Computer Methods and Programs in Biomedicine 221 (2022) 106885

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

A deep learning framework to classify breast density with noisy labels regularization

Hector Lopez-Almazan[a,*], Francisco Javier Pérez-Benito[a], Andrés Larroza[a], Juan-Carlos Perez-Cortes[a], Marina Pollan[b,c], Beatriz Perez-Gomez[b,c], Dolores Salas Trejo[d,e], María Casals[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

ARTICLE INFO

Article history:
Received 21 December 2021
Revised 12 April 2022
Accepted 10 May 2022

Keywords:
Breast density
Noisy labels
Deep learning
Dense tissue classification
Mammography

ABSTRACT

Background and Objective: Breast density assessed from digital mammograms is a biomarker for higher risk of developing breast cancer. Experienced radiologists assess breast density using the Breast Image and Data System (BI-RADS) categories. Supervised learning algorithms have been developed with this objective in mind, however, the performance of these algorithms depends on the quality of the ground-truth information which is usually labeled by expert readers. These labels are noisy approximations of the ground truth, as there is often intra- and inter-reader variability among labels. Thus, it is crucial to provide a reliable method to obtain digital mammograms matching BI-RADS categories. This paper presents RegL (Labels Regularizer), a methodology that includes different image pre-processes to allow both a correct breast segmentation and the enhancement of image quality through an intensity adjustment, thus allowing the use of deep learning to classify the mammograms into BI-RADS categories. The Confusion Matrix (CM) - CNN network used implements an architecture that models each radiologist's noisy label. The final methodology pipeline was determined after comparing the performance of image pre-processes combined with different DL architectures.

Methods: A multi-center study composed of 1395 women whose mammograms were classified into the four BI-RADS categories by three experienced radiologists is presented. A total of 892 mammograms were used as the training corpus, 224 formed the validation corpus, and 279 the test corpus.

Results: The combination of five networks implementing the RegL methodology achieved the best results among all the models in the test set. The ensemble model obtained an accuracy of (0.85) and a kappa index of 0.71.

Conclusions: The proposed methodology has a similar performance to the experienced radiologists in the classification of digital mammograms into BI-RADS categories. This suggests that the pre-processing steps and modelling of each radiologist's label allows for a better estimation of the unknown ground truth labels.

1. Background

Mammogram screening is a highly standardized procedure for breast cancer early detection programs [1,2]. Breast density measures the proportion of fibroglandular tissue over the entire breast. Breast density is widely measured qualitatively using the Breast

* Corresponding author.
E-mail addresses: hlopez@iti.es (H. Lopez-Almazan), fjperez@iti.es (F. Javier Pérez-Benito), alarroza@iti.es (A. Larroza), jcperez@iti.upv.es (J.-C. Perez-Cortes), mpollan@isciii.es (M. Pollan), bperez@isciii.es (B. Perez-Gomez), salas_dol@gva.es (D. Salas Trejo), casals_mar@gva.es (M. Casals), rllobet@iti.upv.es (R. Llobet).

https://doi.org/10.1016/j.cmpb.2022.106885

Image and Data System (BI-RADS) [3] developed by the American College of Radiologists. The BI-RADS system includes four qualitative categories: I (almost fatty), II (scattered fibroglandular densities), III (heterogeneously dense), and IV (extremely dense). Several studies have shown that radiologists exhibit intra- and inter-reader variability in the assessment of breast density into BI-RADS categories [4–6].

Breast density is known to be a risk marker for the development of breast cancer [7–10]. Dense breasts are those given a clinical BI-RADS assessment of either heterogeneously or extremely dense categories (III and IV, respectively). The breast density assessment is usually carried out using a semiautomatic tool such as DM-Scan [11,12].

Before the appearance of deep learning, classical computer vision techniques, such as Gabor filters, Histogram of Oriented Gradients (HOG), etc. [13,14] were used to carry out the breast density classification of digital mammograms. The performance of convolutional neural networks (CNN) compared to typical algorithms in computer vision tasks led to the establishment of the former as the new standard in computer vision problems [15–17]. CNNs give great performance on different areas such as waste classification, pedestrian detection, etc [18–22]. As well as these areas, it has also become the standard for X-Ray computer vision problems in healthcare environments [23–26].

The existing variability in the assessment means there is noise in the labels. Obtaining accurate labels is a challenging task. Some studies used employed a huge number of experts to obtain more accurate labels [27,28]. Training deep learning models with datasets containing noisy labels leads to poor generalization capabilities. Some studies use different deep learning related techniques to improve generalization [29,30], while other works propose more complex frameworks to perform classification via deep learning in presence of noisy labels [31–33].

This current work presents a fully automated framework for dense tissue classification into BI-RADS categories. This framework has been denominated RegL (Labels Regularizer) and it has been applied to regularize the specialists' BI-RADS label variability. It includes breast detection, intensity adjustment, and dense tissue classification. Among the contributions of this work, we want to highlight: (1) a preprocessing algorithm capable of eliminating noise in the background of mammograms, (2) another preprocessing algorithm capable of correcting the range of intensities, (3) a final pre-process to standardize the grey level variability, and (4) the implementation of a CNN architecture that models different radiologists opinions.

2. Methods

2.1. Dataset and participants

A multi-center study covered women from 7 Spanish screening centres which belong to the Spanish breast cancer screening network. This study, called DDM-Spain, recruited 3584 women aged 45 – 68 years to investigate the influence of lifestyle and genetic factors on observed breast density. All participants agreed to their left cranio-caudal mammograms (single view) being used for study purposes [34]. A subset of 1395 full-field digital mammograms from 3 screening centers was used to evaluate the intra- and inter-reader variability in dense tissue estimation. Three experienced radiologists (referred to here as R1, R2, and R3) assessed the breast density using the 4th edition of BI-RADS, which classifies the breast density into four categories based on the density percentage, as shown in Table 1. The BI-RADS scale is the most common breast density assessment method [35–37]. Besides, thinking about future projects, the radiologists subdivided the first BI-RADS category into three subcategories (I :0%; II :110%; III :1025%;) to get the classification into the Boyd scale [38], which also is a percentage-based scale. Nevertheless, the Boyd system was not used in this paper.

Table 1
BI-RADS 4th edition scale for breast density classification. The 4th edition of BI-RADS is percentage-based, where each category comprises a range of percentages.

BI-RADS category	Density percentage
I	< 25%
II	25 - 50%
III	50 - 75%
IV	> 75%

Table 2
Inter-reader variability amongst radiologist labels. Each row shows a comparison considering the labels from the first radiologist as predictions and the labels from the second as ground truth.

Ground Truth vs Predictions	Accuracy (%)	Kappa
R1 vs R2	84.6	0.71
R2 vs R3	75.7	0.57
R1 vs R3	77.1	0.59

As previously mentioned, three radiologists made the breast density assessment, with the unavoidable inclusion of a degree of subjectivity and variability, leading to inexact ground truths. Supervised learning requires a unique ground-truth label. With this in mind, for each mammogram, we calculated the majority vote for those images where at least two readers have matching opinions and the median for those in which the three radiologists have differing opinions. Once a single-labelled dataset was extracted, we realized that there were imbalances in the dataset. The BI-RADS 1 category represents 61.4% of the total images (857 images), BI-RADS 2 25.2% (352 images), BI-RADS 3 11.7% (163 images) and BI-RADS 4 1.6% (23 images). To address this problem we used weighting techniques, as explained in Section 2.4.

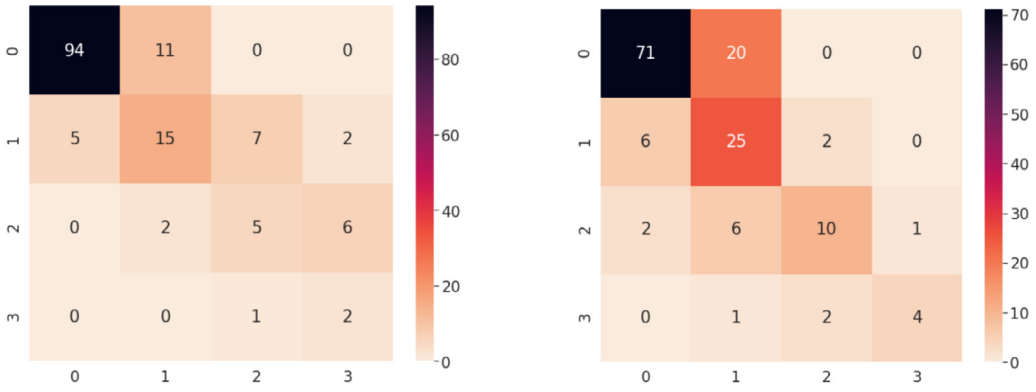
2.2. Intra and inter-reader variability

We calculated the inter-reader variability using the labels of the three radiologists. We defined a series of experiments in which, for each of them, a radiologist's labels were selected as ground truth and another radiologist's labels as predictions, which served to obtain the accuracy and kappa index score as a way to measure the radiologists' concordance. The three most representative comparisons between radiologists are shown in Table 2. These results show that concordance between R1 and R2 is higher than that achieved with R3. The accuracy ranges from 75.7% to 84.6%, while the kappa index ranges from 0.57 to 0.71.

It was also possible to calculate the intra-reader variability as the same radiologists made a second assessment on a subset of 145 mammograms, which have thus been labelled twice. Similarly as with inter-reader variability, intra-reader variability has been measured in terms of accuracy and kappa index, as shown in Table 3. These results show that even the same radiologist does not fully agree with his/her own labels after a certain time. Table 3 shows that both the R1 and R3 experiments obtained the same results, this is pure coincidence considering that their confusion matrices are different, as shown in Fig. 1(b).

2.3. Breast density classification with noisy labels framework

The RegL framework consist of two steps, (1) a preprocessing pipeline called Digital Mammograms Preprocessing Pipeline (DMPP) and (2) dense tissue classification by using the Confusion



(a) Confusion matrix from the intra-reader comparision on R1.

(b) Confusion matrix from the intra-reader comparision on R3.

Fig. 1. Confusion matrices from the intra-reader comparision of both R1 and R3. (a) R1 confusion matrix and (b) R3 confusion matrix.

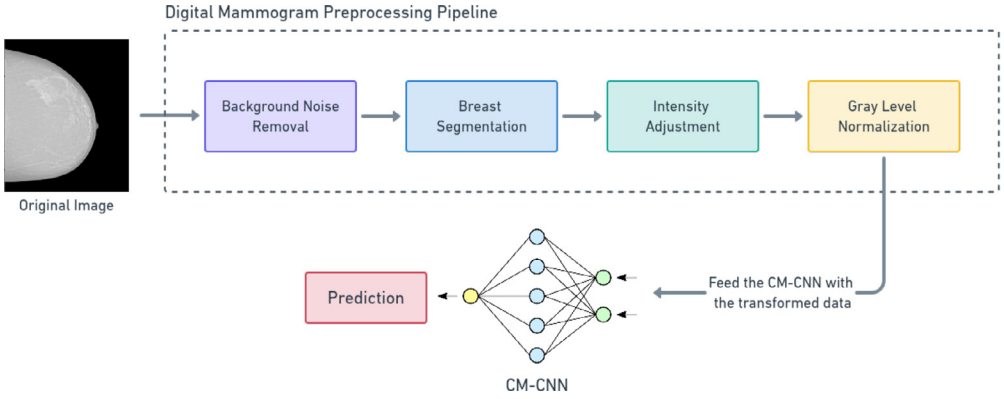


Fig. 2. RegL general diagram. First, the digital mammogram is fed into the DMPP, removing the background noise of the image, segmenting the breast, adjusting the intensity from the breast pixels, and, finally, normalizing its grey level, thus feeding the CM-CNN with the preprocessed mammogram.

Table 3
Intra-reader variability for each radiologist. Each row shows the results from comparing the first radiologist labels with the labels made after a certain time interval.

Radiologist	Accuracy (%)	Kappa
R1	77.3	0.54
R2	80.6	0.63
R3	77.3	0.54

Matrix Convolutional Neural Network (CM-CNN). Fig. 2 shows a general diagram of the classification framework.

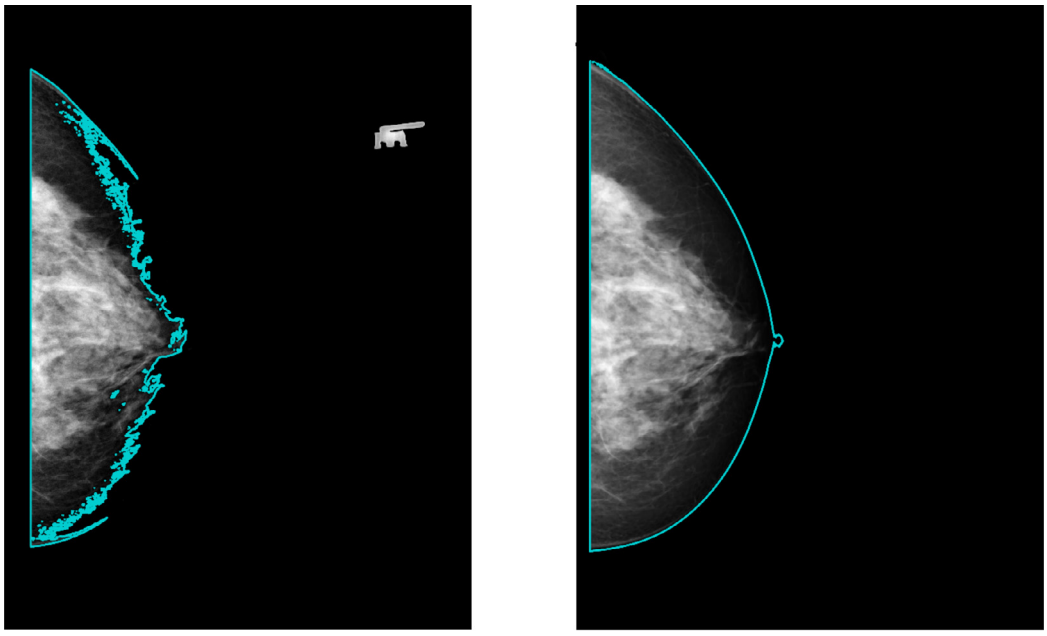
2.3.1. Digital mammograms preprocessing pipeline

The DMPP is made up of four pre-processing steps: (1) noise removal from the mammograms background pixels, (2) breast segmentation, (3) intensity adjustment from the breast pixels and (4) normalization of the grey level variability.

Background noise removal. It has been found that in a significant number of mammograms, the background pixels are not actually black but have a greyish hue, whose value is close to the pixels at the edge of the breast. Due to this problem, the breast segmentation algorithm performs poorly in these images, segmenting the breast edge pixels as if they belonged to the background. We have developed a process that gets the maximum intensity from a subregion of the image. Given that the mammograms are CC views from the left breast, the subregion is near the lower right corner of the image since the image labels and marks, such as the one in Fig. 3(a), only appear in the upper right corner. Then, it obtains the maximum intensity of that subregion and performs an intensity windowed adjustment transformation on the lower intensity values of the image. Therefore, intensities lower than the maximum intensity found are set to 0. Fig. 3 shows the improvement in segmentation by removing the noise in the background.

Breast segmentation. The segmentation algorithm consists of an iterative algorithm based on connected components. This algorithm obtains the grey level threshold that distinguishes the breast from the background. Even though there are some issues concerning the use of connected-components based labelling on binary images [39], homogeneous breast shape makes this kind of algorithm suitable to be used for breast segmentation and exhibits perfect breast detection.

The first step of the breast detection procedure is to assess the histogram of the image. Based on the premise that the most frequent pixel values belong to the background, we defined a range of possible thresholds that separate the breast and the background. Following this, the segmentation algorithm assures that the breast is left-oriented and binarizes the image using the first possible threshold, thus applying the connected component labelling method. We chose the Scan plus Array-based Union-Find (SAUF) algorithm, [40] a two-scan algorithm: the first scan assigns provisional labels to pixels and records labels equivalences. Thus, the second scan replaces the equivalent labels with their represen-



(a) Breast segmentation with noise in the background.

(b) Breast segmentation without noise in the background.

Fig. 3. Comparison between breast segmentation applied to an image with (a) noise in the background and (b) after removing the noise from the background.

tative label [41]. Finally, if only two components are obtained, the threshold is set. If not, the same procedure is applied to the remaining ones.

Intensity Adjustment. A visual inspection of the images showed unusual brightness problems in the breast pixels of some mammograms, such as excess or lack of brightness. To solve this problem, we developed a process to find the optimal intensity window of an image and, then, to make a windowed intensity adjustment on that optimal window.

To find the optimal window values, firstly we applied a median filter to remove noise while preserving the edges. Following this, we obtained the cumulative distribution function (CDF) from the breast pixel values and searched the lower and upper elbows of the function in an iterative way. These elbows correspond to the values that define the optimal window. Finally, we applied a windowed intensity adjustment using the calculated window. Fig. 4 shows the results of adjusting the intensity on a bright mammogram.

Normalizing grey level variability. The pixel size, grey-scale bit resolution, signal-to-noise ratio, or detective quantum efficiency are crucial concepts related to image quality [42]. Factors such as the acquisition devices or the process used to capture the image create a high degree of variability in the quality of the mammograms.

We analysed the grey levels of a random sample of 100 mammograms. Before this analysis, the images were processed applying the transformations detailed in previous sections. The assessment consists of the visualization of the mean density function for each category. Fig. 5 shows the comparison between density functions, confirming that there is a difference in the grey level amongst categories, not only in the brighter pixels, but also in the darker ones.

Mammogram features such as 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 pre-process to normalize the grey-level variability composed of the following preprocessing steps:

1. Shift histogram to set the minimum breast tissue pixel to 0.
2. Normalize the pixel values of the image between $[0, 1]$.
3. Adjust the pixel values so that the mode is 0.

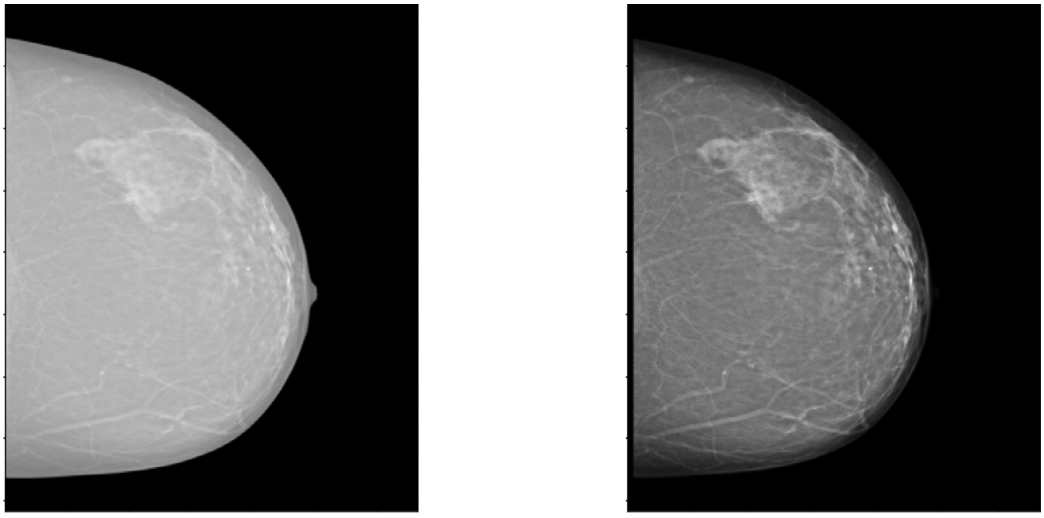
Fig. 6 shows the comparison of the mean density functions per category after normalizing the grey level variability. This way, the higher the category, the fewer pixels there will be on the left of the mode, while there will be a greater number on the right. This makes sense considering that the higher the category, the greater the amount of dense tissue, whose tonal-intensities are close to white.

Fig. 7 shows adjusting the intensity and then normalizing the grey level variability for each BI-RADS category. There are no visual differences between the intensity adjusted and grey level normalized images, which is expected since the grey level normalization does not intend to improve the image's visual quality. Instead, it tries to make the classification task a little bit easier by eliminating the variability produced by external factors.

2.3.2. Confusion matrix convolutional neural network architecture

Although there are four BI-RADS categories, the works detailed in this paragraph turned the problem into a binary classification problem due to the subjectivity that exists in the evaluation. Authors of [43] decided to classify mammograms in the BI-RADS 2 and BI-RADS 3 categories because these are the most critical categories for radiologists to distinguish. Other works grouped BI-RADS 1 and BI-RADS 2 into a new category called non-dense, and BI-RADS 3 and BI-RADS 4 into another called dense, and then performed a binary classification of these categories [44,45].

Other approaches use convolutional neural networks to classify mammograms into the four BI-RADS categories. Some works use transfer learning with well-known pre-trained networks on ImageNet [46], such as the ResNet-50 [47], to classify the mammograms [48]. In another work, firstly, the dense tissue segmentation is carried out to perform the classification only on the segmented tissue [49]. Other work approaches the problem by estimating the percentage density and translating it into the BI-RADS categories [50].



(a) Original image (b) Transformed image

Fig. 4. Example of adjusting the intensity range from an image. (a) Original image with excess of brightness and (b) image with the intensity adjusted.

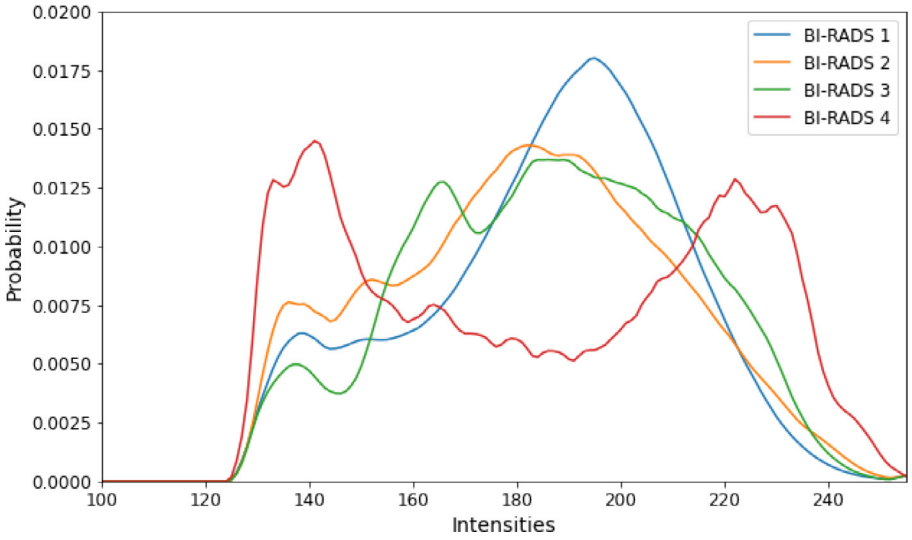


Fig. 5. Mean density functions per category from the random sample of 100 images of the analysis.

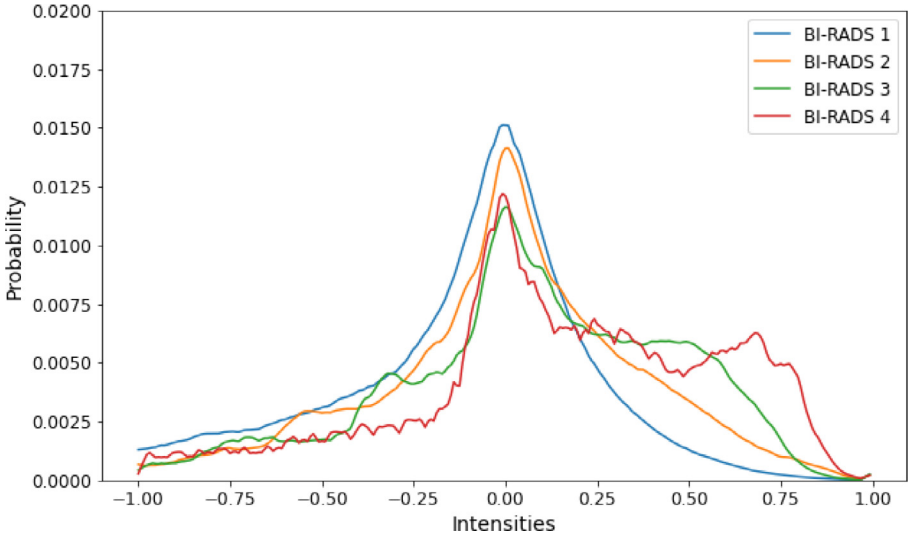


Fig. 6. Mean densities functions per category from the random sample of 100 images after normalizing their gray level.

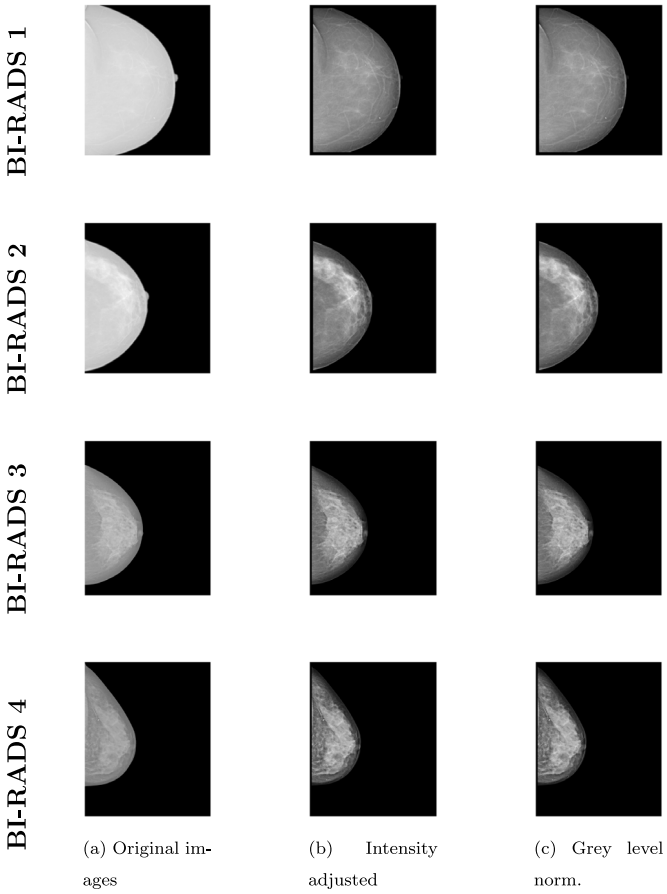


Fig. 7. Example of applying the DMPP to each BI-RADS category. Each row represents a BI-RADS category, while each column represents the output image after a certain preprocess. The first column represents the original images, while the second and third columns represent the images after adjusting the intensity and normalizing the grey-level variability.

Since a correct ground truth does not exist due to the existing intra- and inter-reader variabilities as shown in Section 2.2, we propose a convolutional neural network architecture, named Confusion Matrix Convolutional Neural Network (CM-CNN), capable of modelling multiple labelling opinions individually [51]. An example of the architecture is shown in Fig. 8. This architecture simultaneously learns the criteria of each radiologist together with the ground truth distribution obtained through the majority vote among labels. Each radiologist's opinion is represented by a confusion matrix where each element $[i, j]$ represents the probability that the radiologist classifies an image in category j, under the premise that the ground truth label is the category i. This way, during training, both model weights and confusion matrices are learned. Also, the architecture uses a base classifier to make the predictions, which in this case it is a pre-trained model using transfer learning.

2.4. Categories imbalance problem

As mentioned in Section 2.1, the dataset is quite imbalanced. This problem leads to the generation of suboptimal classification models that have a good coverage of the majority classes, whereas the minority classes are missclassified [52,53]. In our specific case, it leads to a misclassification of the BI-RADS 3 and 4 categories. To solve this, the cross entropy loss was replaced by the weighted cross entropy loss function as shown in Equation 1. Thus, the cost of misclassifying a minority class is much higher than that of misclassifying a majority class, causing the model to focus on correctly classifying minority examples.

$$CE = -\sum_{i \in C} w_i \cdot y_i \cdot \log(\hat{y}_i) \tag{1}$$

Class weights are calculated using the information from the training set, as shown in Equation 2.

$$w_c = \frac{n}{n_c |C|} \tag{2}$$

where n is the total number of samples, n_c the total number of samples in category C and $|C|$ the number of categories.

Table 4
Set of architectures and hyperparameters used in the experimentation.

Model	Optimizer	Learning Rate	L2 Regularization
VGG-19 [54]	Adam	0.000005	0.001
ResNext4D [55]	Adam	0.000005	0.001
DenseNet121 [56]	Adam	0.00001	0.005
WideResNet50 [57]	Adam	0.00001	0.01
EfficientNet-B1 [58]	Adam	0.00003	0.005

Table 5
Distribution of each BI-RADS category in training, validation and test set.

	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4
Training	548	226	104	14
Validation	138	56	26	4
Test	171	70	33	5

Table 6
Class weights for each BI-RADS category on the training set calculated using the Eq. (2).

	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4
N° Images	548	226	104	14
Weight	0.4069	0.9867	2.1442	15.9285

2.5. Experimental design

In this section, we design some experiments aimed at analyzing the importance of the RegL framework. The experimentation consists of a series of experiments that compare the results of five neural networks with and without the RegL framework. The Student's t-test was used to provide statistical analysis for each experiment. The p-values were considered statistically significant at the 0.05 cutoff. All the neural networks were pretrained on the ImageNet dataset. The results comparisions are made considering the majority vote as the ground truth. For the experiments to be reliable, we fixed the hyperparameters of each neural network for all the experiments, as shown in Table 4.

A stratified split of the data based on the majority vote defined the training, validation, and test set. The split ratio is 80% for the training set and 20% for the test set. Also, 20% from the training set were used as the validation set. Finally, from the total of 1395 images, 892 mammograms make up the training set, 224 the validation set and 279 the test set.

Training, validation, and test sets are quite imbalanced, as shown in Table 5. Therefore, as introduced in Section 2.4, the weighted cross entropy was used as the cost function. Table 6 shows the weights for each category on the training set, calculated with Eq. (2). The cost of misclassifying a sample from the BI-RADS 4 category is 39 times greater than that of a sample from the BI-RADS 1 category.

Images were resized to 256×256 px and the pixel values were normalized between $[0, 1]$. Moreover, as the neural networks used in this study (Table 4) were pretrained on the ImageNet dataset

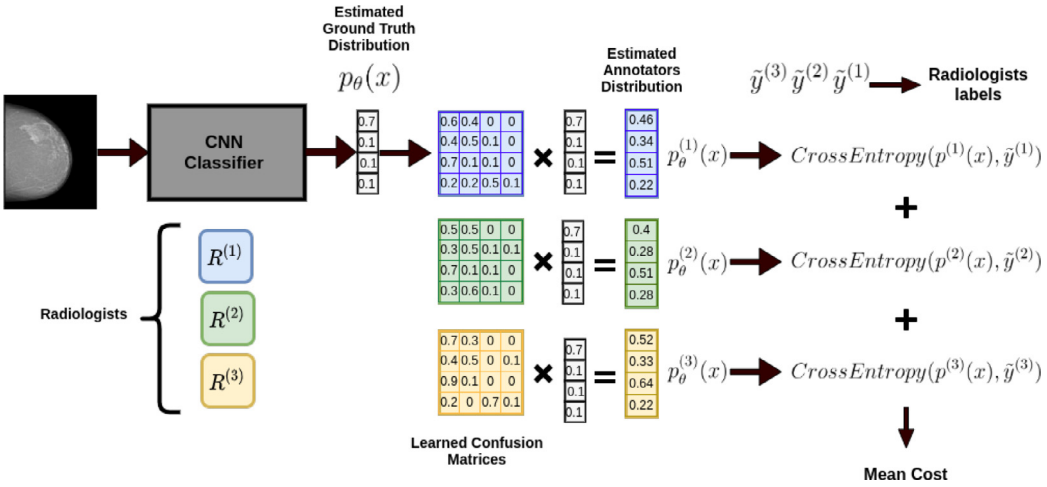


Fig. 8. CM-CNN general diagram. Given an input mammogram, the classifier generates an estimate of the ground truth class probabilities, $p_\theta(x)$. Then the class probabilities for each annotator are computed, $p_\theta^{(r)}(x) = R^{(r)} p_\theta(x)$ for $r \in \{1, 2, 3\}$. Then, the model parameters and annotators confusion matrices are optimized to minimize the mean of three cross-entropy losses between each estimated annotator distribution $p_\theta^{(r)}(x)$ and their respective noisy labels $\hat{y}^{(r)}$ [51].

and, therefore, their input must be 3-channel colour images, the greyscale mammograms were transformed into RGB images by copying the same image on each of the three channels. The maximum number of training epochs was fixed to 30, although no model reached this number of iterations because of early stopping criteria.

The absence of precise ground truth, along with the intra- and inter-reader variability, motivated the use of neural network architectures that modelled each radiologist's opinion, as explained before. We did not expect a model to behave like a specific radiologist, so it is worth noting that the model performance will depend on the expert opinion we choose to compare with.

All the experimentation was done in a remote server with the following specifications: Ubuntu 20.04 as operating system, 16Gb of RAM, and a Nvidia Tesla V100 32Gb GPU. The training was done using Python 3.7 and Pytorch 1.8.

3. Results

This section presents the results of the experimentation defined in Section 2.5. It is organized as follows: first, we show the importance of applying the DMPP and the CM-CNN, then we show the effect of using the entire RegL framework, thus lastly, showing the results of an ensemble of five neural networks applying the RegL framework.

3.1. DMPP importance

In this experiment, we have compared the performance on the test set by two groups of models made up of the neural networks specified in Table 4. The first group of models has been trained and evaluated on images without applying the DMPP, while the second one has been trained and assessed on the same images but applying the DMPP.

The experiment results showed in Table 7 reveal that the DMPP increases the classification capability of the models (statistically significant, p-values < 0.05). Specifically, they achieve an accuracy and kappa index mean improvement of 3.4% and 0.05 points, respectively. The accuracy with DMPP is statistically significant higher (p-value $= 0.017$) than accuracy without DMPP.

Table 7
Comparison experiments to analyze the DMPP influence. The first two columns compare the accuracy obtained when the DMPP was applied or not. On the other hand, the last two columns compare the kappa index when the DMPP was applied or not.

Models	Accuracy (%) no DMPP	Accuracy (%) DMPP	Kappa no DMPP	Kappa DMPP
VGG-19	77.41 ±2.5	80.64 ±2.36	0.61	0.65
ResNext4D	75.98 ±2.55	79.21 ±2.42	0.55	0.60
DenseNet121	67.74 ±2.79	74.55 ±2.6	0.37	0.54
WideResNet50	76.70 ±2.53	78.85 ±2.44	0.59	0.62
EfficientNet-B1	75.26 ±2.58	77.06 ±2.51	0.54	0.57

Table 8
Comparison experiments to analyze the influence of implementing the CM-CNN architecture. The first two columns compare the accuracy obtained when applying or not the CM-CNN architecture. On the other hand, the last two columns compare the kappa index when applying or not the CM-CNN architecture.

Models	Accuracy (%) no CM-CNN	Accuracy (%) CM-CNN	Kappa no CM-CNN	Kappa CM-CNN
VGG-19	77.41 ±2.5	79.21 ±2.42	0.61	0.64
ResNext4D	75.98 ±2.55	79.56 ±2.41	0.55	0.63
DenseNet121	67.74 ±2.79	75.62 ±2.57	0.37	0.58
WideResNet50	76.70 ±2.53	79.21 ±2.42	0.59	0.64
EfficientNet-B1	75.26 ±2.58	77.77 ±2.48	0.54	0.61

3.2. CM-CNN importance

In this section, we analyze the influence of applying the CM-CNN architecture. The first group of models does not implement CM-CNN architecture, while the second one implements the CM-CNN architecture using the corresponding base classifiers. Both groups of models have been trained and evaluated on images without applying the DMPP.

The experiment results showed in Table 8 indicate that modeling each radiologist opinion using the CM-CNN architecture increases the classification capability of the models (statistically significant, p-values < 0.05). Specifically, they achieve an accuracy and kappa index mean improvement of 3.65% and 0.088 points, respectively. The accuracy with CM-CNN is statistically significant higher (p-value $= 0.028$) than accuracy without CM-CNN.

Table 9
Comparison experiments to analyze the influence of applying the RegL framework. The first two columns compare the accuracy obtained when applying or not the RegL. On the other hand, the last two columns compare the kappa index when applying or not the RegL.

Models	Accuracy (%) no RegL	Accuracy (%) with RegL	Kappa no RegL	Kappa with RegL
VGG-19	77.41 ±2.5	81.0 ±2.34	0.61	0.67
ResNext4D	75.98 ±2.55	81.0 ±2.34	0.55	0.66
DenseNet121	67.74 ±2.79	79.21 ±2.42	0.37	0.64
WideResNet50	76.70 ±2.53	81.0 ±2.34	0.59	0.65
EfficientNet-B1	75.26 ±2.58	78.13 ±2.47	0.54	0.62

3.3. Regl importance

Finally, this experiment analyses the influence of applying the entire RegL framework. The first group of models implement the RegL framework, while the second group of models do not.

The experiment results from Table 9 show that applying the RegL framework significantly increases the classification capability of the models (statistically significant, p-values < 0.05). Specifically, they achieve an accuracy and kappa index mean improvement of 5.45% and 0.116 points, respectively. In particular, the increase of the kappa index represents a large increase in performance. The accuracy with RegL is statistically significant higher (p-value $= 0.024$) than accuracy without RegL.

3.4. Convolutional neural network ensemble

Looking for results optimization, we have created an ensemble that averages the output of the five neural networks mentioned. Each network implements the RegL framework.

As aforementioned, we have included class weights into the cross-entropy loss function to solve the categories imbalanced problem. The experimentation demonstrated that networks trained using the "weighted" cross-entropy function perform better on the minority example, to the detriment of the majority ones. We designed the ensemble using three models trained with the weighted cross entropy loss and the other two with the normal one, with optimization purposes.

Table 10 shows that the ensemble reaches an accuracy and kappa index of 84.58% and 0.71 points respectively. Moreover, it is important to highlight the fact that the CM-DenseNet121 achieves similar results but with the drawback that it is unable to classify any examples from the BI-RADS 4 category, therefore, this model would be useless individually.

Fig. 9 shows the confusion matrix for the ensemble predictions and the majority vote. It can be clearly seen that it achieves excellent overall performance in all categories.

We have also contrasted the predictions obtained by the ensemble method and each radiologist's labels. Tables 11, 12 and 13 show a comparison between the ensemble predictions and a radiologist labels using another radiologist's labels as the ground truth.

Table 11 shows that both the ensemble and R2 have similar behaviour in all the categories except on BI-RADS 4. The criteria of R2 in this category is more similar to that of R1 than that of the ensemble.

Table 12 shows that the ensemble method and R3 have similar performance when R2 labels are used as ground truth. Besides, the ensemble method outperforms in some categories.

Table 13 shows that the ensemble method and R3 have similar behaviour when R1 labels are used as ground truth.

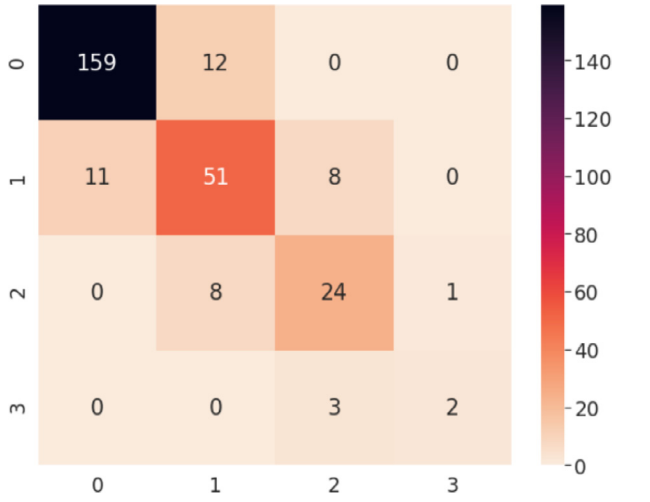


Fig. 9. Confusion matrix obtained from the comparison between the ensemble predictions and the majority vote.

4. Discussion

The experimentation has shown that the implementation of the RegL framework leads to better results. The DMPP leads to an average increase in accuracy and kappa index of 3.4% and 0.05 points. Moreover, the CM-CNN architecture leads to an average increase in accuracy and kappa index of 3.65% and 0.088 points respectively. Both the DMPP and the CM-CNN architecture lead to similar average increases. The kappa index increase from the CM-CNN architecture was slightly better. Finally, the entire RegL framework leads to an average increase in accuracy and kappa index of 5.45% and 0.116 points respectively. The obtained results show that applying the RegL framework significantly improves the density classification task. Besides, the ensemble improves the results obtained by individual models (statistically significant, p-values $<$ 0.05). The ensemble accuracy is statistically significant higher (p-value $= 0.032$) than the accuracy of individual models using RegL.

The comparison of the results of the neural network ensemble against the intra- and inter-reader variabilities shown in Tables 3 and 2 allow us to assure that the ensemble behaves like a radiologist in the task of classifying mammograms according to their breast density. Furthermore, applying the RegL framework reduces the impact of the existing variability in the assessment.

The proposed framework was compared against state-of-the-art breast density classification. All of these studies have in common that they use the entire CC-view mammography to make the classification into the four BI-RADS categories. A summary can be found in Table 14. As shown, the performance of the proposed framework outperformed the state of the arts with an overall accuracy of 84.58%. However, it must be said that each of these studies uses a different data set, which is a limitation of the comparative analysis.

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

1. A preprocess that allows a correct breast segmentation in mammographies with noisy background.
2. A preprocess that adjust the intensities to eliminate problems such as unusual brightness.
3. An intuitive preprocess protocol that normalizes the gray level variability caused by different acquisition devices or different capture processes.

Table 10
Results comparision of the five neural networks that make up the ensemble, along with the results of the ensemble. Each column shows the accuracy of all the networks in a certain BI-RADS category, thus, showing the general accuracy and kappa index in the two last columns.

Model	Class Weights	Accuracy (%) BI-RADS 1 (N=171)	Accuracy (%) BI-RADS 2 (N=70)	Accuracy (%) BI-RADS 3 (N=33)	Accuracy (%) BI-RADS 4 (N=5)	General Accuracy (%) (N=279)	Kappa
CM-VGG-19	Yes	91.22 ±2.16	68.57 ±5.54	57.57 ±8.73	100 ±0.0	81.72 ±2.34	0.68
CM-ResNext4D	Yes	89.47 ±2.34	72.85 ±5.31	72.72 ±7.87	60 ±21.9	82.79 ±2.3	0.69
CM-WideResNet50	Yes	94.73 ±1.7	44.28 ±5.93	78.78 ±7.22	60 ±21.9	79.56 ±2.43	0.62
CM-DenseNet121	No	94.73 ±1.7	67.14 ±5.61	78.78 ±7.22	0 ±0.0	84.22 ±2.19	0.70
CM-EfficientNet-B1	No	91.81 ±2.09	75.71 ±5.12	42.42 ±8.73	0 ±0.0	80.28 ±2.39	0.62
CM-Ensemble		**92.98** ±1.95	**72.85** ±5.31	**72.72** ±7.87	**40** ±21.9	**84.58** ±2.19	**0.71**

Table 11
Comparison between the predictions from the ensemble and R2 labels, considering the R1 labels as the ground truth.

Category	R2 Recall (%)	Ensemble Recall (%)	R2 Precision (%)	Ensemble Precision (%)
BI-RADS 1	96.0	94.0	93.0	92.0
BI-RADS 2	65.0	62.0	79.0	72.0
BI-RADS 3	60.0	57.0	60.0	57.0
BI-RADS 4	100.0	67.0	61.0	22.0

Table 12
Comparison between the predictions from the ensemble and R3 labels, considering the R2 labels as the ground truth.

Category	R3 Recall (%)	Ensemble Recall (%)	R3 Precision (%)	Ensemble Precision (%)
BI-RADS 1	91.0	91.0	87.0	91.0
BI-RADS 2	59.0	65.0	58.0	62.0
BI-RADS 3	52.0	66.0	49.0	66.0
BI-RADS 4	9.0	33.0	100.0	100.0

Table 13
Comparison between the predictions from the ensemble and R3 labels, considering the R1 labels as the ground truth.

Category	R3 Recall (%)	Ensemble Recall (%)	R3 Precision (%)	Ensemble Precision (%)
BI-RADS 1	94.0	94.0	87.0	92.0
BI-RADS 2	56.0	62.0	67.0	72.0
BI-RADS 3	58.0	57.0	54.0	57.0
BI-RADS 4	45.0	67.0	56.0	22.0

Table 14
Accuracy of different breast density classification studies.

Study	Year	No. of Images	Accuracy (%)
Lee and Nishikawa [50]	2018	455	81.0
LIBRA [59]	2012	324	81.0
Lehman et al. [48]	2018	41479	77.0
Our proposed method	2021	892	84.58

4. The implementation of a convolution-based architecture capable of modeling multiple radiologists' opinions, thus reducing the existing variability caused by the noisy labels.
5. A breast density classification framework with similar behavior to that achieved by expert radiologists at classifying the breast density into the BI-RADS categories. The framework could even be useful to standarize the way of evaluating breast density.

4.1. Limitations and future research

The main limitation of this work is that the mammograms used were labeled according BI-RADS 4th edition, which uses percentage ranges to define each category. Currently, BI-RADS 5th edition is *de facto* standard for mammography density classification, which takes into account not only the percentage of dense tissue, but also its distribution throughout the breast. It is expected that the trained models would get poorer results when tested with test images labelled with BI-RADS 5th edition. It would be interesting, therefore, to train and test the proposed models with images labelled according to the latest edition of BI-RADS.

Also, it would be interesting to include mediolateral oblique (MLO) mammograms into the corpus of images, together with images from other caption devices and labelled by other radiologists.

Finally, it would also be interesting to carry out a multicenter study in which several radiologists participate. Once the methodology is validated, it would be interesting to develop a model capable of estimating the breast cancer risk from a set of dense tissue features and other variables. This tools could help in the radiologists decision making.

5. Conclusion

Nowadays, the amount of mammograms that must be analyzed in the screening centers is constantly growing due to the breast cancer early detection programs. This increase in the number of images means more work for the radiologists who analyze the mammograms. In this sense, the availability of a tool that provides automatic classification of dense tissue on digital mammograms becomes crucial.

The work presented in this paper provides an automatic framework aimed at breast density classification and based on deep learning. The proposed methodology: (1) removes the noise in the mammogram background allowing a better breast segmentation, (2) removes unusual brightness problems that influence the image quality, (3) standarizes the grey-level variability of the breast pixels, and finally (4) performs a dense tissue classification into the BI-RADS categories.

Fixing a radiologist's labels as the ground truth, both the precision and the recall of the proposed framework are close to those obtained by other radiologists. These results mean that by applying the framework, the results are similar to those of an experienced radiologist. The validation of the methodology would allow not only to automate, but also to reliably standardize the density reading.

Funding

This work was partially funded by Generalitat Valenciana through IVACE (Valencian Institute of Business Competitiveness) distributed nominatively to Valencian technological innovation centres under project expedient IMAMCN/2021/1.

Ethics approval and consent to participate

This study was approved by the Research Ethics Committee of the Instituto de Salud Carlos III (project name: "Determinantes de la densidad mamográfica en las mujeres participantes de los programas de detección pre- coz del cáncer de mama en España *DDM-Spain*") and consent was obtained from study participants at the time of screening.

Acknowledgements

The authors of this work would like to thank to Inmaculada Martínez for her participation in the image labeling process.

References

[1] N. Ascunce, J. Delfrade, D. Salas, R. Zubizarreta, M. Ederra, Programas de detección precoz de cáncer de mama en españa: características y principales resultados, Medicina Clínica 141 (1) (2013) 13–23.
[2] S.E. de Oncología Médica, Manual de prevencion y diagnostico precoz, del cancer (2020). https://seom.org/manual-prevencion/128/.
[3] E.A. B. L. Sickles, C.J. D'Orsi, ACR BI-RADS mammography. in: ACR BI-RADS atlas, Breast Imaging Reporting and Data System, American College of Radiology (2013).
[4] A. Alikhassi, H.E. Gourabi, M. Baikpour, Comparison of inter- and intra-observer variability of breast density assessments using the fourth and fifth editions of breast imaging reporting and data system, European J Radiol Open 5 (2018) (2018) 67–72, doi:10.1016/j.ejro.2018.04.002.
[5] W. Alomaim, D. O'Leary, J. Ryan, L. Rainford, M. Evanoff, S. Foley, Variability of breast density classification between us and uk radiologists, J Med Imaging Radiat Sci 50 (1) (2019) 53–61, doi:10.1016/j.jmir.2018.11.002.
[6] A. Aloufi, A. Alnaeem, A. Almousa, K. Alzimami, A. Alfuraih, B. Alshahrani, M. Zayed, T. Almashouq, S. Alnasser, K. Aldossari, M. Alomrani, I. Alzahrani, E. Harkness, S. Astley, Breast density in saudi arabia: intra and inter reader variability in screening mammograms assessed visually using bi-rads, and visual analogue scales 11316 (2020), doi:10.1117/12.2548758.
[7] E. Ali, M. Raafat, Relationship of mammographic densities to breast cancer risk, Egypt. J. Radiol. Nucl. Med. 52 (1) (2021), doi:10.1186/s43055-021-00497-y.
[8] S. Advani, W. Zhu, J. Demb, B. Sprague, T. Onega, L. Henderson, D. Buist, D. Zhang, J. Schousboe, L. Walter, K. Kerlikowske, D. Miglioretti, D. Braithwaite, Association of breast density with breast cancer risk among women aged 65 years or older by age group and body mass index, JAMA Network Open 4 (8) (2021), doi:10.1001/jamanetworkopen.2021.22810.
[9] A. Mokhtary, A. Karakatsanis, A. Valachis, Mammographic density changes over time and breast cancer risk: a systematic review and meta-analysis, Cancers (Basel) 13 (19) (2021), doi:10.3390/cancers13194805.
[10] E. Kim, Y. Chang, J. Ahn, J.-S. Yun, Y. Park, C. Park, H. Shin, S. Ryu, Mammographic breast density, its changes, and breast cancer risk in premenopausal and postmenopausal women, Cancer 126 (21) (2020) 4687–4696, doi:10.1002/cncr.33138.
[11] M. Pollán, R. Llobet, J. Miranda-García, J. Antón, M. Casals, I. Martínez, C. Palop, F. Ruiz-Perales, C. Sánchez-Contador, C. Vidal, B. Pérez-Gómez, D. Salas-Trejo, Validation of DM-scan, a computer-assisted tool to assess mammographic density in full-field digital mammograms, Springerplus 2 (1) (2013), doi:10.1186/2193-1801-2-242.
[12] I.M. Gómez, M.C.e. Busto, J.A. Guirao, F.R. Perales, R.L. Azpitarte, Estimación semiautomática de la densidad mamaria con DM-scan, Radiología 56 (5) (2014) 429–434, doi:10.1016/j.rx.2012.11.007.
[13] Z. Gandomkar, K. Tay, W. Ryder, P. Brennan, C. Mello-Thoms, Idensity: an automatic gabor filter-based algorithm for breast density assessment, Progress in Biomedical Optics and Imaging - Proceedings of SPIE 9416 (2015), doi:10.1117/12.2083149.
[14] A. Rampun, B. Scotney, P. Morrow, H. Wang, J. Winder, Breast density classification using local quinary patterns with various neighbourhood topologies, Journal of Imaging 4 (2018) 14, doi:10.3390/jimaging4010014.
[15] Y. Shin, I. Balasingham, Comparison of hand-craft feature based svm and CNN based deep learning framework for automatic polyp classification, 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2017) 3277–3280, doi:10.1109/EMBC.2017.8037556.
[16] S.K. Baranwal, K. Jaiswal, K. Vaibhav, A. Kumar, R. Srikantaswamy, Performance analysis of brain tumour image classification using CNN and SVM, 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) 2020 (2020) 537–542, doi:10.1109/ICIRCA48905.2020.9183023.
[17] M. Hasan, S. Ullah, M.J. Khan, K. Khurshid, Comparative analysis of SVM, ANN and CNN for classifying vegetation species using hyperspectral thermal infrared data, ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4213 (2019) (2019) 1861–1868, doi:10.5194/isprs-archives-XLII-2-W13-1861-2019.
[18] J. Janai, F. Güney, A. Behl, A. Geiger, Computer vision for autonomous vehicles: problems, datasets and state-of-the-art, Found. Trends Comput. Graph. Vis. 12 (2020) (2020) 1–308, doi:10.1561/0600000079.
[19] D. Pop, A. Rogozan, C. Chatelain, F. Nashashibi, A. Bensrhair, Multi-task deep learning for pedestrian detection, action recognition and time to cross prediction, IEEE Access PP 2019 (2019), doi:10.1109/ACCESS.2019.2944792. 1–1
[20] Y. Hu, D. Zhang, G. Cao, Q. Pan, Network data analysis and anomaly detection using CNN technique for industrial control systems security, 2019 IEEE international conference on systems, Man and Cybernetics (SMC) 2019 (2019) 593–597, doi:10.1109/SMC.2019.8913895.
[21] L. Haochen, Z. Bin, S. Xiaoyong, Z. Yongting, Cnn-based model for pose detection of industrial, PCB 2017 (10) (2017) 390–393, doi:10.1109/ICICTA.2017.93.
[22] J. Li, J. Chen, B. Sheng, P. Li, P. Yang, D. Feng, J. Qi, Automatic detection and classification system of domestic waste via multimodel cascaded convolutional neural network, IEEE Trans. Ind. Inf. 18 (1) (2022) 163–173, doi:10.1109/TII.2021.3085669.
[23] E. Luz, P. Silva, R. Silva, L. Silva, J. Guimarães, G. Miozzo, G. Moreira, D. Menotti, Towards an effective and efficient deep learning model for COVID-19 patterns detection in x-ray images, Research on Biomedical Engineering (2021), doi:10.1007/s42600-021-00151-6.
[24] O. Yadav, K. Passi, C. Jain, Using deep learning to classify x-ray images of, potential tuberculosis patients 2018 (2018) 2368–2375, doi:10.1109/BIBM.2018.8621525.
[25] J. Zhao, M. Li, W. Shi, Y. Miao, Z. Jiang, B. Ji, A deep learning method for classification of chest x-ray images, J. Phys. Conf. Ser. 1848 (1) (2021) 012030, doi:10.1088/1742-6596/1848/1/012030.
[26] B. Wang, K. Yager, D. Yu, M. Hoai, X-Ray scattering image classification using deep learning, 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) 2017 (2017) 697–704, doi:10.1109/WACV.2017.83.
[27] J. Seah, C. Tang, Q. Buchlak, X. Holt, J. Wardman, A. Aimoldin, N. Esmaili, H. Ahmad, H. Pham, J. Lambert, B. Hachey, S. Hogg, B. Johnston, C. Bennett, L. Oakden-Rayner, P. Brotchie, C. Jones, Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study, The Lancet Digital Health 3 (2021), doi:10.1016/S2589-7500(21)00106-0.
[28] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) (2017) 115–118, doi:10.1038/nature21056.
[29] D. Karimi, H. Dou, S.K. Warfield, A. Gholipour, Deep learning with noisy labels: exploring techniques and remedies in medical image analysis, Med Image Anal 65 (2020) (2020) 101759, doi:10.1016/j.media.2020.101759.
[30] H. Song, M. Kim, D. Park, J. Lee, Learning from noisy labels with deep neural networks: a survey, CoRR abs/2007.08199 (2020).
[31] S. Liu, J. Niles-Weed, N. Razavian, C. Fernandez-Granda, Early-learning regularization prevents memorization of noisy labels, Advances in Neural Information Processing Systems 2020-December (2020).
[32] K. Gu, X. Masotto, V. Bachani, B. Lakshminarayanan, J. Nikodem, D. Yin, A realistic simulation framework for learning with label noise, CoRR abs/2107.11413 (2021).
[33] C. Northcutt, L. Jiang, I. Chuang, Confident learning: estimating uncertainty in dataset labels, J. Artif. Int. Res. 70 (2021) (2021) 1373–1411, doi:10.1613/jair.1.12125.
[34] B. Pérez-Gómez, F. Ruiz, I. Martínez, M. Casals, J. Miranda, C. Sánchez-Contador, C. Vidal, R. Llobet, M. Pollán, D. Salas, Women's features and inter-/intra-rater agreement on mammographic density assessment in full-field digital mammograms (DDM-SPAIN), Breast Cancer Res. Treat. 132 (1) (2011) 287–295, doi:10.1007/s10549-011-1833-3.
[35] S. Saeed, I. Masroor, H. Iqbal, S. Sufian, M. Awais, Variability of breast den-

sity assessment and the need for additional imaging: a comparison between computed mammography and digital mammography, Journal of the College of Physicians and Surgeons Pakistan 30 (11) (2020) 1213–1216, doi:10.29271/jcpsp.2020.11.1213.
[36] A. Gemici, E. Bayram, E. Hocaoglu, E. Inci, Comparison of breast density assessments according to bi-rads 4th and 5th editions and experience level, Acta Radiol Open 9 (2020), doi:10.1177/2058460120937381.
[37] B. Sprague, K. Kerlikowske, E. Bowles, G. Rauscher, C. Lee, A. Tosteson, D. Miglioretti, Trends in clinical breast density assessment from the breast cancer surveillance consortium, J. Natl. Cancer Inst. 111 (6) (2019) 629–632, doi:10.1093/jnci/djy210.
[38] N. Boyd, J. Byng, R. Jong, E. Fishell, L. Little, A. Miller, G. Lockwood, D. Tritchler, M. Yaffe, Quantitative classification of mammographic densities and breast cancer risk: results from the canadian national breast screening study, J. Natl. Cancer Inst. 87 (9) (1995) 670–675, doi:10.1093/jnci/87.9.670.
[39] 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), doi:10.1016/j.patcog.2017.04.018.
[40] K. Wu, E. Otoo, K. Suzuki, Optimizing two-pass connected-component labeling algorithms, Pattern Analysis & Applications 12 (2009) (2009) 117–135, doi:10.1007/s10044-008-0109-y.
[41] L. He, Y. Chao, K. Suzuki, K. Wu, Fast connected-component labeling, Pattern Recognit 42 (9) (2009) 1977–1987, doi:10.1016/j.patcog.2008.10.013.
[42] J. Boita, R.E. van Engen, A. Mackenzie, A. Tingberg, H. Bosmans, A. Bolejko, S. Zackrisson, M.G. Wallis, D.M. Ikeda, C.V. Ongeval, R. Pijnappel, M. Broeders, I. Sechopoulos, F. Jansen, L. Duijm, H. de Bruin, I. Andersson, C. Behmer, K. Taylor, F. Kilburn-Toppin, M. van Goethem, R. Prevos, S.P. N. Salem, How does image quality affect radiologists' perceived ability for image interpretation and lesion detection in digital mammography? Eur Radiol 31 (7) (2021) 5335–5343, doi:10.1007/s00330-020-07679-8.
[43] A. Abdelrahim, W. Berg, H. Peng, Y. Luo, R. Jankowitz, S. Wu, A deep learning method for classifying mammographic breast density categories, Med Phys 45 (2017), doi:10.1002/mp.12683.
[44] N. Wu, K. Geras, Y. Shen, J. Su, S. Kim, E. Kim, S. Wolfson, L. Moy, K. Cho, Breast density classification with deep convolutional neural, networks 2018 (2018) 6682–6686, doi:10.1109/ICASSP.2018.8462671.
[45] Z. Gandomkar, M. Suleiman, D. Demchig, P. Brennan, M. McEntee, Bi-rads density categorization using deep neural, networks 22 (2019) 2019, doi:10.1117/12.2513185.
[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009) 248–255, doi:10.1109/CVPR.2009.5206848.
[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016 (2016) 770–778, doi:10.1109/CVPR.2016.90.
[48] C. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, R. Barzilay, Mammographic breast density assessment using deep learning: clinical implementation, Radiology 290 (2018) 180694, doi:10.1148/radiol.2018180694.
[49] N. Saffari, H. Rashwan, M. Abdel-Nasser, V.K. Singh, M. Arenas, E. Mangina, B. Herrera, D. Puig, Fully automated breast density segmentation and classification using deep learning, Diagnostics 10 (2020), doi:10.3390/diagnostics10110988.
[50] J. Lee, R.M. Nishikawa, Automated mammographic breast density estimation using a fully convolutional network, Med Phys 45 (2018) (2018) 1178–1190.
[51] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. Alexander, N. Silberman, Learning from noisy labels by regularized estimation of, annotator confusion 2019 (2019) 11236–11245, doi:10.1109/CVPR.2019.01150.
[52] Y.-X. Li, Y. Chai, Y.-Q. Hu, H.P. Yin, Review of imbalanced data classification methods, Kongzhi yu Juece/Control and Decision 34 (4) (2019) 673–688, doi:10.13195/j.kzyjc.2018.0865.
[53] Y. Feng, M. Zhou, X. Tong, Imbalanced classification: a paradigm-based review, Stat Anal Data Min 14 (5) (2021) 383–406, doi:10.1002/sam.11538.
[54] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556 (2015).
[55] S. Xie, R.B. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017 (2017) 5987–5995, doi:10.1109/CVPR.2017.634.
[56] G. Huang, Z. Liu, K.Q. Weinberger, Densely connected convolutional networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017 (2017) 2261–2269, doi:10.1109/CVPR.2017.243.
[57] S. Zagoruyko, N. Komodakis, Wide residual networks, in: in: Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, 2016. Pp. 87.1–87.12
[58] M. Tan, Q. Le, K. Chaudhuri, R. Salakhutdinov, Efficientnet: rethinking model scaling for convolutional neural networks, Proceedings of the 36th International Conference on Machine Learning,Vol. 97 of Proceedings of Machine Learning Research, PMLR 2019 (2019) 6105–6114.
[59] B.M. Keller, D. Nathan, Y. Wang, Y. Zheng, J.C. Gee, E.F. Conant, D. Kontos, Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy C-means clustering and support vector machine segmentation, Med Phys 39 (8) (2012) 4903–4917.