

[3] V. Assi, J. Warwick, J. Cuzick, S. W. Duffy, Clinical and epidemiological issues in mammographic density, Nature Reviews Clinical Oncology 9 (1) (2012) 33 (2012).

[4] C. J. D'Orsi, E. Sickles, E. Mendelson, E. Morris, ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, Reston, VA, American College of Radiology, 2013 (2013).

[5] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Pérez, E. R. Denton, R. Zwiggelaar, A novel breast tissue density classification methodology, IEEE T. Inf. Technol. B. 12 (1) (2008) 55–65 (2008).

[6] F. J. Pérez-Benito, F. Signol, J.-C. Perez-Cortes, M. Pollán, B. Pérez-Gómez, D. Salas-Trejo, M. Casals, I. Martínez, R. LLobet, Global parenchymal texture features based on histograms of oriented gradients improve cancer development risk estimation from healthy breasts, Comput. Meth. Prog. Bio. 177 (2019) 123–132 (2019).

[7] S. Ciatto, N. Houssami, A. Apruzzese, E. Bassetti, B. Brancato, F. Carozzi, S. Catarzi, M. Lamberini, G. Marcelli, R. Pellizzoni, et al.,

Categorizing breast mammographic density: intra-and interobserver reproducibility of bi-rads density categories, The Breast 14 (4) (2005) 269–275 (2005).

[8] P. Skaane, Studies comparing screen-film mammography and full-field digital mammography in breast cancer screening: updated review, Acta Radiologica 50 (1) (2009) 3–14 (2009).

[9] D. van der Waal, G. J. den Heeten, R. M. Pijnappel, K. H. Schuur, J. M. Timmers, A. L. Verbeek, M. J. Broeders, Comparing visually assessed bi-rads breast density and automated volumetric breast density software: a cross-sectional study in a breast cancer screening setting, PLoS One 10 (9) (2015) e0136667 (2015).

[10] S. H. Kim, E. H. Lee, J. K. Jun, Y. M. Kim, Y.-W. Chang, J. H. Lee, H.-W. Kim, E. J. Choi, et al., Interpretive performance and interobserver agreement on digital mammography test sets, Korean journal of radiology 20 (2) (2019) 218–224 (2019).

[11] K. J. Geras, R. M. Mann, L. Moy, Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives, Radiology (2019) 182627 (2019).

[12] R. Miotto, F. Wang, S. Wang, X. Jiang, J. T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Brief. Bioinform. 19 (6) (2017) 1236–1246 (2017).

[13] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436 (2015).

[14] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, et al., Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Proc. Mag. 29 (2012).

[15] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: A survey, Pattern Recogn. Lett. 119 (2019) 3–11 (2019).

[16] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, W. Denk, Connectomic reconstruction of the inner plexiform layer in the mouse retina, Nature 500 (7461) (2013) 168 (2013).

[17] K. Lee, N. Turner, T. Macrina, J. Wu, R. Lu, H. S. Seung, Convolutional nets for reconstructing neural circuits from brain images acquired by serial section electron microscopy, Curr. Opin. Neurobiol. 55 (2019) 188–198 (2019).

[18] M. K. Leung, H. Y. Xiong, L. J. Lee, B. J. Frey, Deep learning of the tissue-regulated splicing code, Bioinformatics 30 (12) (2014) i121–i129 (2014).

[19] J. Zhou, C. Y. Park, C. L. Theesfeld, A. K. Wong, Y. Yuan, C. Scheckel, J. J. Fak, J. Funk, K. Yao, Y. Tajima, et al., Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk, Nat. Genet. 51 (6) (2019) 973 (2019).

[20] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, et al., Unsu-

pervised deep learning applied to breast density segmentation and mammographic risk scoring, IEEE transactions on medical imaging 35 (5) (2016) 1322–1331 (2016).

[21] S. K. Zhou, H. Greenspan, D. Shen, Deep learning for medical image analysis, Academic Press, 2017 (2017).

[22] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al., Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324 (1998).

[23] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: Proc. CVPR. IEEE, 2015, pp. 648–656 (2015).

[24] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proc. CVPR. IEEE, 2014, pp. 1701–1708 (2014).

[25] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, arXiv preprint arXiv:1312.6229 (2013).

[26] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302 (1945).

[27] M. Pollán, R. Llobet, J. Miranda-García, J. Antón, M. Casals, I. Martínez, C. Palop, F. Ruiz-Perales, C. Sánchez-Contador, C. Vidal, et al., Validation of dm-scan, a computer-assisted tool to assess

mammographic density in full-field digital mammograms, Springerplus 2 (1) (2013) 242 (2013).

[28] R. Llobet, M. Pollán, J. Antón, J. Miranda-García, M. Casals, I. Martínez, F. Ruiz-Perales, B. Pérez-Gómez, D. Salas-Trejo, J.-C. Pérez-Cortés, Semi-automated and fully automated mammographic density measurement and breast cancer risk prediction, Comput Methods Programs Biomed 116 (2) (2014) 105–115 (2014).

[29] L. He, X. Ren, Q. Gao, X. Zhao, B. Yao, Y. Chao, The connected-component labeling problem: A review of state-of-the-art algorithms, Pattern Recognit. 70 (2017) 25–43 (2017).

[30] K. Wu, E. Otoo, K. Suzuki, Optimizing two-pass connected-component labeling algorithms, Pattern Anal. Appl. 12 (2) (2009) 117–135 (2009).

[31] R. Lakshmanan, V. Thomas, S. M. Jacob, P. Thara, et al., Pectoral muscle boundary detection in mammograms using homogeneous contours, in: 2015 Fifth International Conference on Advances in Computing and Communications (ICACC), IEEE, 2015, pp. 354–357 (2015).

[32] R. Shen, K. Yan, F. Xiao, J. Chang, C. Jiang, K. Zhou, Automatic pectoral muscle region segmentation in mammograms using genetic algorithm and morphological selection, J. Digit. Imaging 31 (5) (2018) 680–691 (2018).

[33] K. Yin, S. Yan, C. Song, B. Zheng, A robust method for segmenting pectoral muscle in mediolateral oblique (mlo) mammograms, Int. J. Camput. Ass. Rad. 14 (2) (2019) 237–248 (2019).

[34] V. Shinde, B. T. Rao, Novel approach to segment the pectoral muscle in the mammograms, in: Cognitive Informatics and Soft Computing, Springer, 2019, pp. 227–237 (2019).

[35] J. James, The current status of digital mammography, Clin. Radiol. 59 (1) (2004) 1–10 (2004).

[36] C. Sáez, M. Robles, J. M. Garcia-Gomez, Comparative study of probability distribution distances to define a metric for the stability of multi-source biomedical research data, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2013, pp. 3226–3229 (2013).

[37] C. Sáez, M. Robles, J. M. García-Gómez, Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances, Statistical methods in medical research 26 (1) (2017) 312–336 (2017).

[38] W. He, S. Harvey, A. Juette, E. R. Denton, R. Zwiggelaar, Mammographic segmentation and density classification: a fractal inspired approach, in: International Workshop on Breast Imaging, Springer, 2016, pp. 359–366 (2016).

[39] A. K. Jain, Data clustering: 50 years beyond k-means, Pattern Recogn. Lett. 31 (8) (2010) 651–666 (2010).

[40] J. Lee, R. M. Nishikawa, Automated mammographic breast density estimation using a fully convolutional network, Med. Phys. 45 (3) (2018) 1178–1190 (2018).

[41] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[42] T. Buelow, H. S. Heese, R. Grewer, D. Kutra, R. Wiemker, Inter-and intra-observer variations in the delineation of lesions in mammograms, in: Medical Imaging 2015: Image Perception, Observer Performance, and Technology Assessment, Vol. 9416, International Society for Optics and Photonics, 2015, p. 941605 (2015).

[43] W. Alakwaa, M. Nassef, A. Badr, Lung cancer detection and classification with 3d convolutional neural network (3d-cnn), Lung Cancer 8 (8) (2017) 409 (2017).

[44] N. Wu, K. J. Geras, Y. Shen, J. Su, S. G. Kim, E. Kim, S. Wolfson, L. Moy, K. Cho, Breast density classification with deep convolutional neural networks, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 6682–6686 (2018).

[45] C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, R. Barzilay, Mammographic breast density assessment using deep learning: clinical implementation, Radiology 290 (1) (2018) 52–58 (2018).

[46] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE T. Pattern Anal. 35 (8) (2013) 1798–1828 (2013).

[47] G. Wu, M. Kim, Q. Wang, B. C. Munsell, D. Shen, Scalable high-performance image registration framework by unsupervised deep feature

representations learning, IEEE T. Bio-med. Eng. 63 (7) (2015) 1505–1516 (2015).

[48] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241 (2015).

[49] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 565–571 (2016).

[50] T. P. Matthews, S. Singh, B. Mombourquette, J. Su, M. P. Shah, S. Pedemonte, A. Long, D. Maffit, J. Gurney, R. M. Hoil, et al., A multi-site study of a breast density deep learning model for full-field digital mammography and digital breast tomosynthesis exams, arXiv preprint arXiv:2001.08383 (2020).