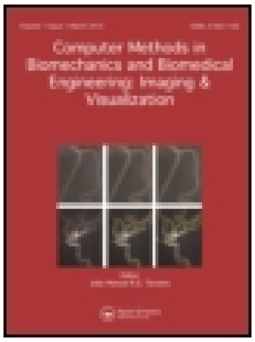

Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization

ISSN: 2168-1163 (Print) 2168-1171 (Online) Journal homepage: http://www.tandfonline.com/loi/tciv20

Lesion classification in mammograms using convolutional neural networks and transfer learning

Ana C. Perre, Luís A. Alexandre & Luís C. Freire

To cite this article: Ana C. Perre, Luís A. Alexandre & Luís C. Freire (2018): Lesion classification in mammograms using convolutional neural networks and transfer learning, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, DOI: 10.1080/21681163.2018.1498392

To link to this article: https://doi.org/10.1080/21681163.2018.1498392

Published online: 26 Jul 2018.

Lesion classification in mammograms using convolutional neural networks and transfer learning

Ana C. Perre[a], Luís A. Alexandre[b] and Luís C. Freire[c]

[a]Instituto de Telecomunicações and Faculdade Ciências da Saúde, Universidade da Beira Interior, Covilhã, Portugal; [b]Instituto de Telecomunicações and Departamento de Informática, Universidade da Beira Interior, Covilhã, Portugal; [c]Escola Superior de Tecnologia da Saúde de Lisboa, Instituto Politécnico de Lisboa, Lisboa, Portugal

ABSTRACT
Convolutional neural networks (CNNs) have recently been successfully used in the medical field to detect and classify pathologies in different imaging modalities, including in mammography. One disadvantage of CNNs is the need for large training datasets, which are particularly difficult to obtain in the medical domain. One way to solve this problem is using a transfer learning approach, in which a CNN, previously pre-trained with a large amount of labelled non-medical data, is subsequently fine-tuned using a smaller dataset of medical data. In this paper, we use such a transfer learning approach, which is applied to three different networks that were pre-trained using the Imagenet dataset. We investigate how the performance of these pre-trained CNNs to classify lesions in mammograms is affected by the use, or not, of normalised images during the fine-tuning stage. We also assess the performance of a support vector machine fed with features extracted from the CNN and the combined use of handcrafted features to complement the CNN-extracted features. The obtained results are encouraging.

ARTICLE HISTORY
Received 15 December 2017
Accepted 2 July 2018

KEYWORDS
Mammographic image; convolutional neural network; transfer learning; support vector machine; breast cancer; lesion classification

1. Introduction

The interpretation of mammographic images can be very difficult to radiologists and, according to Jalalian et al. (2013), they fail to detect 10–30% of breast cancers, mainly because screening is a repetitive and fatiguing task (Sampat et al. 2005).

Therefore, Computer-Aided Detection/Diagnosis tools were created to assist the detection and diagnosis of early stage cancers, decreasing false negative rate and improving radiologists' efficiency (Jalalian et al. 2013; Arevalo et al. 2016; Tang ASTI 2014; Ganesan et al. 2013).

Since 2006, deep learning algorithms have become an important tool in the field of big data and artificial intelligence (Jiao et al. 2016). These algorithms simulate the human visual system and are able to learn complex relationships between labelled data samples; their fields of application include, but are not limited to, image understanding, speech recognition and natural language processing (Arevalo et al. 2016; Jiao et al. 2016).

Convolutional neural networks (CNNs) are one example of deep learning algorithms that proved to be successful (Jiao et al. 2016). They were introduced by Fukushima and later improved by LeCun et al. and are considered the most successful type of deep learning algorithms in image understanding (Arevalo et al. 2016). CNNs have been used in complex tasks such as visual object recognition and image classification (Jiao et al. 2016). In the biomedical image processing field, CNNs are applied in several areas such as electron microscopy images, breast histology images, mammography images and magnetic resonance images of the brain (Jiao et al. 2016; Arevalo et al. 2016).

In medical image classification, CNNs could be trained from scratch. However, that would require large amounts of data and extensive computational/memory resources and could, eventually, lead to overfitting and convergence problems (Tajbakhsh et al. 2016). To prevent these issues, it is possible to fine-tune a pre-trained CNN model that has been trained using a large amount of non-medical labelled data such as the one that can be found in the ImageNet database, which offers more than 1.2 million categorised natural images (Shin et al. 2016).

In this paper, we have applied CNNs to solve the problem of mammographic lesion classification into benign or malign classes.

Figure 1 illustrates the differences between both types of lesions mentioned before. Note that regular contours are compatible with benign lesions, while an irregular shape is often associated with malignancy (Pisco 2003). Therefore, we have studied the performance of three different types of CNN implementations when fine-tuned using images that were, or were not, normalised, allowing us to understand the impact of normalisation on the lesion classification results. We have also analysed the performance of a support vector machine (SVM) fed with features extracted from the CNN. Finally, we have evaluated the use of handcrafted features to complement the CNN-extracted features.

CONTACT Ana C. Perre ✉ ana.perre@ipcb.pt Faculdade Ciências da Saúde and Instituto de Telecomunicações, Universidade da Beira Interior, Covilhã 6201-001, Portugal

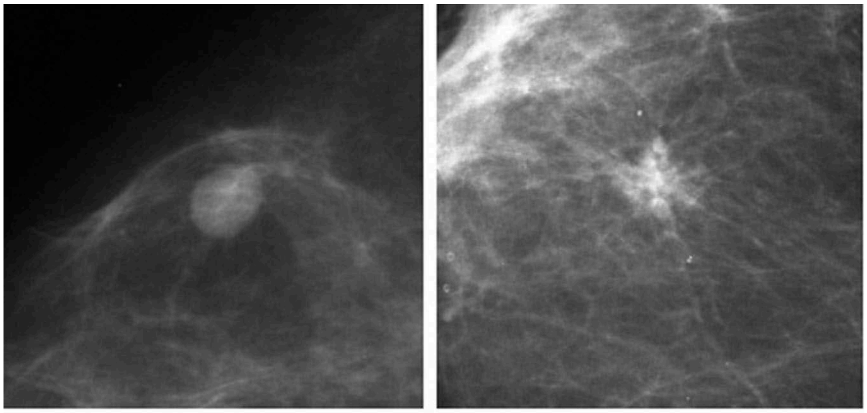

Figure 1. Example of benign lesion on the left and malign lesion on the right (image from the BCDR-FM dataset).

2. State-of-art

Deep learning-based algorithms have recently shown potential for applications in digital pathology. Since 2012, these algorithms are being used in major computer vision competitions, such as the ImageNet Large Scale Visual Recognition Competition, showing the best performance in their class (Wang et al. 2016).

CNNs have already been used by other researchers in the medical image field and, specifically, in the mammographic image field. Table 1 summarises the works of some of them, which are described in detail below.

Our study has initially been guided by the work of Arevalo et al. (2016), who proposed a new method that was applied to the BCDR-F03 (Film Mammography Dataset Number 3) dataset from the Breast Cancer Digital Repository. The method includes baseline descriptors, such as Handcraft features (HCfeats), Histogram of oriented gradients and Histogram of gradient divergence, in a supervised feature learning approach that incorporated a CNN. For image classification, the activations from the penultimate layer were extracted and used as input of an SVM. The authors also used different CNN models (CNN2, which consisted in a single connected layer combined with a fully connected layer, CNN3, which consisted in two convolutional layers and a fully connected layer and DeCAF – a pre-trained model with ImageNet) obtaining AUC mean values of 0.780, 0.820 and 0.820, respectively, when combined with HCfeats, and nearly 0.76, 0.82 and 0.79, respectively, when used standalone.

Table 1. State-of-art summary.

Author	Dataset	Preprocess	Classifier	Classes	Performance (mean)
Arevalo et al. (2016)	BCDR-F03	GCN + LCN	CNN + SVM	2 (B/M)	76.0–82.0% (AUC)
Wichakam and Vateekul (2016)	INBreast	–	CNN + SVM	2 (B/M)	90.63–98.44% (Acc)
Jiao et al. (2016)	DDSM	μ & std	CNN + SVM	2 (B/M)	92.0–97.0% (Acc)
Yi et al. (2017)	DDSM	μ	CNN (Pre-trained)	2 (B/M)	58.0–85.0% (Acc)
Sun et al. (2016)	Own	–	CNN(SSL) &SVM	2 (B/M)	82.36–88.18% (AUC)

Wichakam and Vateekul (2016) published another study that combined deep convolutional networks, used as an automatic feature extraction tool, and an SVM, used as a classifier, for mass detection on digital mammograms and applied them to the INBreast dataset. Different approaches in the deep convolutional networks allowed reaching the best accuracy performance of 98.44% (with the SVM-FC1 of the A3 architecture).

Jiao et al. (2016) used images from the Digital Database for Screening (DDSM) dataset and applied a fine-tuning operation on the trained deep CNN model in LSVRC (dataset containing more than one million labelled natural images), in order to extract middle-level and high-level features from different hierarchical levels and used them to train two classifiers for the decision procedure. In the testing process, if the results of linear SVM applied on high- and mid-level features were consistent, they would add them to the subset – result 1. However, if the outcomes were inconsistent, they would use the original grey information of the training set to calculate the lesion closeness to the benign or malignant class, creating sub-classes that were used to obtain the similarity measure, which helped them to achieve the result 2. The final outcome contained both results 1 and 2. At the end, and with deep features of different layers, they obtained a classification accuracy of 96.7%, surpassing the results from other researchers indicated by them and having better results than if they had used middle and high features separately. Finally, they compared their network with the most used variations of Alex net. So, with the Caffe ref and VGG models, they obtained accuracy values of 92% and 97%, respectively (Jiao et al. 2016).

In their work, Yi et al. (2017) used the DDSM database and the best result was obtained with an ensembled GoogLeNet-based architecture (in parallel training), which achieved an accuracy of 0.85 and an AUC value of 0.91. They concluded that instead of using a network that has an architecture built to process craniocaudal (CC) and medio-lateral oblique (MLO) views independently, better results could be obtained with a single network taking into consideration both views. They created a visualisation method based on what they called *Directed Dream Images*, which would enhance and exaggerate some aspects of the image, creating patterns that corresponded to a high malignant/benign score based on images seen in training data, which were learned by the CNN as clinically relevant features.

Sun et al. (2016), developed a new semi-supervised learning (SSL) algorithm that allows the use of a small amount of labelled data and a relatively small amount of unlabelled data to train a CNN. They concluded that unlabelled data may only be used as supplementary, because it could not replace labelled data, which definitely improves the overall accuracy because it contains more accurate information from the radiologist. However, they presented some results with unlabelled data to train the CNN, getting an AUC value of 0.8818 with the SSL method. Using an SVM with labelled data only or mixed data, the AUC values were 0.8236 and 0.8535, respectively.

3. Lesion classification using CNNs with transfer learning

3.1 *Transfer learning for lesion classification*

In this paper, we propose to study the application of several classification models and preprocessing strategies for mass detection in digitised mammograms.

It is well known that CNNs require large amounts of data to be properly trained. However, in the medical field, it is usually difficult to obtain such large datasets, which is due not only to the limited number of exams produced in a single facility, but also to the amount of work that is needed for hand labelling of the samples. So, our work will be based on a transfer learning approach; we will reuse CNNs that were previously trained for a different task and fine-tune them to our current problem.

The three different pre-trained models studied in this paper were previously used to perform classification in the ImageNet ILSVRC challenge data: CNN-F (Fast, imagenet-vgg-f) model, CNN-M (Medium, imagenet-vgg-m) models (Chatfield et al. 2014), and Caffe reference model (Jia et al. 2014). We fine-tuned these networks in order to achieve the classification of *benign* or *malign* lesions from the mammographic images.

In order to apply the pre-trained models to our problem, we have adapted the software MatConvNet (Vedaldi and Lenc 2015) available for Matlab (System specifications: Matlab R2015a and Intel i7-3820 CPU @ 3.60GHz with 32GB RAM).

3.2 *The networks*

As mentioned above, three different pre-trained models were used in this work: CNN-F, CNN-M and Caffe. The first two were chosen because they are often referenced in the literature; besides, Caffe is a new version of the DeCAF model, which was used by Arevalo et al. (2016). Therefore, using these three networks, inter-comparison of results can be achieved.

Table 2 presents the differences between the three pre-trained networks. In the convolutional layers' columns, indicated as '*Conv#*', the '*num* × *size* × *size*' set indicates the number of convolution filters and their receptive field size. The indications '*st.*' and '*pad.*' represent the convolution stride and the spatial padding, whereas the *LRN* is the Local Response Normalisation with or without the max-pooling down-sampling factor. In the fully connected layers' columns, indicated as ('*Full#*'), the number indicates their dimensionality; besides, '*Full6*' and '*Full7*' are regularised using dropout and the last layer corresponds to the softmax classifier. Except for the last layer, the Rectification Linear Unit is the activation function for all weight layers (Chatfield et al. 2014).

The architecture of the CNN-F model consists 8 learnable layers (5 convolutional layers and 3 fully connected layers), and the fast processing is guaranteed by the 4-pixel stride in the first convolutional layer. On the other hand, the CNN-M architecture, in the first convolutional layer, has a decreased stride and smaller receptive field and, in the second convolutional layer, has a larger stride keeping the computation time reasonable (Chatfield et al. 2014).

The Caffe reference model, like the others mentioned before, has a complete set of layers, which are used for visual tasks such as classification, and trains the model using a standard stochastic gradient descent algorithm (Jia et al. 2014).

3.3 *The dataset*

We have used the BCDR-FM dataset (Film Mammography Dataset) from the Breast Cancer Digital Repository (http://bcdr.inegi.up.pt), which includes 1125 studies with 3703 MLO and CC images of 1010 patient cases, mostly female gender (998), from 20 to 90 years old. The dataset also contains 1044 identified – and clinically described – lesions, 1517 manually made segmentations and BI-RADS classifications carried out by specialised radiologists (Arevalo et al. 2016).

Table 2. CNN pre-trained models used in this work (adapted from Chatfield et al. 2014).

Archit.	Conv1	Conv2	Conv3	Conv4	Conv5	Full6	Full7	Full8
CNN-F	6411 × 11 st.4, pad.0 LRN, × 2pool	256 × 5 × 5 st.1, pad.2 LRN, × 2pool	256 × 3 × 3 st.1, pad.1	256 × 3 × 3 st.1, pad.1	256 × 3 × 3 st.1, pad.1 × 2pool	4096 dropout	4096 dropout	1000 soft-max
CNN-M	96 × 7 × 7 st.2, pad.0 LRN, × 2pool	256 × 5 × 5 st.2, pad.1 LRN, × 2pool	512 × 3 × 3 st.1, pad.1	512 × 3 × 3 st.1, pad.1	512 × 3 × 3 st.1,pad.1 × 2pool	4096 dropout	4096 dropout	1000 softmax
Caffe	96 × 11 × 11 st.4,pad.0 LRN, × 2pool	256 × 5 × 5 st.1, pad.2 LRN, × 2pool	384 × 3 × 3 st.1, pad.1	384 × 3 × 3 st.1, pad.1	256 × 3 × 3 st.1,pad.1 × 2pool	4096 dropout	4096 dropout	1000 softmax

The downloaded dataset, named BCDR-F03 – 'Film Mammography Dataset Number 3', which is a subset of the BCDR-FM, comprises 736 grey-level digitised mammograms (426 benign and 310 malign mass lesions) from 344 patients. These are distributed into MLO and CC views with image size of 720 × 1168 (width × height) pixels and a bit depth of 8 bits per pixel in TIFF format; included are also clinical data and image-based descriptors. Although a digital dataset is available, we have used the digitised dataset to enable the comparison with the work of Arevalo et al. (2016); besides, digital images have a bigger bit depth of 14 bits per pixel.

The preprocessing stage of our work is similar to the one used in Arevalo et al. (2016), namely: (i) *cropping* a ROI of 150 × 150 pixels using the information of the bounding box of the segmented region, being the aspect ratio always preserved even when the lesion's dimensions are bigger than the ROI. However, when the lesion is next to the border of the image, we translate the square crop, thus changing image coordinates and including the surrounding breast pattern, instead of zero-padding the outer portion of the crop; (ii) *data augmentation* using a combination of flipping and 90, 180 and 270 degrees rotation transformations.

3.4 *Image normalisation*

The data normalisation procedure used in this work is similar to the one proposed by Arevalo et al. (2016); it consists in a *Global Contrast Normalisation (GCN)*, obtained by subtracting the mean of the intensities in the image (calculated per image and not per pixel) to each pixel, and a *Local Contrast Normalisation (LCN)* (Arevalo et al. 2016).

We have then divided images into three groups: 50% for training, 10% for validation and 40% for testing. The images' input size for the different models was 224 × 224 pixels; the parameters' exploration space comprised three fully connected layers, 50 epochs, an fc8 initially randomised layer, five learning rate values (1e–2, 1e–3, 1e–4, 5e–2, 5e–3 and 5e–4), the three pre-trained models (vgg-f, vgg-m and caffe) and the use, or not, of normalised images – see Figure 2.

4. Experiments

After the fine-tuning of the three networks using the train and the validation sets (which comprised 2800 and 560 images, respectively) with and without normalisation, we have chosen the best parameters to apply to the test set (comprised of 2240 images); in the subsequent experiments, the training set comprised 3360 images due to the merge of the initial training and validation sets.

Afterwards, we have chosen the network with the best performance and extracted the activations from one or several of the last layers, from the sixteenth to the nineteenth layer. Used separately or combined two-by-two, three-by-three, or all together, the extracted activations were then used to train an SVM in order to assess if the classification performance improved.

After that, we have used handcrafted features given by the dataset authors, once more to assess if there was an improvement in the classification results. These handcrafted features include *Intensity features* (mean, median, maximum, minimum, standard deviation, skewness and kurtosis), *Shape features* (area, perimeter, circularity, elongation, *y*_centre_mass, *x*_centre_mass and form) and *Texture features* (contrast, correlation and entropy). Beyond these, we have also included age and density information, and used all of them together or separately to see which ones had larger influence in the classification performance.

5. Results and discussion

The results of the parameters' exploration are shown in Tables 3 and 4. With normalised training and validation sets, the best AUC mean value was achieved using the Caffe reference model (AUC mean = 0.775, std = 0.014), followed by the CNN-F model (AUC mean = 0.752, std < 0.001) and the CNN-M (AUC mean = 0.743, std = 0.005). This is somewhat surprising given that the CNN-M is a more powerful model (the filters are larger) than the CNN-F.

Relatively to the training and validation sets created without normalisation, the best AUC mean value was achieved by the CNN-M model (AUC mean = 0.785, std = 0.003), followed by the Caffe reference model (AUC mean = 0.769, std = 0.002) and the CNN-F model (AUC mean = 0.763, std = 0.004). The CNN-M did improve by a significant amount with and without normalisation (from an AUC mean value of 0.743–0.785, respectively).

Once the best combination of parameters to each model was determined, new results were obtained using the test set (and the new merged training set), which are presented in Table 5. Figure 3 also shows the graphic for the run that yielded the best AUC value. It is possible to see that we have achieved the best AUC mean value of 0.813 (std = 0.001) with the Caffe reference model and no normalisation, surpassing the result of 0.79 in Arevalo et al. (2016), which was obtained with the combined use of the DeCAF model, normalised images and an SVM instead of a softmax layer (since they considered that the former had better performance as classifier than the latter). Relatively to the computational time, the Caffe model was the fastest, with 22.66 min for training and 2.58 min for testing, totalling 25.24 min; in general, one observed that time increased when normalised image crops were used.

As it happened with the validation set, the best AUC mean values were achieved using images without normalisation, namely 0.776 with CNN-M and 0.767 with CNN-F, which are similar to the ones obtained during the fine-tuning stage. The AUC mean values with normalised images are lower than those obtained with the validation set, especially the one

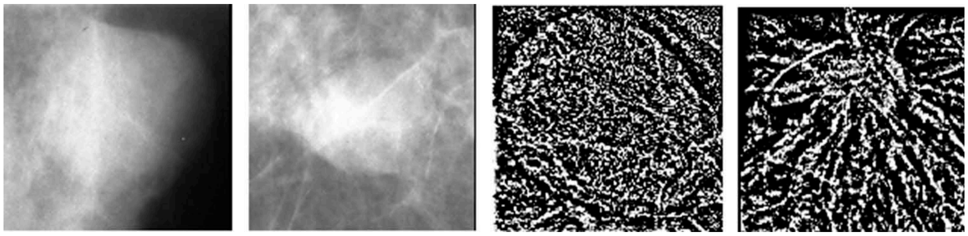

Figure 2. Examples of 150 × 150 crop images; the first two were obtained without normalisation, whereas the last two were obtained with normalisation.

Table 3. Results in bold correspond to the best mean AUC values achieved for each network. CNN parameter exploration, with five repetitions, using normalised images (only the train and validation sets were used).

Network	Batch size	Learning rate	Top1err train mean	Top1err validation mean	AUC mean	AUC std
CNN-F	256	1e−2	0.208	0.267	0.752	0.002
CNN-F	256	1e−3	0.302	0.360	0.720	0.006
CNN-F	256	1e−4	0.371	0.500	0.638	0.007
CNN-F	256	5e−2	0.371	0.500	0.646	0.037
CNN-F	256	5e−3	0.231	0.252	**0.752**	0.000
CNN-F	256	5e−4	0.356	0.479	0.678	0.000
CNN-M	64	1e−2	0.112	0.304	0.734	0.008
CNN-M	64	1e−3	0.186	0.286	0.732	0.001
CNN-M	64	1e−4	0.321	0.381	0.738	0.008
CNN-M	64	5e−2	0.332	0.359	0.739	0.040
CNN-M	64	5e−3	0.133	0.290	0.739	0.003
CNN-M	64	5e−4	0.212	0.263	**0.743**	0.005
Caffe	256	1e−2	0.205	0.251	0.758	0.001
Caffe	256	1e−3	0.298	0.342	0.716	0.004
Caffe	256	1e−4	0.371	0.500	0.687	0.014
Caffe	256	5e−2	0.329	0.334	**0.775**	0.014
Caffe	256	5e−3	0.224	0.251	0.753	0.000
Caffe	256	5e−4	0.337	0.458	0.696	0.036

Table 4. Results in bold correspond to the best mean AUC values achieved for each network. CNN parameter exploration, with five repetitions, using images with no normalisation (only the train and validation sets were used).

Network	Batch size	Learning rate	Top1err train mean	Top1err validation mean	AUC mean	AUC std
CNN-F	256	1e−2	0.203	0.326	**0.763**	0.004
CNN-F	256	1e−3	0.254	0.361	0.748	0.018
CNN-F	256	1e−4	0.371	0.500	0.686	0.013
CNN-F	256	5e−2	0.370	0.483	0.745	0.005
CNN-F	256	5e−3	0.206	0.310	0.762	0.006
CNN-F	256	5e−4	0.336	0.420	0.696	0.013
CNN-M	64	1e−2	0.095	0.265	0.757	0.019
CNN-M	64	1e−3	0.150	0.265	0.781	0.005
CNN-M	64	1e−4	0.232	0.353	0.765	0.003
CNN-M	64	5e−2	0.203	0.334	0.742	0.040
CNN-M	64	5e−3	0.107	0.270	0.760	0.009
CNN-M	64	5e−4	0.170	0.270	**0.785**	0.003
Caffe	256	1e−2	0.177	0.323	0.767	0.003
Caffe	256	1e−3	0.230	0.376	0.765	0.004
Caffe	256	1e−4	0.371	0.500	0.680	0.021
Caffe	256	5e−2	0.345	0.425	0.740	0.013
Caffe	256	5e−3	0.190	0.325	**0.769**	0.002
Caffe	256	5e−4	0.281	0.381	0.756	0.003

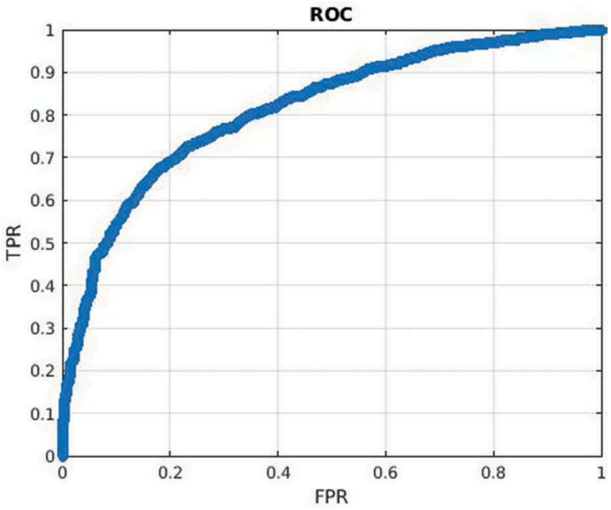

Figure 3. Example AUC for the best run achieved with Caffe using images without normalisation process (AUC = 0.814).

yielded by the Caffe model, which was substantially lower (AUC mean = 0.584; previous one was 0.775).

After obtaining the AUC mean values only with a CNN, we have chosen the network with the best performance, which was the Caffe reference model, and extracted the activations from the last layers, namely from the sixteenth to the nineteenth layers. Furthermore, we have tested the combination of the different extracted features (two-by-two, three-by-three, or all of them). The results are presented in Table 6. In general, the AUC mean values are lower than the ones obtained with the standalone CNN and the best performance was achieved with activations from the sixteenth layer (AUC mean = 0.773; best run, AUC = 0.784), and the combination of the sixteenth and eighteenth layers (AUC mean = 0.772; best run, AUC = 0.786).

Regarding the results presented in Table 6, the best accuracy is obtained using the sixteenth layer or this one combined with the seventeenth layer; the sensibility is rather low for all the tested combinations, with the best result obtained using the features from the seventeenth layer (0.557); regarding specificity, the best result of 0.846 is obtained using the combined features extracted from the sixteenth and eighteenth layers, which also yielded the best result in terms of precision (0.743); the best results in terms of *f*-measure (0.628) is obtained using the features from the sixteenth and seventeenth layers together.

As the best AUC results were obtained using the activations from the sixteenth layer, assembling the handcrafted features, age and breast density, yielded the results that are presented in Table 7. Note that, generally, the AUC mean values are similar or higher than the previous ones. The best performance was achieved when combining shape with texture features (AUC mean = 0.801), shape and texture features with density (AUC mean = 0.800), intensity, shape and texture features (AUC mean = 0.799), and intensity, shape and texture features with density (AUC mean = 0.799). Shape and texture seem to be the handcrafted features that have the biggest influence in the classification process (AUC mean = 0.796 and 0.782, respectively), which is accordance with the fact that irregular shapes and heterogeneous textures are related to malignant lesions (Pisco 2003; Yi et al. 2017).

Table 5. Results in bold correspond to the best mean AUC values achieved for each network. CNN applied to images with and without normalisation. Training on the merged train and validation sets and testing on the test set.

Network	Batch size	Epochs	Learning rate	Norm	Top1err train mean	Top1err test mean	AUC mean	AUC std	Time approx. (min.)
CNN-F	256	50	5e−3	Yes	0.196	0.378	0.721	0.001	28.31
CNN-F	256	50	1e−2	No	0.171	0.296	**0.767**	0.002	25.81
CNN-M	64	50	5e−4	Yes	0.173	0.355	0.733	0.002	76.84
CNN-M	64	50	5e−4	No	0.146	0.288	**0.776**	0.006	76.25
Caffe	256	50	5e−2	Yes	0.324	0.488	0.584	0.004	27.28
Caffe	256	50	5e−3	No	0.190	0.251	**0.813**	0.001	25.24

Table 6. Results in bold correspond to the best mean AUC values achieved for each network. SVM classification mean results with different activation layers obtained with pre-trained model Caffe without image normalisation (performed five times).

Act. layer	AUC	Accuracy	Sensibility	Specificity	Precision	f-Measure
16	**0.773**	0.702	0.538	0.844	0.742	0.625
17	0.749	0.685	0.557	0.794	0.694	0.618
18	0.766	0.689	0.523	0.839	0.724	0.609
19	0.732	0.678	0.549	0.781	0.684	0.609
16&17	0.767	0.702	0.546	0.835	0.735	0.628
16&18	**0.772**	0.700	0.529	0.846	0.743	0.616
16&19	0.760	0.695	0.542	0.817	0.716	0.619
17&18	0.760	0.692	0.535	0.823	0.719	0.613
17&19	0.744	0.679	0.551	0.786	0.684	0.610
18&19	0.755	0.687	0.527	0.822	0.713	0.606
16&17&18	0.769	0.700	0.541	0.836	0.735	0.623
16&17&19	0.761	0.695	0.545	0.817	0.719	0.619
16&18&19	0.762	0.694	0.539	0.823	0.718	0.618
17&18&19	0.754	0.689	0.530	0.823	0.716	0.609
16&17&18&19	0.763	0.693	0.540	0.827	0.728	0.617

Table 7. Results in bold correspond to the best mean AUC values achieved for each network. The influence of age, density and HCfeats (intensity, shape and texture) combined with features from the sixteenth layer activations in the classification performance of SVM (mean values – performed five times).

	Standalone	Age	Density	Age&density
Age	0.766	–	–	–
Density	0.773	–	–	–
Age&density	0.764	–	–	–
Intensity (I)	0.774	0.766	0.772	0.764
Shape (S)	**0.796**	0.788	**0.796**	0.788
Texture (T)	0.782	0.775	0.780	0.773
I&S	**0.796**	0.789	**0.796**	0.788
I&T	0.779	0.772	0.778	0.770
S&T	**0.801**	0.794	**0.800**	0.793
I&S&T	**0.799**	0.794	**0.799**	0.792

Age, density and intensity, when used alone, do not appear to have influence in the classification performance, since no increase in AUC mean values was observed. On the other hand, when combined with others, sometimes they deteriorate the classification performance.

6 Conclusions

In this paper, we studied the application of CNNs to the problem of mammogram lesion classification. We have evaluated three different implementations of CNNs and two approaches of image normalisation.

In terms of the results obtained with the three different CNNs implementations, in the case of the normalised images with the testing set, the results decreased substantially comparatively to the previous results obtained with the validation set, mostly in Caffe model, which yielded an AUC mean value of 0.584. When the images where fed to the networks without normalisation, in the testing set, the Caffe model achieved the best AUC mean value (0.813), followed by CNN-M and CNN-F, 0.776 and 0.767, respectively.

Regarding the image normalisation, the results reveal that the normalisation process proposed by Arevalo et al. (2016) decreases the classification performance of the networks. Perre et al. (2018) considered that the image normalisation method that is chosen can change the classification performance, depending on which type of pre-trained CNN model is selected.

The fact that all crop images were composed by the surrounding breast pattern (e.g. instead of being zero padded) and, in some cases, that the lesion was not centred, may have been an advantage for the CNN learning process without confounding factors.

After we extracted the different activations' layers and applied them to an SVM, we found that the best performance was achieved using the sixteenth layer, with an AUC mean value equal to 0.773. However, the classification result is lower than the one obtained with the respective CNN model (Caffe), which goes against the opinion of Arevalo et al. (2016), since they consider that the SVM has a better performance than the softmax classifier used in the CNN.

The handcrafted features increased the classification performance. Besides, their combined use allowed to achieve an AUC mean value of 0.799. Shape and texture were the preferred information to the SVM and their combined use resulted in an AUC mean value of 0.801. Breast density showed great influence in the classification when combined with shape and texture for example, with an AUC mean value of 0.800. Intensity and age did not show great influence in the classification performance.

As future work, we intend to: test another type of classifiers with the CNN features as input; train the CNN in one database and testing them in a different database, and instead of doing a binary classification, try to include another label called 'normal' to detect normal tissue.

Acknowledgements

This work was supported by National Funding from the FCT – Fundação para a Ciência e a Tecnologia, through the UID/EEA/50008/2013 Project. The database used in this work was a courtesy of MA Guevara and coauthors, Breast Cancer Digital Repository Consortium.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the FCT – Fundação para a Ciência e a Tecnologia, through the UID/EEA/50008/2013 Project.

ORCID

Ana C. Perre http://orcid.org/0000-0001-6668-2620

References

Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. 2016. Representation learning for mammography mass lesion classification with convolutional neural networks. Computer Methods and Programs in Biomedicine. 127:248–257.

Chatfield K, Simonyan K, Vedaldi A, Zisserman A 2014. Best scientific paper award return of the devil in the details: delving deep into convolutional nets. British Machine Vision Conference. Accessed arxiv.org. https://arxiv.org/pdf/1405.3531v4.pdf.

Ganesan K, Acharya U, Chua C, Min L, Abraham K, Ng K. 2013. Computer-aided breast cancer detection using mammograms: a review. IEEE Reviews in Biomedical Engineering. 6:77–97.

Jalalian A, Mashohor S, Mahmud H, Saripan M, Ramli A, Karasfi B. 2013. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. Clinical Imaging. 37:420–426.

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T 2014. Caffe: convolutional architecture for fast feature embedding. Cornell University Library. Report No. : arXiv:1408.5093, Accessed: https://arxiv.org/abs/1408.5093 .

Jiao Z, Gao X, Wang Y, Li J. 2016. A deep feature based framework for breast masses classification. Neurocomputing. 197:221–231.

Perre A, 2018. The influence of image normalization in mammographic classification with CNNs. APPIS 2018 – 1st International Conference on Applications of Intelligent Systems; Las Palmas de Gran Canaria, Spain.

Pisco JM. 2003. Imagiologia basica texto e atlas. 1st ed. lisboa: LIDEL.

Sampat M, Markey M, Bovik A. 2005. Computer-aided detection and diagnosis in mammography. Handbook of image and video processing. Cambridge, MA: Academic Press Books - Elsevier

Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging. 35(5):1285–1298.

Sun W, Tseng TLB, Zhang J, Qian W. 2016. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Computerized medical imaging and graphics 57: 4–9.

Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Transactions on Medical Imaging. 35(5):1299–1312.

Tang JS, Agaian S, Thompason I. 2014. Guest editorial: computer-aided detection or diagnosis (CAD) systems. IEEE Systems Journal. 8:907–909.

Vedaldi A, Lenc K. 2015. Matconvnet – convolutional neural networks for Matlab. Proceeding of the ACM Int Conf on Multimedia.Cornell University Library. Report No.: arXiv:1412.4564, https://arxiv.org/abs/1412.4564

Wang D, Khosla A, Gargeya R, Irshad H, Beck A 2016. Deep learning for identifying metastatic breast cancer. Cornell University Library. Report No.: arXiv: submit/1591239. Accessed: https://arxiv.org/pdf/1606.05718.pdf.

Wichakam I, Vateekul P 2016. Combining deep convolutional networks and svms for mass detection on digital mammograms. 2016 8th International Conference on Knowledge and Smart Technology (KST): 239–244.

Yi D, Sawyer RL, Cohn III D, Dunnmon J, Lam C, Xiao X, Rubin D 2017. Optimizing and visualizing deep learning for benign/malignant classification in breast tumors. Report No.: arXiv:1705:06362v1. Accessed: http://search.ebscohost.com/login.aspx?direct=true&site=eds-live&db=edsarx&AN=1705.06362.