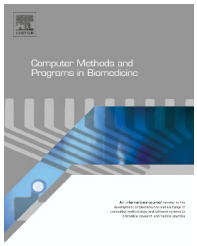
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE XXX (2016) XXX–XXX

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Representation learning for mammography mass lesion classification with convolutional neural networks

John Arevalo[a], *Fabio A. González*[a], *Raúl Ramos-Pollán*[b], *Jose L. Oliveira*[c], *Miguel Angel Guevara Lopez*[d,*]

[a] *Universidad Nacional de Colombia, Bogotá, Colombia*
[b] *Universidad Industrial de Santander, Bucaramanga, Colombia*
[c] *DETI-IEETA, Universidade de Aveiro, Portugal*
[d] *CCG, Computer Graphics Center, Portugal*

ARTICLE INFO

Article history:
Received 8 July 2015
Received in revised form
18 December 2015
Accepted 21 December 2015

Keywords:
Breast cancer
Feature learning
Convolutional neural networks
Computer-aided diagnosis
Mammography

ABSTRACT

Background and objective: The automatic classification of breast imaging lesions is currently an unsolved problem. This paper describes an innovative representation learning framework for breast cancer diagnosis in mammography that integrates deep learning techniques to automatically learn discriminative features avoiding the design of specific hand-crafted image-based feature detectors.

Methods: A new biopsy proven benchmarking dataset was built from 344 breast cancer patients' cases containing a total of 736 film mammography (mediolateral oblique and craniocaudal) views, representative of manually segmented lesions associated with masses: 426 benign lesions and 310 malignant lesions. The developed method comprises two main stages: (i) preprocessing to enhance image details and (ii) supervised training for learning both the features and the breast imaging lesions classifier. In contrast to previous works, we adopt a hybrid approach where convolutional neural networks are used to learn the representation in a supervised way instead of designing particular descriptors to explain the content of mammography images.

Results: Experimental results using the developed benchmarking breast cancer dataset demonstrated that our method exhibits significant improved performance when compared to state-of-the-art image descriptors, such as histogram of oriented gradients (HOG) and histogram of the gradient divergence (HGD), increasing the performance from 0.787 to 0.822 in terms of the area under the ROC curve (AUC). Interestingly, this model also outperforms a set of hand-crafted features that take advantage of additional information from segmentation by the radiologist. Finally, the combination of both representations, learned and hand-crafted, resulted in the best descriptor for mass lesion classification, obtaining 0.826 in the AUC score.

Conclusions: A novel deep learning based framework to automatically address classification of breast mass lesions in mammography was developed.

* *Corresponding author*. Tel.: +351 253510580.
E-mail addresses: jearevaloo@unal.edu.co (J. Arevalo), fagonzalezo@unal.edu.co (F.A. González), rramosp@uis.edu.co (R. Ramos-Pollán), jlo@ua.pt (J.L. Oliveira), mguevara@ccg.pt (M.A. Guevara Lopez).

http://dx.doi.org/10.1016/j.cmpb.2015.12.014
0169-2607/

1. Introduction

Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012 (second most common cancer overall); this represents about 12% of all new cancer cases and 25% of all cancers in women.[1] Breast cancer has a known asymptomatic phase that can be detected with mammography, and therefore, mammography is the primary imaging modality for screening. Double-reading (two radiologists independently read the same mammograms) has been advocated to reduce the proportion of missed cancers and it is currently included in most screening programs [1]. However, double-reading incurs in additional workload and costs. Alternatively, computer-aided diagnosis (CADx) systems can assist a single radiologist when reading mammograms providing support for their decisions. These systems can be used as second opinion criteria by radiologists, playing a key role in the early detection of breast cancer and helping to reduce the death rate among women with breast cancer in a cost-effective manner [2].

A successful approach to build CADx systems is to use machine learning classifiers (MLC). MLC are learned from a set of labeled data samples capturing complex relationships in the data [3–5]. In order to train a MLC for breast cancer diagnosis, a set of features describing the image is required. Ideally, features should have high discriminant power that allows inferring whether a given image is from a malignant finding or not. This is, however, a challenging topic that has gathered the focus of research in several sciences, from medicine to computer vision. Thus, several types of features may be used to infer the diagnosis. Many CADx systems use hand-crafted features based on prior knowledge and expert guidance. In particular, strategies based on feature selection [6] and hand-crafted features that characterize geometry and textures [7] has been proposed for mass classifications. As an alternative, the use of machine learning strategies to learn good features directly from the data is a new paradigm that has shown successful results in different computer vision tasks. One such paradigm is *deep learning*.

Deep learning methods have been widely applied in recent years to address several computer perception tasks [8]. Their main advantage lies in avoiding the design of specific feature detectors. In turn, deep learning models look for a set of transformations directly from the data. This approach has had remarkable results, particularly in computer vision problems such as natural scene classification and object detection [9]. Deep learning models have also been adapted to different medical tasks such as tissue classification in histology and histopathology images [10,11], Alzheimer disease diagnosis [12–15], and knee cartilage segmentation [16] among others.

However, only few works have explored deep learning methods to address the automatic classification of identified lesions in mammography images [17]. In [18] stacked deep auto-encoders were used to estimate breast density score using multiscale features. Lately, this has been extended by including breast tissue segmentation and scoring of mammographic texture [19] with a convolutional neural network (CNN) model. CNN model is the most successful deep learning strategy applied to image understanding [9]. In [20,21] CNNs are used as representation strategy to characterize microcalcifications. Finally, the most recent work developed in this area was done in [22] which uses Adaptive Deconvolutional Networks to learn the representation in order to classify malign/benign breast lesions. Such strategy was evaluated on 245 lesions in a bootstrap fashion, reporting the area under the ROC curve (AUC) $AUC = 0.71$. In this work, we also use convolutional architectures, however the features are learned in a supervised way during CNN training, taking advantage of expert knowledge represented by previously identified lesions in breast imaging, manually segmented by expert radiologists in both mammographic views (mediolateral oblique and craniocaudal).

The remainder of the paper is organized as follows: Section 2 describes the proposed approach to perform classification of identified lesions in mammography images. Section 3 details the experimental setup used to evaluate the proposed approach. Finally, Sections 4 and 5 show results and present the main conclusions of this work.

2. Material and methods

2.1. *Breast cancer digital repository*

The benchmarking dataset used in this study is available on the Breast Cancer Digital Repository (BCDR).[2] BCDR is a wide-ranging annotated public repository composed of Breast Cancer patient' cases in the northern region of Portugal. The BCDR is subdivided in two different repositories: (1) a Film Mammography-based Repository (BCDR-FM) and (2) a Full Field Digital Mammography-based Repository (BCDR-DM). Both repositories were created with anonymous cases from medical archives (complying with current privacy regulations as they are also used to teach regular and postgraduate medical students) supplied by the Faculty of Medicine – Centro Hospitalar São João, at University of Porto (FMUP–HSJ). BCDR provides normal and annotated patient cases of breast cancer including mammography lesions outlines, anomalies observed by radiologists, pre-computed image-based descriptors and related clinical data. The BCDR-FM is composed by 1010 patient cases (998 female and 12 male, with ages between 20 and 90 years old), including 1125 studies, 3703 mediolateral oblique (MLO) and craniocaudal (CC) mammography incidences and 1044 identified lesions clinically described (820 already identified in MLO and/or CC views). With this, 1517 segmentations were manually made and BI-RADS classified by specialized radiologists. MLO and CC images are grey-level digitized mammograms with a resolution of 720 (width) by 1168 (height) pixels and a bit depth of 8 bits per pixel, saved in the TIFF format. The BCDR-DM, still in construction, at the time of writing is composed by 724 Portuguese patient cases (723 female and 1 male, with ages between 27 and 92 years old), including 1042 studies, 3612 MLO and/or CC mammography

[1] World Cancer Research Fund International http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics, Accessed May 20, 2015

[2] http://bcdr.inegi.up.pt

Please cite this article in press as: J. Arevalo, et al., Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. (2016), http://dx.doi.org/10.1016/j.cmpb.2015.12.014

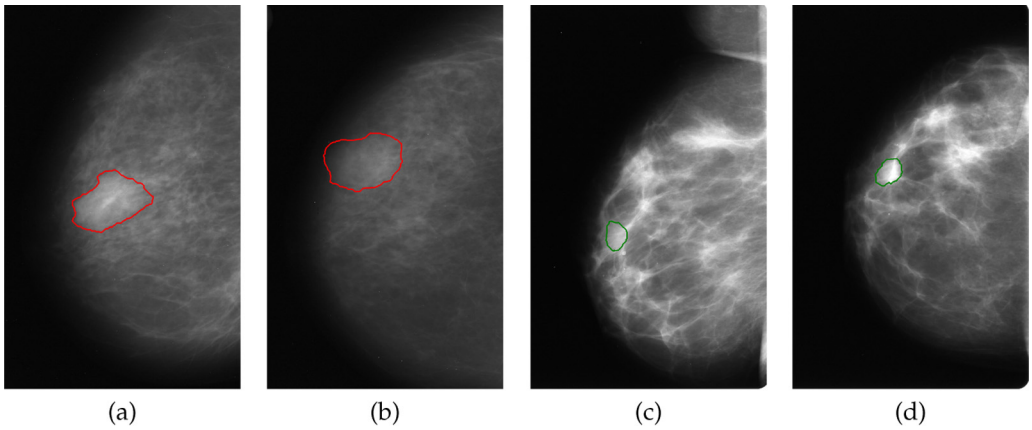


Fig. 1 – Samples of lesions presented in the dataset. Malignant lesion in (a) oblique view and (b) craneo-caudal view. Benign lesion in (c) oblique view and (d) craneo-caudal view.

incidences and 452 lesions clinically described (already identified in MLO and CC views). With this, 818 segmentations were manually made and BI-RADS classified by specialized radiologists. The MLO and CC images are grey-level mammograms with a resolution of 3328 (width) by 4084 (height) or 2560 (width) by 3328 (height) pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). The bit depth is 14 bits per pixel and the images are saved in TIFF format. As described below, this work is focused on the BCDR-FM Repository.

2.1.1. *Benchmarking dataset*

A new dataset of the BCDR-FM repository has been made publicly available, at http://bcdr.inegi.up.pt, for comparison and research reproducibility purposes. The 8-bit resolution "Film Mammography Dataset Number 3" (BCDR-F03) was built as a subset of the BCDR-FM and it is composed of 344 patients with 736 film images containing 426 benign mass lesions and 310 malign mass lesions, including clinical data and image-based descriptors. Such lesions are associated with masses. The motivations to choose 8-bit resolution images over 12-bit or 14-bit are twofold: Firstly, in contrast to the BCDR-DM (currently under construction), almost all lesions in the BCDR-FM repository have a proven biopsy; and secondly, digital mammography (high resolution images) are not as widely available as film mammography images since the former are more expensive to acquire [23]. For all the experimentation clinical data were not included as features. Fig. 1 shows examples of both classes with their respective segmentations. The dataset contains MLO and CC views.

2.1.2. *Baseline descriptors*

Based on the systematic evaluation presented by Moura et al. [3], the histogram of oriented gradients (HOG) and the histogram of gradient divergence (HGD) were selected as descriptors for our baseline since they showed the best performance against other traditional descriptors. Additionally, a set of 17 hand-crafted features extracted from the segmented lesions (representative of shape, texture and intensities of the mammograms) are used for comparative purposes.

Table 1 – Set of hand-crafted features. For details see [3].

Type	Features
Intensities	Mean, median, maximum, minimum, standard deviation, skewness, kurtosis
Shape	Area, perimeter, circularity, elongation, y_center_mass, x_center_mass, form
Textures	Contrast, correlation, entropy

Hand-crafted features (HCfeats): HCfeats is a set comprising 17 features selected from produced sets of high performance features proposed by Perez et al. [24] that demonstrated a high impact in characterizing lesions corresponding to masses. Table 1 lists the features and their description. HCfeats is composed by intensity descriptors computed directly from the grey-levels of the pixels inside the lesion' s contour identified by the radiologists; texture descriptors computed from the grey-level co-occurrence matrix related to the bounding box of the lesion' s contour; and shape descriptors computed from the lesion' s contour. Notice that computing this set of features requires not only the region of interest (ROI) detection, but also the manual segmentation provided by the expert.

Histogram of oriented gradients (HOG): HOG describes images through the distribution of the gradients. Images are divided into a grid of blocks (e.g. 3×3), and each block is described by a histogram of the orientation of the gradient. Each histogram has a predefined number of bins dividing the range of possible orientations (from 0 to 2π radians, or from 0 to π radians), and the value of each bin is calculated by summing the magnitude of the gradient of the pixels which have gradient direction within the limits of the bin.

Histogram of gradient divergence (HGD): Gradient divergence in a point i, j is measured as the angle between the vector of the intensity gradient on i, j and a vector pointing to the center of the image with origin in i, j. HGD describes images through the distribution of the gradient divergence. Images are divided into concentric regions, and each region is described by a histogram of the gradient divergence.

2.2. *Proposed method*

Image representation is fundamental for automatic classification of lesions in mammography images. The goal is to

Please cite this article in press as: J. Arevalo, et al., Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. (2016), http://dx.doi.org/10.1016/j.cmpb.2015.12.014

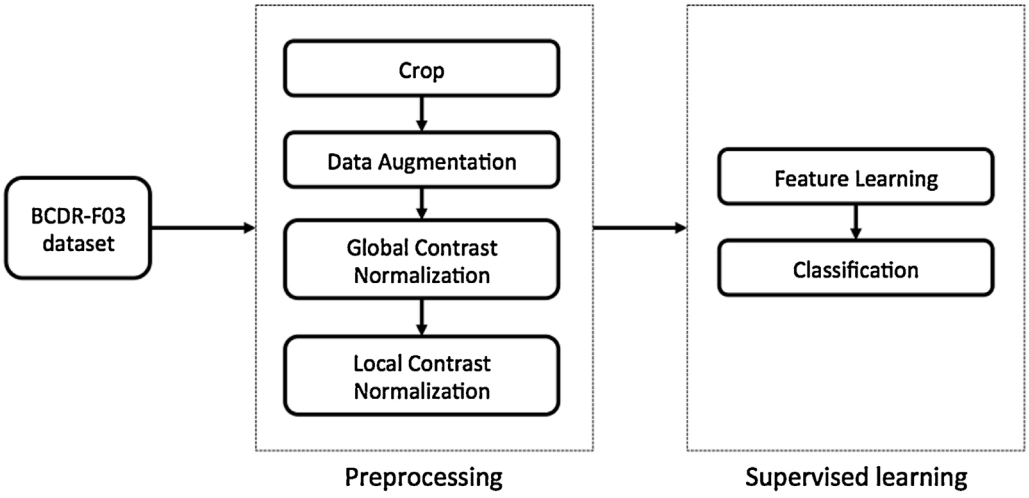


Fig. 2 – Workflow diagram of the proposed method.

describe the content of the image in a compact and discriminative way. Traditional CADx systems represent images with a carefully selected set of mathematical and heuristic features aiming to characterize the lesion. Recent studies have replaced this hand-crafted process with a learning-based approach where a model is trained in an unsupervised way using deep learning, to transform the raw pixels in a set of features that feeds a classifier algorithm [22,19]. In contrast to previous work, we have herewith applied a hybrid approach in which CNNs are used to learn the representation in a supervised way. That is, we used lesions previously classified (labeled as benign or malignant) to guide the feature learning process.

Fig. 2 shows an overall view of the proposed method. It comprises two main stages: preprocessing and supervised training. The *preprocessing* stage aims to prepare the data in better conditions through a set of transformations so that the next stage takes advantage of relevant characteristics. *Supervised learning* is the second stage that involves two processes: feature learning and classification training. *Feature learning* is performed by training a CNN. It is noteworthy that feature learning is a supervised stage since the CNN training is guided by the labeled samples. The final stage is the SVM *classifier training* with the penultimate layer of the CNN as features.

2.2.1. *Preprocessing*

Preprocessing is a common stage in CADx systems. Its main goal is to enhance the characteristics of the image by applying a set of transformations that could help to improve performance in following stages. The first step in this work is to extract the ROI from the image. Secondly, an oversampling strategy is used to both get more samples artificially and help to prevent overfitting during training. Finally, a normalization process is carried out to prepare data for learning algorithms. It is widely known that feature learning and deep learning methods usually perform better when the input data has some properties such as decorrelation and normalization, mainly because such properties help gradient-based optimization techniques to converge [25].

Cropping: CADx systems aim at classifying a previously identified ROI in the whole film image. This ROI can be obtained by a manual segmentation or automatically detected by a computer aided detection system. Because of lesions in BCDR-03 dataset are manually segmented, we fixed the input size to ROIs of $r \times r$ pixels. With this, ROIs can be easily extracted by taking the bounding box of the segmented region. Specifically, images were cropped to the bounding box of the lesions and rescaled to $r \times r$ pixels preserving the aspect ratio when either width or height of the bounding box are greater than r, otherwise the lesion is centered without scaling and preserving the surrounding region.

Data augmentation: The expressiveness of neural network models, and particularly deep ones, comes mainly from the large number of parameters to learn. However, more complex models also increase the chance of overfitting the training data. Data augmentation is a good way to help to prevent this behavior [26]. Data augmentation is the process of artificially create new samples by applying transformations to the original data. In a lesion classification problem, data augmentation makes sense because a lesion can be presented in any particular orientation. Thus, the model also should be able to learn from such transformations. In particular, For each training image, we have artificially generated 7 new label-preserving samples using a combination of flipping and 90,180 and 270 degrees rotation transformations.

Global contrast normalization: Due to the digitalization process, the lighting conditions between different film images will be different, and all pixel values of the image are affected by that. A common way to overcome this effect, is to perform a global contrast normalization (GCN) by subtracting the mean of the intensities in the image to each pixel. Notice that the mean is not calculated per pixel, but per image, so it is perfectly fine to subtract it without worrying about whether the current image belongs to train, validation, or test set. Let $\mathbf{X} \in \mathbb{R}^{r \times r}$ be the image, the element-wise transformation is

$$\hat{\mathbf{X}}_{i,j} = \mathbf{X}_{i,j} - \bar{x} \tag{2.1}$$

with $\bar{x} \in \mathbb{R}; \bar{x} = \frac{1}{r^2}\sum_{i,j}\mathbf{X}_{i,j}$, the mean of the $\mathbf{X}$ image intensities, and $\mathbf{X}_{i,j} \in \mathbb{R}$ the intensity in the i, j pixel.

Please cite this article in press as: J. Arevalo, et al., Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. (2016), http://dx.doi.org/10.1016/j.cmpb.2015.12.014

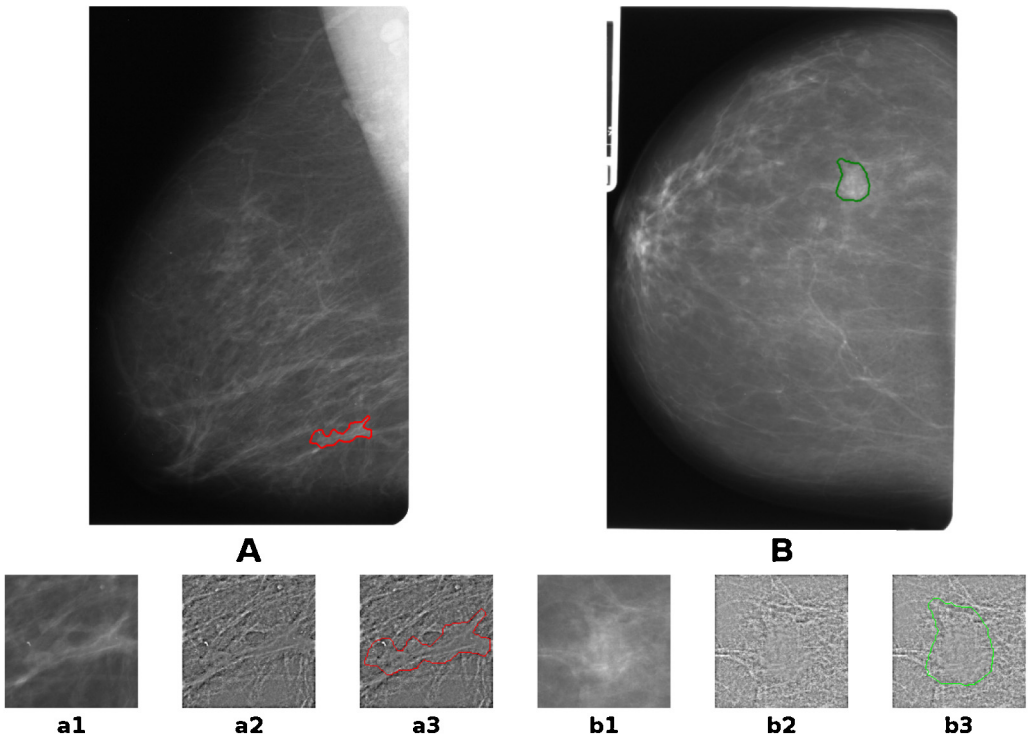


Fig. 3 – Mammography images after the preprocessing step. Images A and B represent malignant and benign lesions respectively. Images a1 and b1 are the bounding box of the lesions. Images a2 and b2 show the output of global and local contrast normalizations. Images a3 and b3 show outline of the lesions over the normalized images.

Local contrast normalization: Local contrast normalization (LCN) is a transformation inspired by computational neuroscience models [27]. Its main idea is to mimic the behavior the V1 visual cortex. It is implemented by defining a $\mathbf{G} \in \mathbb{R}^{k\times k}$ normalized Gaussian window, i.e $\sum_{p,q}\mathbf{G}_{p,q} = 1$. Then, for each pixel in the global contrast normalized image $\hat{\mathbf{X}}$, the mean of its $k \times k$ neighborhood is removed:

$$\mathbf{V}_{i,j} = \hat{\mathbf{X}}_{i,j} - \sum_{p,q} \mathbf{G}_{p,q} \cdot \hat{\mathbf{X}}_{i+p,j+q} \tag{2.2}$$

with $\mathbf{V} \in \mathbb{R}^{k\times k}$ as the local normalized patch. Then the norm of each $k \times k$ neighborhood is scaled to 1 when it is greater than 1:

$$\tilde{\mathbf{X}}_{i,j} = \frac{\mathbf{V}_{i,j}}{\max\left(c, \sigma_{i,j}\right)} \tag{2.3}$$

where $\sigma_{i,j} \in \mathbb{R}; \sigma_{i,j} = \sqrt{\sum_{p,q} \mathbf{G}_{p,q} \cdot v^2_{i+p,j+q}}$ is the norm of the $k \times k$ neighborhood, and $c \in \mathbb{R}$ is a tolerance parameter to avoid floating point precision problems. It has been empirically shown that such divisive normalization reduces statistical dependencies [28,25], which in turn accentuates differences between input features and accelerates gradient-based learning [29].

Improvement in both performance and training time when using such normalizations has been reported when the stochastic gradient descent algorithm is used to train deep networks [25]. This has been explained by the fact that, in the same way as whitening and other decorrelation methods, all variables end up with similar variances, making the model more likely to discover non-trivial relationships between spatially near inputs [30]. Also, it has been shown that similar strategies to locally normalize contrast in mammograms have enhanced performance of automatic analysis [31]. Fig. 3 shows an original image and its corresponding output after applying the preprocessing stage. Again, this preprocessing is performed in an image-wise fashion, thus it is not necessary to store parameters in the training procedure.

2.2.2. *Supervised feature learning*

A CNN is a neural network that shares connections between hidden units yielding low computational time and translational invariance properties. CNNs have been successfully applied in shape recognition problems [32] as well as medical diagnosis that involved texture as a discriminant feature [11]. Because mass characterization is highly correlated with shape and texture features [17,3], a CNN model becomes a suitable strategy for mass lesion classification. The main components of the CNN and the applied strategies to train it are detailed below.

Architecture: A CNN comprises 3 main components: a convolutional layer, an activation function and a pooling layer. To improve the capability of the model the three components are stacked iteratively so that the output of one component is the input for the next one, and the output of one set of components is the input for the next set, building a deep neural network with many layers. The convolutional layer is composed of several small matrices or "*kernels*" that are convolved

Please cite this article in press as: J. Arevalo, et al., Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. (2016), http://dx.doi.org/10.1016/j.cmpb.2015.12.014

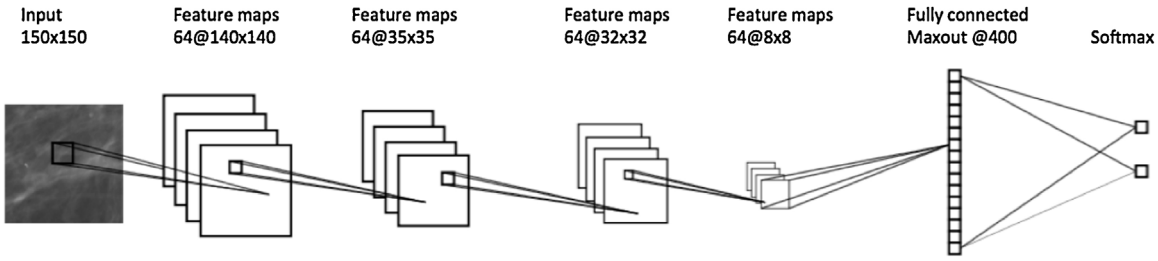


Fig. 4 – Best convolutional neural network evaluated on mass classification.

throughout the whole input image working as filters. The output of this convolution is called "*feature map*". These feature maps are the input for the activation function which applies a non-linear transformation in an element-wise fashion. Finally, the pooling layer aggregates contiguous values to one scalar with functions like mean or max.

The proposed architecture, depicted in Fig. 4, has 11×11 local kernels and the rectifier linear as activation function in the first convolutional layer followed by a 5×5 pooling layer with stride of 4×4 pixels. The second convolutional layer has 4×4 local kernels with the rectifier linear as activation function, with 4×4 pooling layer without overlapping. Then a fully connected layer with 400 units with maxout activation function is stacked to finally add a softmax classifier. In particular, the maxout activation function $h_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is a defined as:

$$h_i(\mathbf{s}) = \max_{j \in [1,k]} z_{i,j} \tag{2.4}$$

where $\mathbf{s} \in \mathbb{R}^n$ is the input vector, $z_{i,j} = \mathbf{s}^T \mathbf{W}_{\ldots ij} + \mathbf{b}_{ij}$ is the output of the j-th linear transformation of the i-th hidden unit, and $\mathbf{W} \in \mathbb{R}^{d \times m \times k}$ and $\mathbf{b} \in \mathbb{R}^{m \times k}$ are learned parameters. It has been shown that maxout models with just 2 hidden units behave as universal approximators, while are less prone to saturate units [33].

Since it is our intention to measure how the network's depth affects the performance of the model, we first evaluate the architecture with a single convolutional layer with a fully connected layer and called it CNN2 in the experiments. Consequently, the whole architecture, i.e. two convolutional layers plus a fully connected layer, is referenced as CNN3.

Regularization: The number of parameters in the model is directly related to capability to overfit the training data. Usually neural networks require different strategies to control this behavior. In this work dropout and max-norm regularization were used. Dropout randomly set to 0 the input of a unit, while max-norm regularization forces the norm of each vector of incoming weights in a unit to a maximum value. In [34] it was empirically shown that these two strategies help prevent co-adaptation between units, e.g., during error back propagation, a unit should not rely on other units to correct its mistakes since there is no certainty about their activations.

Optimization: The proposed architecture has approximately 4.6 million of parameters. Training large models has to scale in both, memory requirements and computational time. The strategy used in this work to train the CNN is stochastic gradient descent with momentum. An early stopping strategy monitoring the area under the ROC curve (AUC) on the validation set was chosen as stop criterion. The implementation of the whole framework was carried out with the Pylearn2 framework [35]. This library uses the Theano framework [36], which in turn takes advantage of GPU technology obtaining up to 140*times* speedup with respect to CPU implementations, making feasible the training of architectures with millions of parameters.

2.2.3. *Classification*

Following the previous work, a linear SVM was selected as classification strategy. Train and validation sets were used to fine-tune the *C* parameter. To evaluate the CNN as a representation strategy, images are propagated through the network, then the penultimate layer activations are extracted and used as representation. This process is done to reduce processing time because, in terms of computational cost, training a single SVM is cheaper than training the whole CNN network. This stage can be seen as a fine-tuning process of the last layer, where a smaller model is adjusted.

3. Experimental setup

The dataset was split in training (50%), validation (10%) and test (40%) sets following a stratified sampling per patient, that is, we make sure all computed instances of a particular patient belongs to only one of the three subsets. This setup warranties that the model is not tested using patients seen during the training stage.

In the preprocessing stage, the size of the cropped region was fixed to $r = 150$ according to the distribution of the lesion size and computational capability; and the filter size for LCN is $k = 11$ pixels. Following previous results [3], 5×5 and 3×3 blocks sizes for HOG and 4 and 8 regions for HGD were explored. Histograms for both 8 and 16 bins were evaluated. The best configuration in train-validation setup was used to report test results.

The CNN parameter exploration was performed by training 25 models with random hyperparameter initializations and the best was chosen according to validation performance. It has been reported that this strategy is preferable over grid search when training deep models [37]. Exploration was conducted using the CETA-CIEMAT[3] Research Center infrastructure. Bigger models that requires more intensive

[3] http://www.ceta-ciemat.es/, accessed on February 17, 2015

Please cite this article in press as: J. Arevalo, et al., Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. (2016), http://dx.doi.org/10.1016/j.cmpb.2015.12.014

computation were carried out using a NVidia Tesla K40 GPGPU card.

Before training the SVM model, a zero-mean unit-variance normalization process is carried out. Train and validation sets were used to fine-tune the C parameter for the SVM classifier. Final performance is reported in terms of AUC in the test set.

Comparison of the methods was based on the average AUC of 5 runs using different random seeds for dataset splitting for each run. Experiments were supported by the Wilcoxon signed rank test to determine whether differences have statistical evidence ($\rho > 0.1$).

4. Results

4.1. *Learned features*

Recall that the CNN weights in the first layer are equivalent to local kernels that work as filters over the image. Thus, visualizing them would allow to describe the patterns that the model is looking for. Fig. 5 shows the weights of the best learned model. This image exposes a set of edges in different orientations as well as some texture patterns. It seems the learned filters are affected by noise, probably because it is still few data for this kind of models. We experimentally found that normalization preprocessing was fundamental to obtain good-looking features and ultimately, good performances in the classification. Without normalization the models were not able to surpasses 0.7 of AUC.

4.2. *Classification results*

Fig. 6 shows ROC curves for all the evaluated representations for the best run. The HCfeats set, which uses segmentation information, performs slightly better than HOG-based descriptors. This confirms the importance of shape information for mass characterization. Interestingly, CNN models, which use only the raw pixels, outperform the state-of-the-art features [3]. The training of CNN3 model took 1.4 h on the Tesla K40 GPGPU card. It is also worthwhile noting that adding

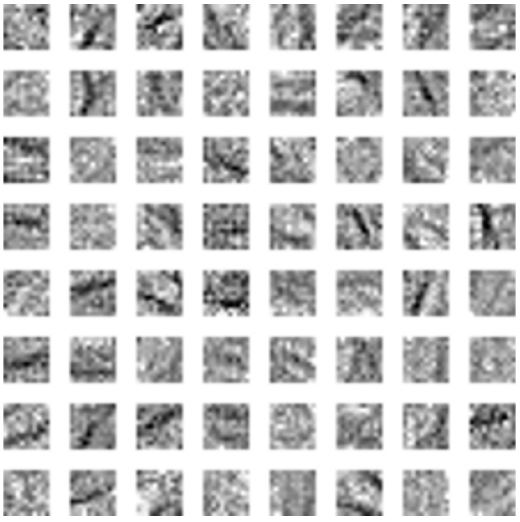


Fig. 5 – Filters learned in the first layer of the CNN model.

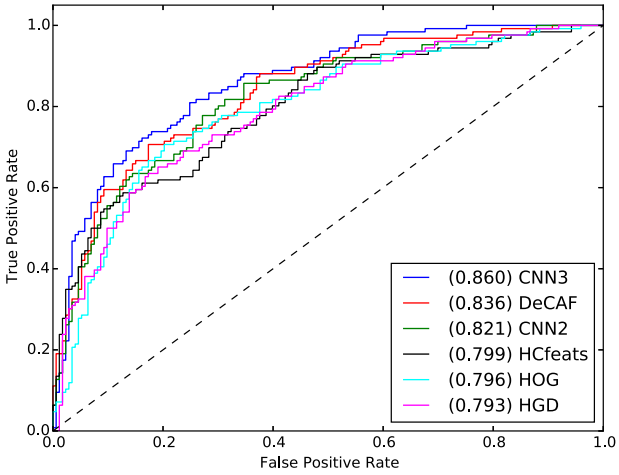


Fig. 6 – ROC curve for evaluated representations in test set for the best run.

a second hidden layer to the CNN model improves the representation capability producing better results. Such behavior is consistent with theoretical foundations to choose deep architectures over shallow ones [38].

For comparative purposes, we included the evaluation of DeCAF [39], a pre-trained model with the Imagenet dataset [40]. DeCAF is a model with greater complexity than all the other representations evaluated on this work. Thus, it is expected to perform better than using hand-crafted features. However, a smaller CNN model trained with the images of the domain performs the best. This behavior, similarly reported when CNNs are trained with small datasets [41], leads to the two main conclusions of this work: On one hand, CNN models outperform state-of-the-art representations for automatic lesion classification in mammography image analysis. On the other hand, such automatic mammography image analysis is a problem with its own particularities, and thus it is not enough to learn the representation using a large CNN model. The learning process should also be guided by a training set with a wide visual variability to show the model texture and shape features presented in mass lesions. Fig. 7 shows boxplots results in terms of AUC for each representation. According to the Wilcoxon test hypothesis, the CNN3 model performs best as compared to other evaluated representations ($\rho < 0.1$).

In order to combine the image-based features with additional information given in the segmentation, HCfeats, described in Section 2.1, were concatenated to each CNN representation and baseline descriptors (HOG, HGD and DeCAF). The resultant vector feature of each image has 417 elements, 400 from the last fully connected layer in the CNN plus 17 features from the HCfeats set. Table 2 shows a summary of these experiments. In general, this combination improves the results. It specially helps to augment the performance of the hand-crafted representations, while CNN models are not very affected. This suggests that CNN models are already capable of capturing shape information, which is consistent with the learned filters depicted in Fig. 5, and thus giving such information explicitly could be redundant. Again, this

Please cite this article in press as: J. Arevalo, et al., Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. (2016), http://dx.doi.org/10.1016/j.cmpb.2015.12.014

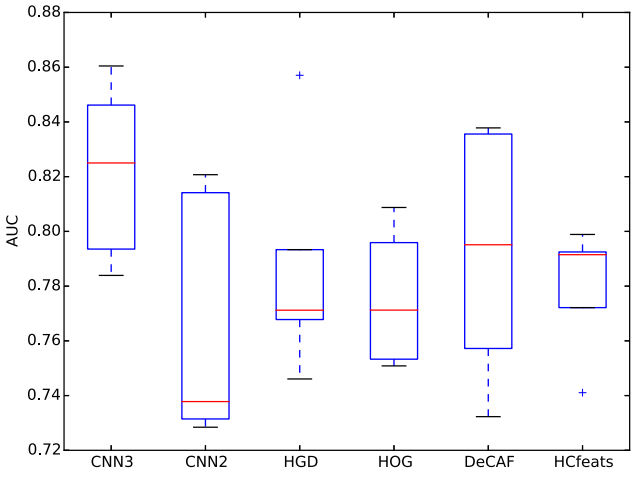


Fig. 7 – Boxplots of different runs for each representation method.

Table 2 – Summary of results in terms of AUC in the test set. Best results are shown in bold typeface and (*) signals scores with no evidence of differences from the highest ($\rho < 0.1$).

Representation	Standalone	Combined with HCfeats
CNN3	**0.82 ± 0.03**	0.82 ± 0.03 (*)
CNN2	0.76 ± 0.05	0.78 ± 0.04
HGD	0.78 ± 0.04	**0.83 ± 0.04**
HOG	0.77 ± 0.03	0.81 ± 0.03 (*)
DeCAF	0.79 ± 0.05	0.82 ± 0.03 (*)
HCfeats	0.77 ± 0.02	–

experimentation was supported by the Wilcoxon test, which showed no significant statistical evidence in the differences between representations combined with HCfeats. However, comparing standalone vs combined with HCfeats, all representations except CNN3, obtained evidence for a statistically significant improvement ($\rho < 0.05$).

An open question regarding these results is how this method would perform in high resolution images (12 or 14-bit images). Based on preliminary experimentation, we hypothesize that the model would obtain superior performance using higher resolution images, since the learning model will have more available information. However, we still do not have enough data to report statistically significant results. On the other hand, it is noteworthy that the neural network design would face new challenges such as higher dimensional input, fewer number of examples and different primitive patterns, among others. Thus, we believe new network architectures should be explored to address high resolution images.

5. Conclusions

This paper presented a framework to address classification of mass lesions in mammography film images. Instead of designing particular descriptors to explain the content of mammography images, the proposed approach learns them directly from data in a supervised way. CNNs were used as the representation learning strategy. The proposed neural network architecture takes the raw pixels of the image as input, to learn in a hierarchical way a set of nonlinear transformations to represent the visual content of an image. The model is composed of a set of local filters with a rectified linear unit activation function, maxpooling layers, a fully-connected layer with maxout activation function and a softmax layer. Our approach outperformed the state-of-the-art image features, HOG and HGD descriptors [3], increasing the performance from 0.787 to 0.822 in terms of AUC. Interestingly, this model also outperforms a set of hand-crafted features that take advantage of additional information from segmentation by the radiologist. Finally, the combination of both representations, learned and hand-crafted, resulted in the best descriptor for mass lesion classification, obtaining 0.826 in the AUC score.

Our future work includes larger architectures as well as the inclusion of other image modalities to enhance the representation. It also would be worth to evaluate the proposed strategy on BCDR-DM images since this suppose a new challenge due to the high resolution images.

Acknowledgements

This work was mainly supported by the Cloud Thinking project (CENTRO-07-ST24-FEDER-002031), co-funded by QREN, "Mais Centro" program. Also, this work was partially funded by projects "Multimodal Image Retrieval to Support Medical Case-Based Scientific Literature Search", ID R1212LAC006 by Microsoft Research LACCIR, "Diseño e implementación de un sistema de cómputo sobre recursos heterogéneos para la identificación de estructuras atmosféricas en predicción climatológica" number 1225-569-34920 through Colciencias contract number 0213-2013, "Programa nacional de proyectos para el fortalecimiento de la investigación, la creación y la innovación en posgrados de la Universidad Nacional de Colombia 2013–2015" with proposal number 18722 and the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the Government of Spain. Arevalo also thanks Colciencias for its support through a doctoral grant in call 617/2013. The authors also thank for K40 Tesla GPU donated by NVIDIA and which was used for some feature learning experiments.

REFERENCES

[1] L. Tabár, B. Vitak, T.H.-H. Chen, A.M.-F. Yen, A. Cohen, T. Tot, S.Y.-H. Chiu, S.L.-S. Chen, J.C.-Y. Fann, J. Rosell, H. Fohlin, R.A. Smith, S.W. Duffy, Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades, Radiology 260 (3) (2011) 658–663, http://dx.doi.org/10.1148/radiol.11110469.

[2] T. Ayer, M.U. Ayvaci, Z.X. Liu, O. Alagoz, E.S. Burnside, Computer-aided diagnostic models in breast cancer screening, Imaging Med. 2 (3) (2010) 313–323.

[3] D.C. Moura, M.A. Guevara López, An evaluation of image descriptors combined with clinical data for breast cancer diagnosis, Int. J. Comp. Assist. Radiol. Surg. 8 (4) (2013) 561–574, http://dx.doi.org/10.1007/s11548-013-0838-2.

[4] R. Ramos-Pollán, M.A. Guevara-López, C. Suárez-Ortega, G. Díaz-Herrero, J.M. Franco-Valiente, M. Rubio-del Solar, N. González-de Posada, M.A.P. Vaz, J. Loureiro, I. Ramos, Discovering mammography-based machine learning classifiers for breast cancer diagnosis, J. Med. Syst. 36 (4) (2012) 2259–2269, http://dx.doi.org/10.1007/s10916-011-9693-2.

[5] R. Ramos-Pollán, M.A. Guevara-López, E. Oliveira, A software framework for building biomedical machine learning classifiers through grid computing resources, J. Med. Syst. 36 (4) (2012) 2245–2257, http://dx.doi.org/10.1007/s10916-011-9692-3.

[6] X. Liu, J. Tang, Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method, Syst. J. IEEE 8 (3) (2014) 910–920, http://dx.doi.org/10.1109/JSYST.2013.2286539.

[7] M. Dong, X. Lu, Y. Ma, Y. Guo, Y. Ma, K. Wang, An efficient approach for automated mass segmentation and classification in mammograms, J. Digit. Imaging 28 (5) (2015) 613–625, http://dx.doi.org/10.1007/s10278-015-9778-4.

[8] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828, doi:10.1109/TPAMI.2013.50.

[9] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117, http://dx.doi.org/10.1016/j.neunet.2014.09.003.

[10] J. Arevalo, A. Cruz-Roa, F.A. González, Hybrid image representation learning model with invariant features for basal cell carcinoma detection, Proc. SPIE 8922 (2013), http://dx.doi.org/10.1117/12.2035530, pp. 89220M-89220M-6.

[11] A.A. Cruz-Roa, J.E.A. Ovalle, A. Madabhushi, F.A.G. Osorio, A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection, in: in: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2013, Springer, 2013, pp. 403–410, http://dx.doi.org/10.1007/978-3-642-40763-5_50.

[12] H.I. Suk, D. Shen, Deep learning-based feature representation for AD/MCI classification Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 8150 LNCS, 2013, pp. 583–590, http://dx.doi.org/10.1007/978-3-642-40763-5_72.

[13] H.-I. Suk, S.-W. Lee, D. Shen, Latent feature representation with stacked auto-encoder for AD/MCI diagnosis, Brain Struct. Funct. (2013) 1–19, http://dx.doi.org/10.1007/s00429-013-0687-3.

[14] H.-I. Suk, S.-W. Lee, D. Shen, Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, Neuroimage 101 (0) (2014) 569–582, http://dx.doi.org/10.1016/j.neuroimage.2014.06.077.

[15] F. Li, L. Tran, K.-H. Thung, S. Ji, D. Shen, J. Li, Robust deep learning for improved classification of AD/MCI patients, in: G. Wu, D. Zhang, L. Zhou (Eds.), Machine Learning in Medical Imaging, Vol. 8679 of Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 240–247, http://dx.doi.org/10.1007/978-3-319-10581-9_30.

[16] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, M. Nielsen, Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network, in: K. Mori, I. Sakuma, Y. Sato, C. Barillot, N. Navab (Eds.), in: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013, Vol. 8150 of Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, 2013, pp. 246–253, http://dx.doi.org/10.1007/978-3-642-40763-5_31.

[17] A. Jalalian, S.B. Mashohor, H.R. Mahmud, M.I.B. Saripan, A.R.B. Ramli, B. Karasfi, Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review, Clin. Imaging 37 (3) (2013) 420–426, http://dx.doi.org/10.1016/j.clinimag.2012.09.024.

[18] K. Petersen, K. Chernoff, M. Nielsen, A.Y. Ng, Breast density scoring with multiscale denoising autoencoders, in: in: STMI workshop at MICCAI 2012 (15th International Conference on Medical Image Computing and Computer Assisted Intervention), 2012.

[19] K. Petersen, M. Nielsen, P. Diao, N. Karssemeijer, M. Lillholm, Breast tissue segmentation and mammographic risk scoring using deep learning, in: H. Fujita, T. Hara, C. Muramatsu (Eds.), Breast Imaging, Vol. 8539 of Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 88–94, http://dx.doi.org/10.1007/978-3-319-07887-8_13.

[20] X.-S. Zhang, A new approach for clustered MCs classification with sparse features learning and TWSVM, Sci. World J. (2014) 970287, http://dx.doi.org/10.1155/2014/970287.

[21] J. Ge, B. Sahiner, L.M. Hadjiiski, H.-P. Chan, J. Wei, M.A. Helvie, C. Zhou, Computer aided detection of clusters of microcalcifications on full field digital mammograms, Med. Phys. 33 (8) (2006) 2975–2988.

[22] A.R. Jamieson, K. Drukker, M.L. Giger, Breast image feature learning with adaptive deconvolutional networks, 2012, http://dx.doi.org/10.1117/12.910710.

[23] G.I. Andreea, R. Pegza, L. Lascu, S. Bondari, Z. Zoia Stoica, A. Bondari, The role of imaging techniques in diagnosis of breast cancer, J. Curr. Health Sci. 37 (2) (2011) 241–248.

[24] N.P. Pérez, M.A.G. López, A. Silva, I. Ramos, Improving the Mann–Whitney statistical test for feature selection: an approach in breast cancer diagnosis on mammography, Artif. Intell. Med. (2014), http://dx.doi.org/10.1016/j.artmed.2014.12.004.

[25] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition? in: in: Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 2146–2153, http://dx.doi.org/10.1109/ICCV.2009.5459469.

[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, Curran Associates Inc., 2012, pp. 1097–1105.

[27] N. Pinto, D.D. Cox, J.J. DiCarlo, Why is real-world visual object recognition hard? PLoS Computat. Biol. 4 (1) (2008) e27, http://dx.doi.org/10.1371/journal.pcbi.0040027.

[28] S. Lyu, E. Simoncelli, Nonlinear image representation using divisive normalization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, 2008, pp. 1–8, http://dx.doi.org/10.1109/CVPR.2008.4587821.

[29] Y. LeCun, Learning invariant feature hierarchies, in: A. Fusiello, V. Murino, R. Cucchiara (Eds.), Computer Vision – ECCV. Workshops and Demonstrations, Vol. 7583 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, Florence, Italy, 2012, pp. 496–505, http://dx.doi.org/10.1007/978-3-642-33863-2_51.

[30] A. Krizhevsky, Learning multiple layers of features from tiny images, Tech. rep., University of Toronto, Toronto, 2009.

[31] W.J. Veldkamp, N. Karssemeijer, Normalization of local contrast in mammograms, IEEE Trans. Med. Imaging 19 (7) (2000) 731–738, http://dx.doi.org/10.1109/42.875197.

[32] Q. Ke, Y. Li, Is rotation a nuisance in shape recognition? in: in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 2014, pp. 4146–4153, http://dx.doi.org/10.1109/CVPR.2014.528.

[33] I. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, in: S. Dasgupta, D. Mcallester (Eds.), Proceedings of the 30th International Conference on Machine Learning (ICML-13), Vol. 28, JMLR Workshop and Conference Proceedings, 2013, pp. 1319–1327.

Please cite this article in press as: J. Arevalo, et al., Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. (2016), http://dx.doi.org/10.1016/j.cmpb.2015.12.014

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting J. Mach. Learn. Res. 15 (2014) 1929–1958.
[35] I.J. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien, Y. Bengio, Pylearn2: a machine learning research library, arXiv preprint arXiv:1308.4214.
[36] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I.J. Goodfellow, A. Bergeron, N. Bouchard, Y. Bengio, Theano: new features and speed improvements, in: Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.
[37] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.
[38] L. Deng, D. Yu, Deep learning: methods and applications, Found. Trends Signal Process. 7 (3–4) (2014) 197–387.
[39] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, arXiv preprint arXiv: 1310.1531.
[40] O. Russakovsky, J. Deng, Z. Huang, A.C. Berg, L. Fei-Fei, Detecting avocados to zucchinis: what have we done, and where are we going? in: in: International Conference on Computer Vision (ICCV), 2013.
[41] D. Rueda-Plata, R. Ramos-Pollán, F.A. González, Supervised greedy layer-wise training for deep convolutional networks with small datasets, in: M. Nú nez, N. Nguyen, D. Camacho, B. Trawiński (Eds.), Computational Collective Intelligence, Vol. 9329 of Lecture Notes in Computer Science, Springer International Publishing, 2015, pp. 275–284, http://dx.doi.org/10.1007/978-3-319-24069-5_26.

Please cite this article in press as: J. Arevalo, et al., Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. (2016), http://dx.doi.org/10.1016/j.cmpb.2015.12.014