Original Article

A Novel Solution Based on Scale Invariant Feature Transform Descriptors and Deep Learning for the Detection of Suspicious Regions in Mammogram Images

Alessandro Bruno[1], Edoardo Ardizzone[2], Salvatore Vitabile[3], Massimo Midiri[3]

[1]Faculty of Media and Communication, Department - NCCA (National Centre for Computer Animation) at Bournemouth University, Poole, Dorset, United Kingdom, [2]Department of Engineering at Palermo University, [3]Department of Biomedicine, Neuroscience and Advanced Diagnostic at Palermo University, Palermo, Italy

Abstract

Background: Deep learning methods have become popular for their high-performance rate in the classification and detection of events in computer vision tasks. Transfer learning paradigm is widely adopted to apply pretrained convolutional neural network (CNN) on medical domains overcoming the problem of the scarcity of public datasets. Some investigations to assess transfer learning knowledge inference abilities in the context of mammogram screening and possible combinations with unsupervised techniques are in progress. **Methods:** We propose a novel technique for the detection of suspicious regions in mammograms that consist of the combination of two approaches based on scale invariant feature transform (SIFT) keypoints and transfer learning with pretrained CNNs such as PyramidNet and AlexNet fine-tuned on digital mammograms generated by different mammography devices. Preprocessing, feature extraction, and selection steps characterize the SIFT-based method, while the deep learning network validates the candidate suspicious regions detected by the SIFT method. **Results:** The experiments conducted on both mini-MIAS dataset and our new public dataset Suspicious Region Detection on Mammogram from PP (SuReMaPP) of 384 digital mammograms exhibit high performances compared to several state-of-the-art methods. Our solution reaches 98% of sensitivity and 90% of specificity on SuReMaPP and 94% of sensitivity and 91% of specificity on mini-MIAS. **Conclusions:** The experimental sessions conducted so far prompt us to further investigate the powerfulness of transfer learning over different CNNs and possible combinations with unsupervised techniques. Transfer learning performances' accuracy may decrease when the training and testing images come out from mammography devices with different properties.

Keywords: *Classification, computer-assisted image processing, computing methodologies, deep learning, digital mammography*

Submitted: 15-Jun-2019 **Revised:** 01-Oct-2019 **Accepted:** 06-May-2020 **Published:** 03-Jul-2020

Introduction

Among global female population, breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death.[1] The scientific community made a lot of efforts over the last decades to improve the diagnostic accuracy of breast cancer in women. Reading mammograms is a time-demanding and tiring job; about 30% of cancers are missed on mammograms (false negatives), but recent tests and studies showed that computer-aided diagnosis (CAD) software for mammography allows for increase in radiologist sensitivity.[2,3] The risk of dying from breast cancer has dropped by >20%, according to International Agency for Research on Cancer scientific papers, in areas where screening mammograms programs have been conducted, and by as much as 40% among women who undergo screening mammograms regularly.[4] The objective of CAD systems is to draw radiologist attention to possible abnormalities in mammography, reducing the number of false positives and false negatives; according to latest scientific studies, computer-aided detection of breast cancer can improve the detection rate from 4.7% to 19.5% compared to radiologists.[5] It is observed that breast cancer is characterized with mass showing irregular appearance, linear spicules, and blurred boundaries. On the other side, benign masses usually have a well-circumscribed border. The Breast Imaging Reporting and

How to cite this article: Bruno A, Ardizzone E, Vitabile S, Midiri M. A novel solution based on scale invariant feature transform descriptors and deep learning for the detection of suspicious regions in mammogram images. J Med Signals Sens 2020;10:158-73.

Address for correspondence:
Dr. Alessandro Bruno, NCCA (National Centre for Computer Animation) at Bournemouth University, Bournemout, Dorset, UK. E-mail: abruno@bournemouth.ac.uk

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.JMSS_31_19

Quick Response Code:

Data System showed that descriptors such as shape, size, and margins are useful to characterize the abnormalities or masses present in breast cancer.[6] Moreover, according to scientific studies, masses are mainly grouped with respect to their size: small size (3–15 mm), middle size (15–30 mm), and large size (30–50 mm).[7] The scientific community put a lot of effort into biomedical imaging tasks which showed alternating results; several approaches for suspicious region detection have been proposed.[8-12]

Image features such as interest keypoints, local and edge descriptors, intensity, perimeter, area, geometrical shape, compactness, and orientation are often used to perform mammogram patch classification.[13-15] Ali and Hamed[16] give a the state-of-the-art survey for early breast cancer detection on mammograms. Min *et al*.[17] used features such as area, perimeter, circularity, and density to characterize the shape of a mass. Texture features are widely adopted to detect clustered micro-calcification in digitized mammograms.[18] In a study,[19] the authors adopted Gabor filters to detect architectural distortions and abnormalities in mammograms. Anitha and Dinesh[20] automatically detected and segmented the suspicious mass regions of mammogram using a modified transition rule named maximal cell strength updation in cellular automata. Tingting *et al*.[21] used Kernel principal component analysis to improve the discriminating power of each single feature extracted from the image. A team of researchers[22] discriminated fatty from dense mammograms by using correlation-based feature selection and sequential minimal optimization. In a study,[23] the authors quantified and estimate the size of abnormalities in mammograms with Scale Invariant Feature Transform (SIFT). Aize *et al*.[24] proposed a robust information clustering algorithm incorporating spatial information for breast mass detection. Kai *et al*.[25] used a combination of adaptive global thresholding segmentation and adaptive local thresholding segmentation on a multiresolution representation of the original mammogram. Pereira *et al*.[26] achieved good results in terms of segmentation and detection of suspicious regions on mammograms by using a combination of wavelet analysis and genetic algorithms. Sampaio *et al*.[27] adopted cellular neural network (CelNN) and support vector machine (SVM) as tools for the detection of masses on mammograms. A method for mass enhancement using piece-wise linear operator in combination with wavelet processing from mammographic images is proposed by Vikhe and Thool.[28] A group of researchers proposed a method based on Dual-Stage Adaptive Thresholding (DuSAT).[29] In greater detail, they detected suspicious mass region by using global histogram and local window thresholding method. The global thresholding is done based on the histogram peak analysis of the entire image, and the threshold is obtained by maximizing the proposed threshold selection criteria. The topic of the detection of suspicious regions in mammograms has been widely addressed by the biomedical community[30] on several application areas such as neuro, retinal, pulmonary, digital pathology, breast, cardiac, abdominal, and musculoskeletal. A lot of water passed under the bridge since Berkman *et al*.[31] proposed convolutional neural networks (CNNs) to classify regions of interest in mammograms. Since then, much of progress has been done in the matter of hardware throughput computation, making deep learning methods[32] (which are very resource demanding) more accessible. Because of the aforementioned reason, a growing number of biomedical imaging methods recently addressed the detection of suspicious regions in mammograms using deep learning solutions. Accordingly, we give a brief list of different state-of-the-art methods as follows. Pengcheng *et al*.[33] employed CNNs to build a classifier for detecting and localizing the abnormalities in digital mammography; they reported VGGNet results achieving the best accuracy up to 92.53% in patch classification. Some scientists[34] assessed the performance of several CNN architectures over Digital Database for Screening Mammography (DDSM) dataset for the task of mass classification. In the recent study by Tsochatzidis *et al*.,[35] the authors addressed the CNN evaluation on mammograms with two different training scenarios where pretrained weights and random fashion weights are adopted to train the nets. Jung *et al*.[36] proposed a mass detection on mammograms based on a deep learning object detector called RetinaNet with good results. The method by Cai *et al*.[37] is focused on the study of calcification clusters as early sign of cancer; they characterized calcification with descriptors obtained from deep learning and handcrafted descriptors. Richa *et al*.[38] showed comparisons between different CNN architectures such as VGG16, ResNet50, and IcenptionV3 for the purpose of mass detection in mammograms. Thijs *et al*.[39] presented a comparison between CNN and a CAD system on a large mammograms dataset. Arfan[40] used the combination of CNN and SVM to detect suspicious regions in mammograms. Wang *et al*.[41] dealt with the discrimination of breast cancer with microcalcifications. Michiel *et al*.[42] applied unsupervised deep learning to address Breast Density Segmentation and Mammographic Risk Scoring evaluation. Yamashita *et al*.[43] provided an extensive survey on the CNNs applications over different tasks in radiology.

Ribli *et al*.[44] proposed a CAD system based on Faster R-CNN, which allows for the detection and classification of malignant or benign lesions on a mammogram in a fully automatic way. A limitation of Ribli *et al*.'s method is the small size of the publicly available dataset. Tavakoli *et al*.[45] came up with a new method based on CNNs and a decision scheme (CNNs + DS). In greater detail, the authors first used a preprocessing block around each pixel that, then, is fed into a trained CNN to determine whether the pixel belongs to normal or abnormal tissues on images from the Mini-MIAS dataset.[46] A classifier based on the

combination of cascade of deep learning and random forest is proposed by Dhungel *et al*.[47] First, a multi-scale deep belief network selects suspicious regions, which are processed by a cascade of deep CNNs (DCNN). Only those regions which are detected by this deep learning analysis go through a two-level cascade of random forest classifiers. The resulting regions are then combined using connected component analysis. As well as, Ribli *et al*. and Dunghet *et al*. conducted their experiments over both DDSM[48] and Inbreast[49] datasets. Akila *et al*.[50] recently proposed a method called MA-CNN (Multiscale All CNN) to classify normal, benignant, and malignant tissue (cancerous) over the Mini-MIAS dataset. Dhungel *et al*.[51] also proposed a deep learning technique based on a two-step training process which employs the learning of a regressor that is function of the values of handcrafted features from the Inbreast dataset. The previous steps are followed by a fine-tuning stage that learns the breast mass classifier. Arevalo *et al*.[52] developed a method comprising two main stages: the first one is a preprocessing to enhance image details, whereas the second one is a supervised training for learning both the features and the breast imaging lesion classifier from Breast Cancer Digital Repository.[53] In the study by Teare *et al*.,[54] the authors provided a solution to detect suspicious regions on images from DDSM based on the use of genetic search of image enhancement methods and a Dual DCNNs. Huynh *et al*.[55] compared SVM based on image features extracted by a CNN and their prior computer-extracted tumor features, aiming to discriminate benign from malignant breast lesions.

They processed images from a University of Chicago Medical Center Dataset. Levy *et al*.[56] focused their efforts on using transfer learning and techniques such as data augmentation and preprocessing to overcome the training data limitations in DDSM. Agarwal *et al*.[57] compared the widely adopted CNNs such as VGG16, ResNet50, and InceptionV3 over DDSM and showed Inception V3 overcoming the other networks.

Much progress has been made over the last decades, as reported in the current section. A list of several state-of-the-art methods is also reported in Table 1.

Remarkably, some methods can reach out to high accuracy levels (using different reported results) over the task of the detection of suspicious regions on mammograms. On the other side, we want to point out that most of the scientific literature methods focus on the mentioned task over mammograms belonging to datasets with very similar properties, such as of spatial resolution and image dimensions. Because of the above reasons, we focus our work over two datasets equipped with pictures having dissimilar properties. We want to stress out the performances of transfer learning over the detection task.

In our method, we consider as suspicious all those regions that include abnormalities such as calcification, well-defined and circumscribed masses, spiculated masses, ill-defined masses, architectural distortion, and asymmetries. The main contributions of our paper can be summed up as it follows: a new solution for the detection of suspicious regions in mammogram images using the integration of a new SIFT-based approach and a deep learning technique with transfer learning; an experimental investigation on the transfer learning paradigm ability to predict suspicious region models from different mammograms; a comparison with some state-of-the-art methods over Mini-MIAS; and, the last but not the least, the sharing of our own new public mammogram dataset called[58] Suspicious Region Detection on Mammogram from PP (SuReMaPP) hand-labeled by three expert radiologists.

Materials and Methods

In this section, first we give, in order, the overall architecture of our integrated solution, a more detailed description of the two techniques that compose the integrated solution. Before moving to the description of each technique, we want to point out that the objectives of our work are mainly three as follows:

- To provide a new solution to detect suspicious regions based on the integration of a SIFT-based algorithm and transfer learning
- To investigate the prediction power of transfer learning method in biomedical imaging comparing two CNN architectures fine-tuned over two different datasets
- To provide a new publicly available and hand-labelled mammogram dataset (SuReMaPP).

Suspicious Region Detection on Mammogram from PP

SuReMaPP consists of 343 mammograms hand-labeled by expert radiologists dealing with the identification of suspicious regions such as abnormalities (benignant and malignant) and calcifications.

SuReMaPP contains mammograms with standard bilateral craniocaudal and mediolateral oblique views. The spatial resolution depends on the mammography device used, in order, GIOTTO IMAGE SDL/W and FUJIFILM FCR PROFECT CS. The former generates images with a spatial resolution of 3584 × 2816 pixels; it is equipped with a detector size of 24 cm × 30 cm. The pixel size is 85 μm.

The latter generates images with a spatial resolution of 5928 × 4728; it is provided with a detector of size 24 × 30 cm. The pixel size is 50 μm.

We want to share SuReMaPP dataset with the scientific community to be used as "Gold Standard" for biomedical imaging methods and algorithms. The images are accessible through the link.[58]

The 343 images from SuReMaPP involve a number 145 patients; 100 mammograms are related to 25 patients and include negative cases (no suspicious regions in them); the remaining 243 mammograms are with positive cases

Table 1: A list of some state-of-the-art methods

Author	Method	Dataset	Reported results	Pros and cons
Cao *et al*.	Spatial clustering[24]	Mini-MIAS	Sensitivity 88.7%	The method is based on a RIC algorithm. It would be interesting to assess the performance over other datasets either
Ribli *et al*.	Faster R-CNN[44]	INbreast	AUC 0.85, sensitivity 90%, and 0.14 false-positive marks per image	It sets the state-of-the-art classification performance on INbreast. The size of the publicly available dataset is small
Hu *et al*.	Adaptive thresholding[25]	Mini-MIAS	Sensitivity 91.8%	The global and local thresholds are chosen adaptively without artificial intelligence. Tests over mammograms with different spatial resolutions are missing
Huynh *et al*.	Transfer learning from deep convolutional neural networks[55]	Dataset from the University of Chicago Medical Center. 219 digital mammograms and 607 ROIs	AUC 0.86 and true positive and false-positive fractions	The article shows performances of several architectures The results are reported only on their own dataset
Xi *et al*.	Deep convolutional neural networks[33]	DDSM dataset	Accuracy 95%	The authors investigated the powerfulness of some state-of-the-art CNNs
Pereira *et al*.	Multilevel thresholding[26]	Mini-MIAS	Sensitivity 90%	The method runs both detection and segmentation tasks. The detection accuracy rate is high, while segmentation performs a bit lower
Dunghel *et al*.	Combination of cascade of deep learning and random forest[47]	DDSM and INbreast	True positive rate of 0.96±0.03 at 1.2 false positive per image on INbreast. True positive rate of 0.75 at 4.8 false positive per image on DDSM	The method achieves very good performances on both datasets. The computational burden of the method seems to be quite expensive (the execution time is almost 20 s)
Tavakoli *et al*.	CNNs and a decision scheme[45]	Mini-MIAS	Sensitivity 93.33%	ROIs in the proposed method are not rescaled to preserve the quality of the image
Burçin *et al*.	Havrda and Charvat entropy and Otsu N thresholding[73]	Mini-MIAS	Sensitivity 90.2%	The method detects abnormalities from mammograms using an unsupervised approach. A check of the robustness of the features extracted over another dataset is missing
Akila Agnes *et al*.	Multiscale all convolutional neural network[50]	Mini-MIAS	Sensitivity 96% and 0.99 AUC	The method exhaustively exploits the powerfulness of CNN over Mini-MIAS reaching out impressive performances
Sampaio *et al*.	Cellular neural network[27]	Mini-MIAS	Sensitivity 90.9%	The method allows for detecting and segmenting suspicious regions even though the latter task has some drawbacks (10% of masses were lost)
Levy and Jain	Deep convolutional neural networks[56]	DDSM	Accuracy 92.9%, precision 92.4%, recall 93.4%	Preprocessing, data augmentation, and transfer learning steps are run to obtain state-of-the-art performances
Vikhe and Thool	Wavelet processing and adaptive thresholding[28]	Mini-MIAS and DDSM	Sensitivity 91%	The method runs suspicious region detection on two subsets of the existing databases reaching out more or less the same accuracy levels
Anitha *et al*.	WPAT, Dual-Stage adaptive thresholding[29]	Mini-MIAS	Sensitivty 93%	The method relies upon dual-stage adaptive thresholding which is, at the same time, dependent on pectoral muscle removal step
Teare *et al*.	Genetic search of image enhancement methods and a dual deep convolutional neural networks[68]	DDSM and ZMDS	Specificity 91% Specificity 80%	False-color enhancement technique to mammography images and utilizing a dual deep CNN engine. Some details on the reliability of the whole system are missing

Contd...

Table 1: Contd...

Author	Method	Dataset	Reported results	Pros and cons
Jaffar	DuSAT, deep convolutional neural network with support vector machine[40]	Mini-MIAS, DDSM	Sensitivity 93.25%	Performances over two different datasets are very similar. Comparisons over high-resolution images are missing

CNNs–Convolutional neural networks, WPAT–Wavelet processing and adaptive thresholding, DuSAT–Dual-stage adaptive thresholding, DDSM–Digital Database for Screening Mammography, MIAS–Medical image analysis, RIC–Robust information clustering, AUC–Area under the curve, ROIs–Regions of Interest, ZMDS–Zebra Mammography Dataset

(both malign and benign suspicious areas) of 120 patients (2–3 views per case).

Patients whose mammograms are in SuReMaPP are aged 41–62 years. Both SuReMaPP undergo data augmentation to increase the number of patches for the fine-tuning step in transfer learning. More details will be given in the next section.

Data augmentation

Data augmentation[59] is a well-suited technique for reducing overfitting; there exist several methods for data augmentation; we adopt geometric transforms to increase the size of the original images from SuReMaPP. We extract patches from both Mini-MIAS and our dataset to be used with CNNs (PyramidNet and AlexNet). The first layers of the PyramidNet and AlexNet CCNs are, in order, designed with a spatial of 224 × 224 and 227 × 227 pixels. We want to point out that transfer learning needs many images for fine-tuning the CNN over a specific image category. The data augmentation we use generates image transforms such as translations, horizontal reflections, and crop.

Furthermore, the standard approach for data augmentation suggests to extract random patches (and their horizontal reflections) from the original images and train the network on these extracted patches. A reasonable requirement regarding the data augmentation technique is that the number of images per category has to be well balanced. Although the extraction of nonsuspicious patches from the mammographies with no labeled regions is a straightforward process, the extraction of patches with suspicious regions is neither a straightforward nor an immediate step. Our algorithm starts the extraction of patches from the centroids of regions labelled by the radiologists [the blue dot in Figure 1a] using a partially overlapped window-sized 224 x 224 pixels (or 227 by 227). In greater detail, a given window centred on the centroid of a labelled suspicious region is further divided into 16 sub-blocks. The red vertices [Figure 1a] will be in turn, the centres of the new patches used for the data augmentation. For a given mammogram, the window centered on the centroid of the suspicious region is uniformly divided into 16 sub-blocks whose red vertices [Figure 1a] will be, in turn, the centers of all the patches with a suspicious region. This algorithm allows us to extract from 9 to 17 patches per mammogram depending on the size and the coordinates of the suspicious

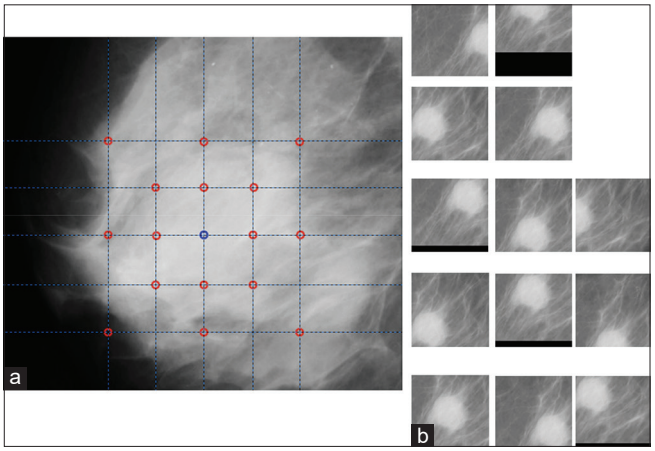

Figure 1: A patch sample from Suspicious Region Detection on Mammogram from PP (a) and a sample of patches generated with data augmentation (b)

region in the image (we do not include the patches that partially fall outside the mammogram). In addition to the translation, we also applied horizontal reflection to increase the size of the dataset [an example of data augmentation is in Figure 1b].

The Mini-MIAS dataset consists of 322 images, digitized at 50-μ pixel edge, with different sized suspicious regions: small regions with radium lower than 28 pixels; medium regions with radium larger than 28 pixels and lower than 57 pixels; and large regions with radium larger than 57 pixels.

The SuReMaPP dataset is composed of 343 mammograms with high spatial resolutions depending on the mammography device (in order, the first device generates mammogram with a spatial resolution of 3584 × 2816 pixels and the second one makes mammograms with a spatial resolution of 5928 × 4728).

To respect the proportions of the region of interest of the mammograms, we decide to extract patches with the following spatial resolutions: 448 × 448 pixels for the mammograms with an original spatial resolution of 3584 × 2816; 774 × 774 pixels for the mammograms whose original spatial resolution is 5928 × 4728. Then, the patches are resized to the spatial resolutions of the first layers of the CNNs (224 × 224 or 227 × 227).

We extract a total of 3206 patches from Mini-MIAS and 4914 patches from SuReMaPP. The data are well balanced

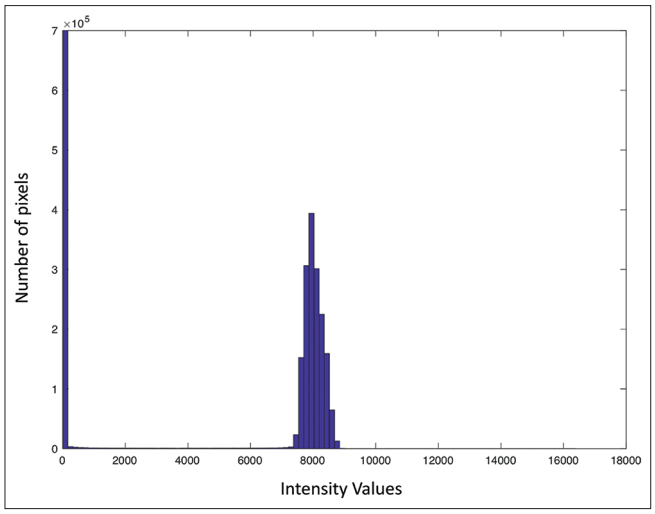

Figure 2: The histogram of a sample from the Suspicious Region Detection on Mammogram from PP dataset is given (two main modalities can be observed)

with an approximately equal number of patches along with suspicious and nonsuspicious areas.

Scale invariant feature transform-based technique

For a given mammogram [Figure 1a], all pixel values are converted in double data type, then the image histogram is analyzed. Two main modalities are usually shown in the histogram of mammograms [Figure 2]; the first one is related to the background black pixels [the first peak of the histogram in Figure 2], which can be filtered out for our purpose because we are only interested in patches containing breast profile, and the second one [see the second peak of the histogram as shown in Figure 2] contains information about the foreground pixels that describe the pixel intensity for each breast profile region. We simply discard all pixels belonging to the first modality of the histogram. It is well known that SIFT descriptors are extracted along boundaries, edges, spikes. and, more generally. the local maxima of Laplacian of Gaussian across different scales of the same image.[14] We want SIFT keypoints to be extracted on both image details (borders, edges, and spikes) and structural components; to address this matter, we decide to treat two different versions of the mammogram, a version with lower gray levels dynamic range image and a version with higher gray levels dynamic range. The image with high dynamic range in the histogram will show a lot more of details and edges than the one with low dynamic range in the histogram, which in turn will highlight the structural component of the image. We transform the image using two histogram-fitting functions: the logistic fitting function [Figure 3a] and a nonparametric kernel-smoothing distribution [Figure 3b]. The logistic distribution, described as in Eq. 1, is used for growth models and in logistic regression. The logistic distribution equation is characterized by mean (μ) and sigma (σ) parameters of the pixel gray levels. As far as it concerns, the nonparametric kernel-smoothing distribution described in Eq. 2, K is a nonnegative function (the kernel function) and $h > 0$ is a smoothing parameter called the bandwidth that controls the smoothness of the resulting probability density curve. In our method, h is set 0.337 which best approximates a standard normal distribution of data.[60]

After all pixel values from an input mammogram [Figure 4a] are converted in double data type, then the image histogram is analyzed. It is observed in Figure 4b that the mentioned histogram specifications allow in order for obtaining an image with lower dynamic gray-level range [upper row in Figure 4b] and an image with higher dynamic gray-level range from the original mammogram image [lower row in Figure 4b]. When the histogram specifications are applied, we move forward to image inspection with SIFT local keypoints and descriptors. SIFT keypoints are extracted on both two versions of the image considering different aspects that will be described below. We deliberately discard the keypoints having negative Laplacian values because of points located close to the edge of the breast.

$$f(x/\mu,\sigma)=\frac{\exp\left\{\frac{x-\mu}{\sigma}\right\}}{\sigma\left(1+\exp\left\{\frac{x-\mu}{\sigma}\right\}\right)^2} \quad (1)$$

$$f(x;h)=\frac{1}{nh}\sum_{i=1}^{n}k\left\{(x-x_i)/h\right\} \quad (2)$$

The extraction of SIFT[61] is mainly characterized by two parameters: the peak threshold and the edge threshold. The edge threshold allows eliminating peaks of the Difference of Gaussians (DoG) scale space with small curvature. The peak threshold parameter filters out the peaks of the DoG space scale, showing low absolute values. Both settings, as suggested by the scientific literature,[61,62] are needed to be set experimentally for the specific task of interest. In order, we set 0.01 as the value of the peak threshold and five as the value of the edge threshold.

We set the number of octaves to four and the number of scale levels to five as suggested to be the optimal values for the SIFT algorithm.[61]

We also select the keypoints with radius parameter larger than 3 mm and lower than 50 mm, as scientific literature[7] suggests that regions with size smaller than 3 mm or >50 mm not being significant for diagnosis [Figure 4c].

Furthermore, a validation step based on Euclidean distance set to 50 is needed on both the mammogram to establish the spatial coherence between the matching points between two specified versions of the mammogram.

The intersections between pairs of keypoints which pass through the above-mentioned steps detect candidate suspicious regions.

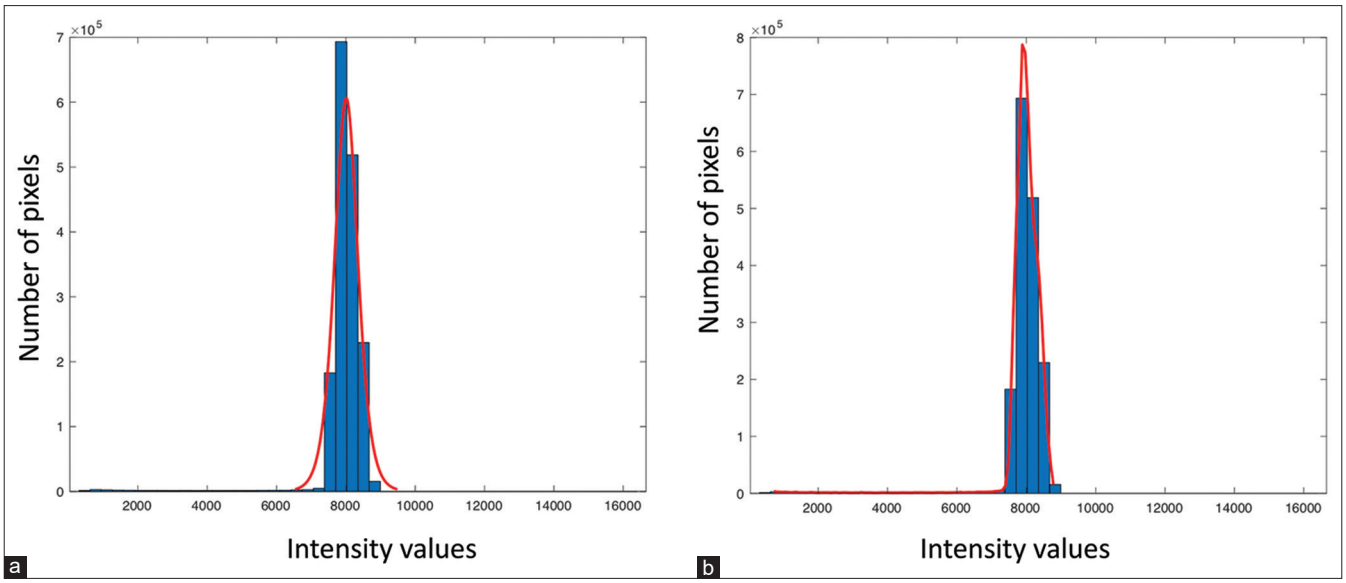

Figure 3: The logistic function (a) and the nonparametric kernel-smoothing distribution (b) are used as fitting functions of the histogram for the Breast profile regions. These functions are, then, used to generate two new versions of the given mammogram

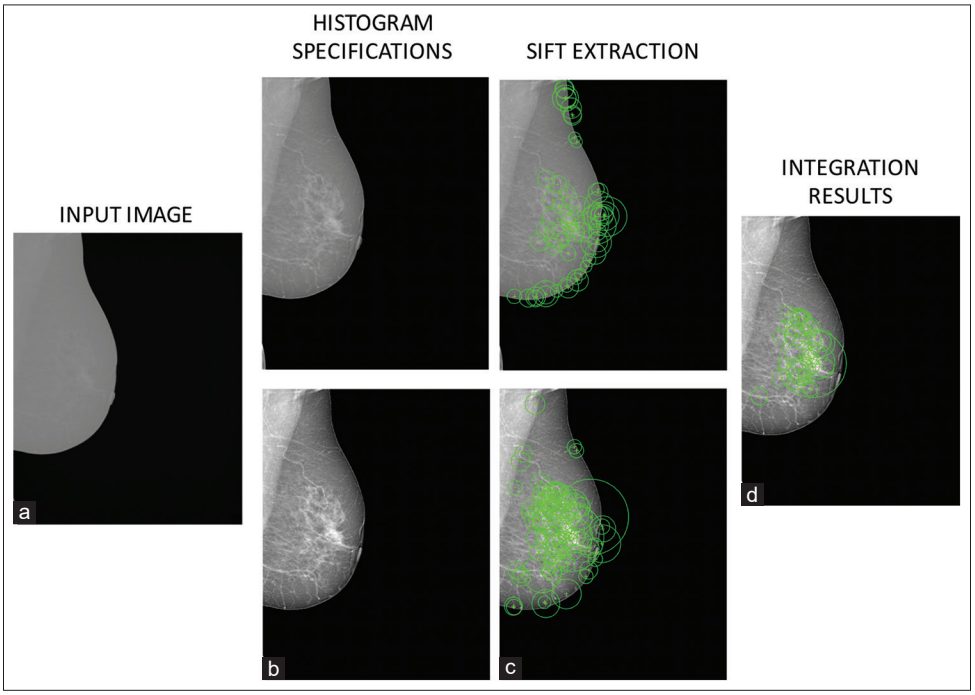

Figure 4: The overall working scheme of the Scale Invariant Feature Transform based module is represented with respect to all the steps which it is made of: (a) the input image is specified into two new mammograms (b) using the logistic function and the nonparametric kernel-smoothing distribution [Figure 2]; Scale invariant feature transform keypoints are extracted on both the mammogram versions considering the radius parameter and discarding those keypoints with negative Laplacian (c), then the intersection between all the keypoints extracted on both mammogram versions is performed as a sort of result integration (d)

As we are interested in assessing the effectiveness of this unsupervised approach, we test it over a subset of 200 images of our own public dataset SuReMaPP, which will be further described in the next sections. The SIFT-based technique returns keypoints, which should be located in suspicious regions. Providing that the datasets we employed in this study have been manually labeled by radiologists, we have knowledge of where the suspicious regions are in the images. Therefore, after the running of the technique over data, we evaluate this method by counting the number of true positive and false positives. For the sake of clarity, keypoints returned by the SIFT-based method which fall within suspicious regions in the image are considered true positive, whereas keypoints falling within nonsuspicious regions are considered false positives. The experimental results show 85% of specificity in spite of a nonnegligible average number

of false positives per image (10 keypoints are on the average detected in nonsuspicious regions per image). We count the number of keypoints and compare the locations with respect to our own dataset SuReMaPP to be used as gold standard. The output of the SIFT-based technique is a set of SIFT keypoints that identify the candidate suspicious regions. To make the output of the SIFT-based method compliant with transfer learning, we extract square patches centered on the keypoints. Therefore, the square patches are the candidate suspicious regions to be validated by the transfer learning module described in the next section.

Here, we want to remark that the first module of the novel technique is an unsupervised method, and the parameter tuning for the extraction of SIFT keypoints does not have any impact on the second module, which is a supervised deep learning technique.

Transfer learning

Transfer learning[63] provides a framework to leverage the already-existing and trained network in a related domain over a new task domain. In our case, we want to reuse two CNNs such as AlexNet[64] and PyramidNet[65] pretrained over ImageNet[66] to be fine-tuned over biomedical data as depicted in the overall scheme in Figure 5. In practical terms, we retrieve the pretrained versions of AlexNet and PyramidNet, and then we apply the transfer learning paradigm with data augmentation, regularization, and fine-tuning on the mammogram domain. AlexNet and PyramidNet, which are the adopted CNNs for our purpose, are designed with different architectures; this is of related interest for our study because we want to investigate the performance of transfer learning in mammogram domain by analyzing the impact of the network depth on the classification task.

CNN is a hierarchical architecture made up of several kernel filters, which allow for extracting local features from images. A standard CNN architecture is equipped with convolutional, pooling, and fully connected layers. Each structure needs to abide by some rules and constraints given by their layers' size (to mention one of the essential properties), the number of layers, pooling, stride, and hyper-parameters, which characterize the overall structure of the CNN stack. Researchers such as Zhang *et al*.,[67] Ioffee *et al*.,[68] Zhang *et al*.[69] gave in order, their contributions over utilizations of Deep Learning, CNN features off-the-shelf, Stochastic gradient descent algorithm, Accelerating Deep Networks. They represent the theoretical basis for the application of Deep Learning approaches over the biomedical domain. A feature map of a CNN is the result of the filtering of an input image. For each layer of the CNN stack, the corresponding feature map shows the partial output of the network.[70]

Because a feature map is the result of spatial filtering of an input matrix and a kernel filter, it needs to stick to filtering rules. It means that as long as convolutional filters increase their dimension with stride and pooling size, it comes out as a decrease of the size of feature maps. The latter one is the conventional method of stacking several convolutional filters. As a side effect, this approach tends to sharply downsample the input images along with the layers of CNN toward the output layer.[65]

AlexNet, as well as the most of CNN architectures, approaches the classification task with stride and pooling which sharply down sample their input loading the computational burden over the first layers in the network. The innovation brought by the PyramidNet architecture is to increase the feature maps, gradually distributing the computational burden across all network units.[65]

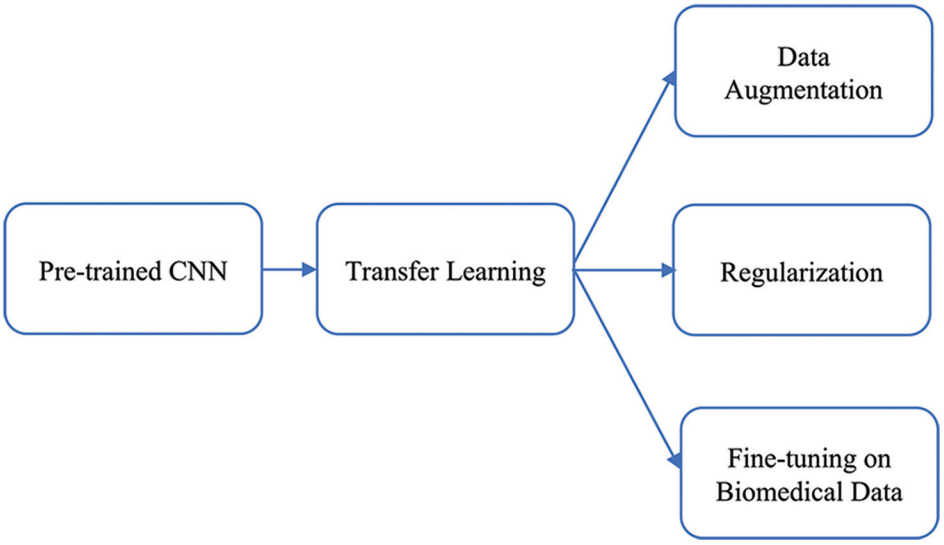

Figure 5: The overall scheme of the deep learning technique we adopt for our purpose: employment of pretrained convolutional neural networks, transfer learning, data augmentation, regularization, and fine-tuning on biomedical data

Zooming in the CNN architectures, AlexNet consists of five convolutional layers, max-pooling, dropout, and three fully connected layers counting nearly 60 million parameters to be tuned during training. AlexNet is trained on a subset of ImageNet[66] data, made of 1.5 million annotated images falling within nearly 1000 categories. The PyramidNet version we choose to carry out the experiments is the one pretrained on ImageNet; it consists of 272 layers and 62.1 million parameters, and the network includes convolutional, max-pooling, dropout layers. Other than AlexNet, residual units, batch normalization, and different positions for rectified linear unit (ReLU) are used in PyramidNet for the purpose of improving the knowledge inference abilities with deeper stack. Although it is deeply and finely described in their reference papers,[64,65] we want to remind that the ReLU is used for the nonlinearity functions, while the dropout layers allow for addressing the problem of overfitting on training data. The key idea of dropout[71] technique is to randomly drop units from the neural network during training to avoid co-adapting. During training, dropout samples form an exponential number of thinned networks. The effect of all the thinned networks is approximated with using a single unthinned network with smaller weights. The dropout technique allows for decreasing the error by a 4%. Each output neuron is modeled by using ReLU rather than the standard hyperbolic tangent because of its velocity. After applying ReLU for modeling each neuron of the networks, then the neuron output $a^i_{x,\ y}$ is normalized by using local response normalization as follows:

$$b^i_{xy} = a^i_{xy}\left(k + a \sum_{j=\max(0,\frac{n}{2})}^{\min(N-1,i+\frac{n}{2})} (a^i_{xy})^2\right) \quad (3)$$

In the next sections, some more details about the experimental configuration of transfer learning in our case study are given.

Results

This section is focused on the analysis and the assessment of our method performances on different datasets such as Mini-MIAS and SuReMaPP as mentioned in the previous sections. The analysis is conducted by looking at all the steps required to design the full stack made up of SIFT-based and deep learning modules. In the next

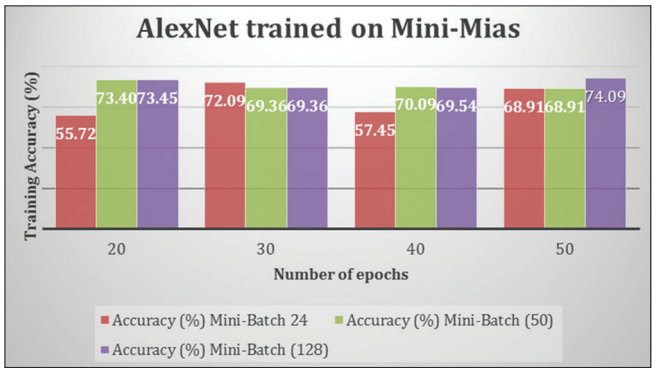

Figure 6: The training accuracy rates of AlexNET on Mini-MIAS are shown with respect to different number of epochs and Mini-Batch

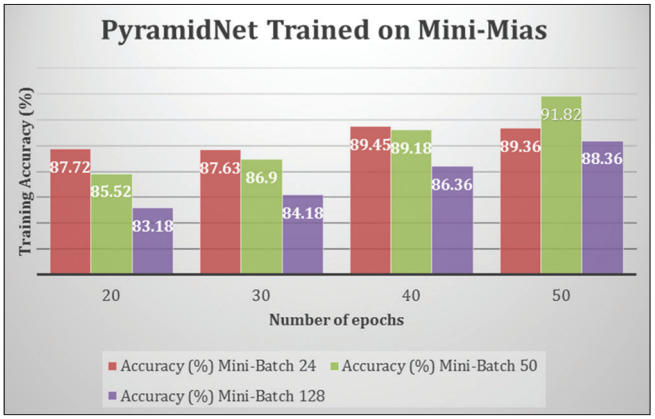

Figure 7: The training accuracy rates of PyramidNet on Mini-MIAS are shown with respect to different number of epochs and Mini-Batch

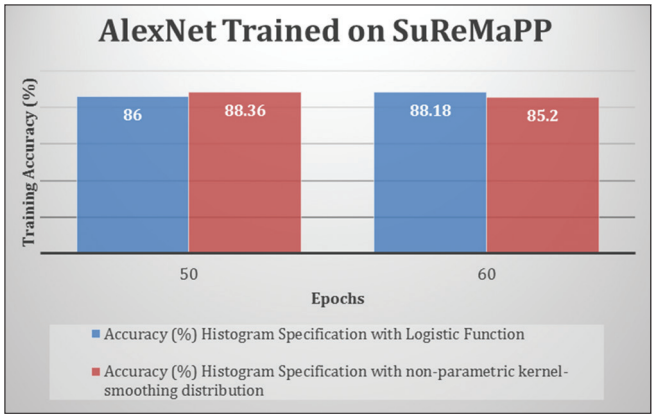

Figure 8: The training accuracy rates of AlexNET on Suspicious Region Detection on Mammogram from PP are shown with respect to different number of epochs and histogram specifications

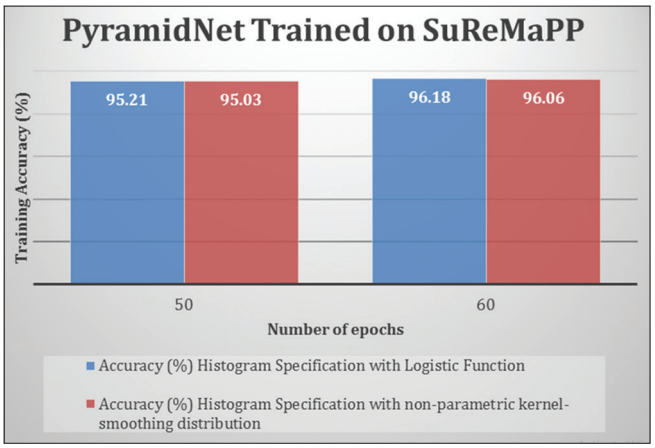

Figure 9: The training accuracy rates of PyramidNet on Suspicious Region Detection on Mammogram from PP are shown with respect to different number of epochs and histogram specifications

subsections, we describe the fine-tuning of CNNs over the histogram specified images from both Mini-MIAS and SuReMaPP and the tests conducted with the integration of the SIFT and transfer learning modules. We also analyze the pros and cons of our method by evaluating its performances with respect to other state-of-the-art methods based on different features and principles such as the ones mentioned in various studies.[24-29,40,72]

Fine-tuning of convolutional neural network

In this section, we give you details concerning the fine-tuning of PyramidNet and AlexNet to show which architecture is the more suitable for this purpose. First, we want to point out that we run two training sessions for both PyramidNet and AlexNet to fine-tune them over the mammogram domain using, in order, a subset of SuReMaPP and a subset of Mini-MIAS. In greater detail, we used a subset of three-fourth of the dataset as training set, while one-fourth is used as a test set. The performance rates you can see in Figures 6-9 are referred to the so-called validation accuracy, that is, the classification accuracy of the model over a subset of the training set (called validation set).

Training sessions are carried out using publicly available versions of AlexNet and PyramidNet, which are pretrained on the ImageNet dataset.[66] Weights of a pretrained model are preinitialized, that is a different way than as it was trained from scratch.

The fine-tuning configuration of AlexNet is as it follows. The iteration number of the fine-tuning is set to 10^4. The learning rate is 10^{-3}. The momentum parameter is set as 0.9, and weight decay is set to 5×10^{-4}. All parameters apart from the above mentioned are set to the default configuration of AlexNet.

Table 2: 5-fold cross validation performance of PyramidNet over 4916 patches from SuReMaPP

Fold test	Suspicious (%)		Nonsuspicious (%)	
	True	False	True	False
1st fold	461 (93.8)	30 (6.2)	458 (93)	34 (7)
2nd fold	463 (94.2)	28 (5.8)	464 (94.3)	28 (5.7)
3rd fold	465 (94.7)	26 (5.3)	467 (94.9)	25 (5.1)
4th fold	466 (94.9)	25 (5.1)	469 (95.3)	23 (4.7)
5th fold	466 (94.9)	25 (5.1)	471 (95.7)	21 (4.3)
Average (%)	94.5±0.48	5.5±0.48	94.64±1.05	5.36±1.05

We have a total number of 4916 patches from SuReMaPP. To apply 5-fold cross-validation, we split up the whole dataset into five subsets counting 983 patches. The remaining amount of patches are used as the training set over the category suspicious and non-suspicious regions. The process aims to detect the capabilities of the model to infer knowledge over the classification task. We repeat the steps on each of the five subsets. The results for each of the 5 groups are described in each table row using the true positives and false positives. The average percentage in the bottom row gives us a measure of the knowledge inference of the Deep Learning Model

As far as it concerns PyramidNet, weight decay is applied to all weights and biases instead of just the weights of the convolution layers. The networks are trained using backpropagation by stochastic gradient descent over ImageNet. The initial learning rate is set to 10^{-3}. The weight decay is set to 10^{-4}, and the momentum parameter is set to 0.9. The filter parameters are set using msra.[73]

As noticeable from Figures 6-9, PyramidNet outperforms AlexNet on both the datasets during the fine-tuning phase as shown in the plots in Figures 6-9. The best performances are achieved by PyramidNet with the following configurations:

1. Fifty epochs and Mini-Batch 50 across the experiments conducted on Mini-MIAS
2. Sixty epochs and the histogram specification with the nonparametric kernel-smoothing distribution across the experiments conducted over SuReMaPP.

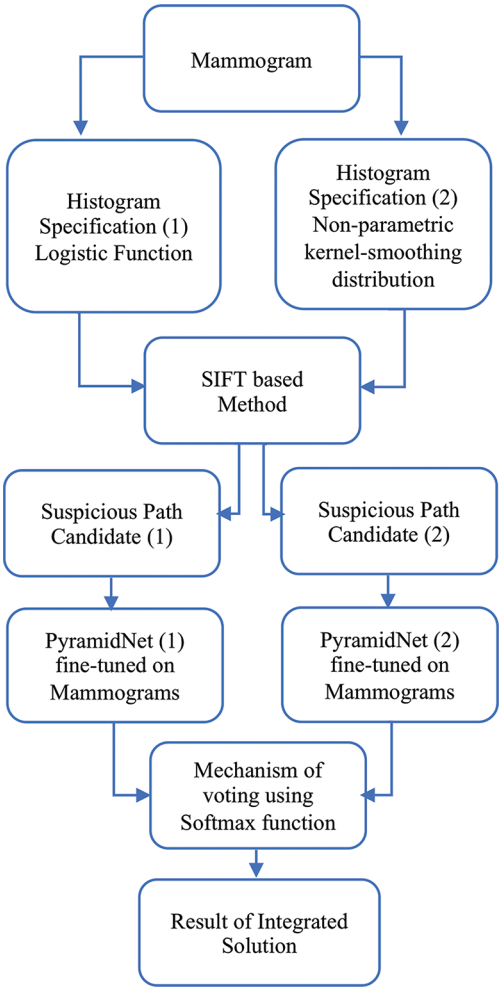

Figure 10: The overall scheme of our novel technique which consists of the integration of a scale invariant feature transform-based method and a deep learning module with transfer learning: input (mammogram image); histogram specifications (logistic function and nonparametric kernel-smoothing distribution); scale invariant feature transform-Based method which extract keypoints on candidate suspicious regions; PyramidNet fine-tuned over mammogram images (specified with the same histogram specification as in scale invariant feature transform-based method); Mechanism of voting using Softmax function

The performance analysis of the fine-tuning [Figures 6-9] prompts us to proceed by collecting our experiments using only PyramidNet as the CNN of our integrated solution. Figures 6 and 7 show the training accuracy of AlexNet and PyramidNet fine-tuned on Mini-MIAS dataset with respect to the size of Mini-Batch and the number of epochs. Mini-Batch is mainly based on the principle of running the training over image subset groups rather than over the entire dataset to extract the accuracy trend of the training step, it is widely used to save time during the training phase in deep learning methods.

Other than in Figures 6 and 7, in Figures 8 and 9, we focus our attention on the impact of the two histogram specifications on PyramidNet and AlexNet trained over SuReMaPP dataset. On the average, PyramidNet outperforms AlexNet in training accuracy over our own dataset.

A 5-fold cross-validation step is applied as a statistical method to estimate the skill of the deep learning model.

We compare our model abilities as described in Table 2. As mentioned above we adopt the 5-fold cross validation as a means to assess the discrimantion skills of a machine learning method. Following the standard steps, we first shuffle the whole dataset randomly. Then we split the images into 5 groups. For each group we take the group itself as a hold-out group or test-set.

The remaining groups represent the training set, we fit a model on the current training set and retain the performance score of the model.

To assess the discrimination capabilities of the model over our data we repeat the steps above using all five groups as hold-out and the remaining four as training set.

More specifically, we have a total number of 4916 patches from SuReMaPP. To apply a 5-fold cross validation we split up the whole dataset into 5 subsets counting 983 patches, the remaining number of patches are used as training set over the category suspicious and non-suspicious regions. The process aims to detect the capabilities of the model to infer knowledge over the classification task. We repeat the steps on each of the 5 subsets. The results of the 5-fold cross validation are in Table 2. For each fold test, the total number of patches of the training set is 3933. In Table 2, we test the training set over the hold-out set of patches. The average correct detections for suspicious and nonsuspicious patches are, respectively, 94.5% and 94.64%, whereas the false detection for the two classes are 5.5% and 5.36%, respectively. We want to stress out that the best parameter configuration achieved during the fine-tuning step is set to be used over the test images to assess the performance of the model over new pictures. No further parameter is changed during the experimental sessions over the test-set.

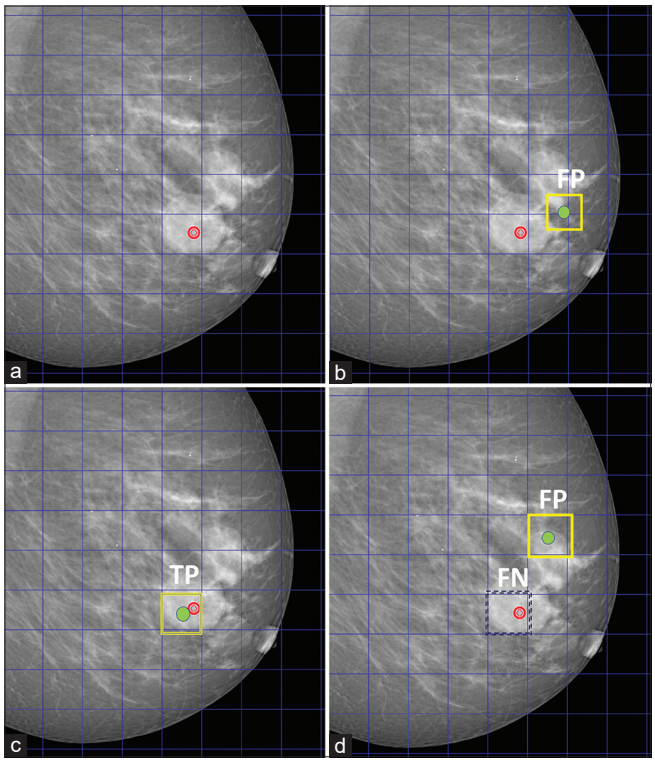

Figure 11: Blue grids above are with the same size (224 × 224) of the input layer of PyramidNet. The red circle spots the suspicious region detected by the radiologists (a). A patch (yellow square) centered on a keypoint (green dot) that does not intercept any suspicious region (b) is a false positive (FP). A patch (yellow square) centered on a keypoint that intercepts the suspicious region (c) is a true positive (TP). In the last case (d), we count the blue dotted patch containing the suspicious region as a false negative (the system was able to detect only a false positive, see the yellow square)

Integrated solution results

In our method, a patch is considered suspicious only when is detected by the SIFT-based method and then by the transfer learning technique. As it can be observed [Figure 10] in the overall scheme, a mammogram is given as input to two histogram specifications, then the SIFT-based method is applied to detect the suspicious patches from both the versions of the mammogram; the SIFT-based method returns pairs of keypoints that fall within suspicious regions to be validated by the deep learning module. Therefore, our deep learning module is composed by two PyramidNet stacks, in order, fine-tuned with the transfer learning paradigm on mammograms processed with the two histogram specifications described as in the SIFT-based technique section.

In this section, we want to give measurements about the performance of our integrated solution by using statistical metrics such as accuracy and sensitivity as defined in Eqs. 5–7. We counted the number of false positives, false negatives, and true positives to give an objective measure of the performance of the method we propose.

For the sake of clarity, we give a graphical description in Figure 11 of what we consider as a true positive, a false positive, and a false negative.

All the experiments have been conducted on both mini-MIAS and SuReMaPP. As said in the previous sections, we need to work with a patch size of 224 × 224 pixels from the mammograms. Hence, we extract a number of 3206 patches from Mini-MIAS dataset and 4914 patches from our new dataset that fit the size requirements to work with PyramidNet (224 × 224 pixels).

As described in Figure 10, the output of each CNN is a probability value given by the Softmax Activation function (it forces the output of the network to represent the probability the input falls into each of the classes). In order, we conduct several experiments by using different bottom and upper threshold values (for each of the CNNs) applied to the Softmax function. For each experiment, a given patch is considered suspicious only when the output of the Softmax function is larger than the given upper threshold value, and conversely, a patch is not considered suspicious when the Softmax Activation function shows values lower than the given bottom threshold. During our experimental sessions, we set 30% as bottom threshold value and 70%, 95%, and 99% as upper probability threshold values. The best performances of our method are registered with 30% and 95% as lower and upper threshold values, respectively. It is necessary to highlight that our deep learning module consists of two PyramidNet CNNs fine-tuned on mammograms (related to the histogram specifications previously mentioned); a patch is voted as suspicious by the deep learning module only when both Softmax outputs are larger than the given upper threshold.

Metrics such as sensitivity and accuracy (Eqs. 5–7) are given to evaluate the effectiveness of our integrated solution.

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (5)$$

$$\text{Specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}} \quad (6)$$

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{True positives} + \text{True negatives} + \text{False positives} + \text{False negatives}} \quad (7)$$

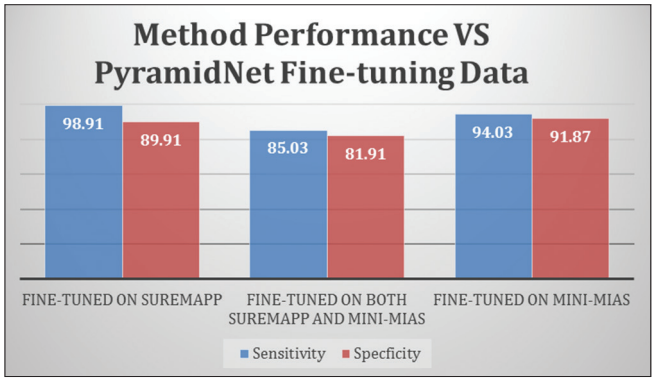

Figure 12: A comparison between different fine-tuning data combination is given with respect to sensitivity and specificity metrics

The integration of the results obtained by our novel proposed solution dramatically reduces the number of false positives with respect to the results obtained only with the SIFT-based method rising from 85% up to 90% of specificity.

A further investigation is necessary to evaluate and assess the level of abstraction from data achieved by transfer learning. For this purpose, we conduct several experiments with respect to different combinations in fine-tuning and test-set images. As briefly mentioned in the previous section, all images in a dataset, both Mini-MIAS and SuReMaPP, have been organized as it follows: 75% of the dataset is used as fine-tuning set, and the rest 25% is used as test set. A list of all experimental case studies is given below:

- Both fine-tuning and test are conducted by using only images belonging to SuReMaPP dataset [Figures 8, 9 and 12]
- Both fine-tuning and test are conducted by using SuReMaPP and Mini-MIAS datasets [Figure 12]
- Both the fine-tuning and test are conducted by using Mini-MIAS dataset [Figures 6, 9 and 12];

Considering the list above, we want to analyze the level of abstraction of transfer learning with different solutions. For our purposes, we want to point out that the images coming from SuReMaPP have quite different sizes (see dataset and data augmentation sections). Furthermore, other than SuReMaPP, the Mini-MIAS images have undergone resampling before to be digitized; this is a limit case study.

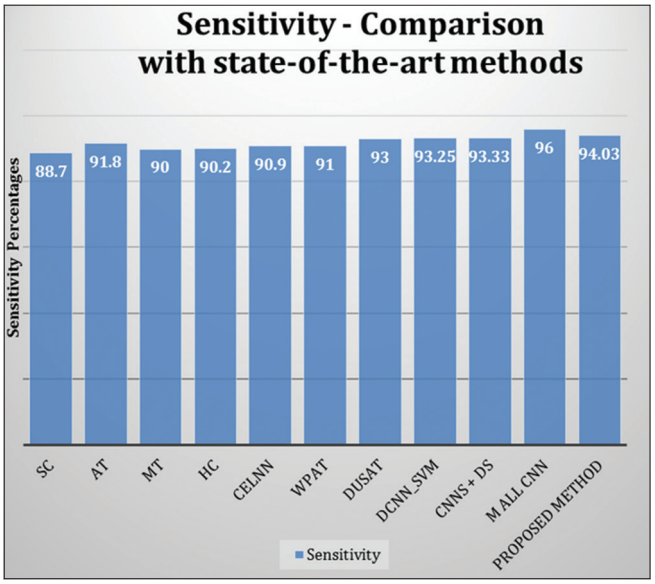

Figure 13: The proposed method is compared with spatial clustering[24] (SC), adaptive thresholding[25] (AT), multilevel thresholding[26] (MT), Havrda and Charvat entropy and Otsu N thresholding[73] (HC), cellular neural network[27] (CelNN), wavelet processing and adaptive thresholding[28] (WPAT), dual-stage adaptive thresholding[29] (DuSAT), deep convolutional neural network with support vector machine[40] (DCNN_SVM), convolutional neural networks and a decision scheme[45] (convolutional neural networks + DS), multiscale all convolutional neural Network[50] (M All convolutional neural network) and our proposed method

We notice [Figure 12] that our novel solution achieves high sensitivity and specificity values; it means that transfer learning allows to predict high-level information through different layers in the PyramidNet stack.

As proof of the effectiveness of our novel solution, we compare our method against several state-of-the-art methods using Mini-MIAS as public gold standard [Figure 13].

We assess the performance of our method on Mini-MIAS with respect to some state-of-the-art methods, which are mainly based on spatial clustering,[24] adaptive thresholding (AT),[25] multilevel thresholding,[26] Havrda and Charvat entropy and Otsu N thresholding,[72], CelNN,[27] wavelet processing and adaptive thresholding[28], DuSAT,[29] DCNN with SVM,[40] CNNs + DS,[45] and multiscale All CNN[50] (M All CNN). Because only sensitivity and false-positive number per image are available from other methods, we compare our method by looking at sensitivity values as shown in plot in Figure 12. Our method achieves a higher sensitivity than most of the comparison methods while keeping a low number of false positives per image. Our method has been implemented and integrated with Caffe's[74] deep learning framework developed by Berkeley AI Research and by community contributors. Caffe's deep neural networks which we worked through are implemented with C++ language. We also used some MATLAB and Python functions, which we interfaced with Caffe framework. All experiments are performed on a computer with a Core i7 950 3.06 GHz processor, 24 GB of RAM, and four GTX 580 graphics cards.

Discussion

The integrated solution we propose for detecting suspicious regions on mammograms reaches high rates of sensitivity and specificity on two very different datasets. Despite the histogram specifications and the preprocessing applied in SIFT-based module allow for reaching up to 85% of specificity, this rate cannot be compared with the state-of-the-art methods. We use the 85% specificity rate of the first module as reference for two main reasons: (1) we want to improve the performance rate when combined with PyramidNet and (2) we want to investigate the inference abilities of transfer learning when adopted over two different types of images and CNN architectures. It is noticed from the experimental sessions that PyramidNet outperforms AlexNet in all the experimental comparisons on the mammogram images we process. This may be explained because of the way PyramidNet increases the feature maps gradually instead of increasing them sharply at unit with downsampling as in AlexNet. Furthermore, it is necessary to mention that we treat high-resolution images coming out from two mammogram devices as in SuReMaPP and medium-resolution images generated in Mini-MIAS. It is observed [Figures 12 and 13] that the performance of the proposed solution decreases when PyramidNet is fine-tuned on both Mini-MIAS and SuReMaPP images, while it keeps ranking high when PyramidNet is fine-tuned only on each dataset. A further discussion about the overall performance with respect to the images used, however, should be conducted and some interesting aspects need to be highlighted: despite the images of SuReMaPP dataset have been generated by two different mammographs (as described in the Dataset section), the integrated solution achieves very high percentages of sensitivity and specificity as shown in Figure 12; the integrated solution also achieves good results also in case of fine-tuning and testing on Mini-MIAS dataset [Figure 12]. As shown by the experimental results, the integration of the SIFT-based method and transfer learning comes out to be a good and valid tool in the CAD perspective [Figure 14].

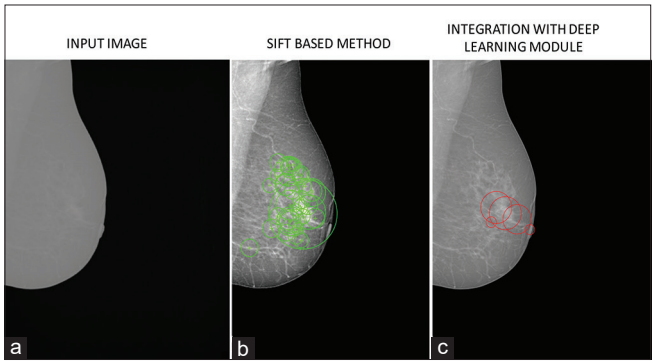

Figure 14: The most important steps of our solution are resumed with green keypoints (b) (scale invariant feature transform-based module) and red keypoints (c) (validated by transfer learning module) overlaid on a mammogram. Only the smaller red circles in the integrated results (c) turn out to be true positives

Conclusions

Our findings suggest that transfer learning paradigm has very powerful ability to infer knowledge from biomedical data reaching good performances in suspicious region detection on mammogram.

Comparisons between the experimental results also show that transfer learning performances slightly drop when the fine-tuning dataset consists of images with quite different size and spatial resolution (it contains both low/middle and high spatial resolution).

The results of our novel solution show how much deep learning helps increasing the performance of the SIFT-based method in such a good combination [Figure 14].

Our method outperforms many state-of-the-art techniques for suspicious region detection in mammograms [Figure 13] based on different approaches such as machine learning, clustering, classification, wavelet, and AT. Even when compared with a deep learning method[40] based on CNN and SVM, our method gets high sensitivity rates.

However, all the considerations above suggest further investigations on the ability of transfer learning and its relations to the acquiring device and spatial resolution. In this respect, the experiments we have conducted so far tell us that transfer learning can be used as a good validation tool to reduce the number of false positives of other

methods on different kinds of mammograms (acquired with different devices), but it is noticeable that if we want to achieve high sensitivity rates, it is recommended to fine-tune the CNN on a single dataset provided with a single acquiring device or with similar acquiring devices.

A common problem in biomedical imaging community is the lack of public dataset with huge amount of biomedical data to be used for scientific purpose. In this perspective, an experimental setup with tens of thousands of labeled images might be used as dataset for the training from scratch of several deep learning architectures to be compared. In that case, some interesting answers could be expected about the knowledge inference upper limit of deep learning techniques. Furthermore, comparing two different deep learning approaches such as CNNs with transfer learning paradigm and CNNs trained from scratch would make possible to find a good trade-off between computational resources and detection accuracy.

In future works, we will be investigating the effectiveness of semantic segmentation with fully connected neural networks and comparing their performances against an integrated solution like the one we propose in this article.

Financial support and sponsorship

None.

Conflicts of interest

There are no conflicts of interest.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Muramatsu C, Hara T, Endo T, Fujita H. Breast mass classification on mammograms using radial local ternary patterns. Comput Biol Med 2016;72:43-53.
3. Watanabe AT, Lim V, Vu HX, Chim R, Weise E, Liu J, *et al*. Improved cancer detection using artificial intelligence: A retrospective evaluation of missed cancers on mammography. J Digit Imaging 2019;32:625-37.
4. -Secretan B, Scoccianti C, Loomis D, Benbrahim-Tallaa L, Bouvard V, Bianchini F, *et al*. Breast-cancer screening–viewpoint of the IARC Working Group. N Engl J Med 2015;372:2353-8.
5. Dheeba J, Albert SN, Tamil SS. Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. J Biomed Inform 2014;49:45-52.
6. Surendiran B, Vadivel A. Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer. Int J Med Eng Inform 2012;4:36-54.
7. Kom G, Tiedeu A, Kom M. Automated detection of masses in mammograms by local adaptive thresholding. Comput Biol Med 2007;37:37-48.
8. Nishikawa RM, Giger ML, Doi K, Vyborny CJ, Schmidt RA. Computer-aided detection of clustered microcalcifications on digital mammograms. Med Biol Eng Comput 1995;33:174-8.
9. Hela B, Hela M, Kamel H, Sana B, Najla M. Breast Cancer Detection: A Review on Mammograms Analysis Techniques. 10th International Multi-Conference on Systems, Signals and Devices (SSD); 2013. p. 1-6.
10. Ganesan K, Acharya UR, Chua CK, Min LC, Abraham KT, Ng KH. Computer-aided breast cancer detection using mammograms: A review. IEEE Rev Biomed Eng 2013;6:77-98.
11. Li Y, Chen H, Cao L, Ma J. A survey of computer-aided detection of breast cancer with mammography. J Health Med Inf 2016;7:4.
12. Mustra M, Grgic M, Rangayyan RM. Review of recent advances in segmentation of the breast boundary and the pectoral muscle in mammograms. Med Biol Eng Comput 2016;54:1003-24.
13. Ardizzone E, Bruno A, Mazzola G. Scale detection via keypoint density maps in regular or near-regular textures. Pattern Recognit Lett 2013;34:2071-8.
14. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vision 2004;60:91-110.
15. Bay H, Ess A, Tuytelaars T, Luc VG. Speeded-up robust features (SURF). Comput Vis Image Underst 2008;110:346-59.
16. Ali Y, Hamed S. Early breast cancer detection using mammogram images: A review of image processing techniques. Biosci Biotech Res Asia 2015;12:225-34.
17. Min D, Xiangyu L, Yide M, Yanan G, Yurun M, Wang K. An efficient approach for automated mass segmentation and classification in mammograms. J Digital Imaging 2015;28:613-25.
18. Kook KJ, Mi PJ, Sik SK, Wook PH. Detection of clustered microcalcifications on mammograms using surrounding region dependence method and artificial neural network. J VLSI Signal Proce 1998;18:251-62.
19. Rangaraj MR, Fabio JA. Gabor filters and phase portraits for the detection of architectural distortion in mammograms. Med Biol Eng Comput 2006;44:883-94.
20. Anitha J, Dinesh PJ. Mammogram segmentation using maximal cell strength updation in cellular automata. Med Biol Eng Comput 2015;53:737-49.
21. Tingting M, Asoke KN, Rangaraj MR. Classification of breast masses via nonlinear transformation of features based on a kernel matrix. Med Biol Eng Comput 2007;45:769-80.
22. Vipul S, Sukhwinder S. CFS--SMO based classification of breast density using multiple texture models. Med Biol Eng Comput 2014;52:521-9.
23. Qiu G, Jianhua Z, Shengyong C, Andrew TP. Automatic segmentation of micro-calcification based on sift in mammograms. Int Conf Biomed Eng Inform 2008;2:13-7.
24. Aize C, Qing S, Xulei Y. Robust information clustering incorporating spatial information for breast mass detection in digitized mammograms. Comput Vis Image Underst 2008;109:86-96.
25. Kai H, Xieping G, Fei L. Detection of suspicious lesions by adaptive thresholding based on multiresolution analysis in mammograms. IEEE Trans Instrum Meas 2011;60:462-72.
26. Pereira DC, Ramos RP, do Nascimento MZ. Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. Comput Methods Programs Biomed 2014;114:88-101.
27. de Sampaio WB, Diniz EM, Silva AC, de Paiva AC, Gattass M. Detection of masses in mammogram images using CNN, geostatistic functions and SVM. Comput Biol Med 2011;41:653-64.
28. Vikhe PS, Thool VR. Mass detection in mammographic images using wavelet processing and adaptive threshold technique. J Med Syst 2016;40:82.
29. Anitha J, Dinesh PJ, Pandian S, Alex I. A dual stage adaptive thresholding (DuSAT) for automatic mass detection in mammograms. Comput Methods Programs Biomed

2017;138:93-104.
30. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, *et al*. A survey on deep learning in medical image analysis. Med Image Analysis 2017;42:60-88.
31. Berkman S, Heang-Ping C, Petrick N, Datong W, Helvie MA, Adler DD, *et al*. Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging 1996;15:598-610.
32. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
33. Pengcheng X, Chang S, Rafik G. Abnormality detection in mammography using deep convolutional neural networks. IEEE Int Symp Med Meas Appl 2018:1-6.
34. Daniel L, Arzav J. Breast Mass Classification from Mammograms using Deep Convolutional Neural Networks. arXiv preprint arXiv: 1612.00542; 2016.
35. Tsochatzidis L, Zagoris K, Arikidis N, Karahaliou A, Costaridou L, Pratikakis I. Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach. Pattern Recognit 2017;71:106-17.
36. Jung H, Kim B, Lee I, Yoo M, Lee J, Ham S, *et al*. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. PLoS One 2018;13:e0203355.
37. Cai H, Huang Q, Rong W, Song Y, Li J, Wang J, *et al.* Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. Comput Math Methods Med 2019;2019:2717454.
38. Richa A, Oliver D, Xavier L, Hoon YM, Robert M. Automatic mass detection in mammograms using deep convolutional neural networks. J Med Imaging 2019;6:31409.
39. Thijs K, Geert L, Bram VG, Albert GM, Clara IS, Ritse M, *et al*. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Analy 2017;35:303-12.
40. Arfan JM. Deep learning based computer aided diagnosis system for breast mammograms. Int J Adv Comput Sci Appl 2017;8:286-90.
41. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep 2016;6:27327.
42. Michiel K, Kersten P, Mads N, Andrew YN, Pengfei D, Christian I, *et al*. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE Trans Med Imaging 2016;35:1322-31.
43. Yamashita R, Nishio M, Do RK, Togashi K. Convolutional Neural Networks: An Overview and Application in Radiology. Insights into Imaging; 2018. p. 1-19.
44. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep 2018;8:4165.
45. Tavakoli N, Karimi M, Norouzi A, Karimi N, Samavi S, Soroushmehr SM Reza. Detection of abnormalities in mammograms using deep features. J Ambient Intell Humaniz Comput 2019. doi: 10.1007/s12652-019-01639-x.
46. Suckling J, Parker J, Dance D, Astley S, Hutt I, Boggis C, *et al*. The mammographic image analysis society digital mammogram database, exerpta medica. Int Congress Series 1994;1069:375-8.
47. Neeraj D, Gustavo C, Andrew BP, Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests. 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA); 2015. p. 1-8.
48. Bowyer K, Kopans D, Kegelmeyer WP, Moore R, Sallam M, Chang K, *et al*. The digital database for screening mammography, Third international workshop on digital mammography. 1996:58. p. 27.
49. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: Toward a full-field digital mammographic database. Acad Radiol 2012;19:236-48.
50. Akila AS, Anitha J, Pandian SI, Peter JD. Classification of mammogram images using multiscale all convolutional neural network (MA-CNN). J Med Syst 2020;44:30.
51. Dhungel N, Carneiro G, Bradley AP. The automated learning of deep features for breast mass classification from mammograms. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2016. p. 106-14.
52. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. Representation learning for mammography mass lesion classification with convolutional neural networks. Comput Methods Programs Biomed 2016;127:248-57.
53. BDCR. Available from: https://bcdr.eu/information/about. [Last accessed on 2020 May 25].
54. Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J Digital Imaging 2017;30:499-505.
55. Benjamin HQ, Hui L, Maryellen ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J Med Imaging 2016;3:1-5.
56. Levy D, Jain A. Breast Mass Classification from Mammograms using Deep Convolutional Neural Networks, CoRR, vol. abs/1612.00542; 2016. Available from: http://arxiv.org/abs/1612.00542, arXiv. [Last accessed on 2016 Dec 02].
57. Agarwal R, Diaz O, Llad'o X, Yap MH, Marti R. Automatic mass detection in mammograms using deep convolutional neural networks. J Med Imaging 2019;6:31409.
58. SuReMaPP Dataset: Suspicious Regions on Mammograms Dataset from PP; 2019. Available from: https://mega.nz/#F!Ly5g0agB!-QL9uBEvoP8rNig8JBuYfw. [Last accessed on 2019 Jun 18].
59. Wang J, Perez L. The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Networks Vis Recognit 2017;1-8. Available from: https://arxiv.org/pdf/1712.04621.pdf. [Last accessed on 2017 Dec 13].
60. Rahmat B, Joelianto E, Purnama I, Purnomo MH. An improved mean shift using adaptive fuzzy Gaussian kernel for Indonesia vehicle license plate tracking. IAENG Int J Comput Sci 2018;45:458-471.
61. Vedaldi A, Fulkerson B. VLFeat: An open and portable library of computer vision algorithms. Proceedings of the 18th ACM International Conference on Multimedia; 2010. p. 1469-72.
62. VLFeat Library SIFT Tutorial. Available from: http://www.vlfeat.org/overview/sift.html. [Last accessed on 2018 Jan 08].
63. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, *et al*. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35:1285-98.
64. Krizhevsky A, Sutskever I, Hinton GE. Imagenet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems; 2012. p. 1097-105.
65. Dongyoon H, Jiwhan K, Junmo K, Deep Pyramidal Residual Networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 5927-35.
66. Deng J, Dong W, Socher R, Li-Jia L, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition CVPR; 2009. p. 248-55.

67. Zhang T. Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms. Proceedings of the Twenty-first International Conference on Machine Learning ICML; 2004. p. 116.
68. Ioffee S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, arXiv preprint arXiv: 1502.03167; 2015.
69. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding Deep Learning Requires Rethinking Generalization. arXiv preprint arXiv: 1611.03530; 2016.
70. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE (1998) Vol 86 pag. 2278-324.
71. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929-58.
72. Burçin K, Nabiyev VV, Turhan K. A novel automatic suspicious mass regions identification uasing Havrda & Charvat entropy and Otsu's N thresholding. Comput Methods Programs Biomed 2014;114:349-60.
73. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. In: ICCV; 2015.
74. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, *et al*. Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint arXiv: 1408.5093; 2014.

BIOGRAPHIES

Alessandro Bruno received his master degree in computer science and his PhD in computer vision at Palermo University respectively in 2008 and 2012. During his PhD scholarship at Palermo University (IT) he focused on texture and local keypoint analysis. After his PhD, he worked from 2012 to 2017 at CVIP (Computer Vision and Image Processing Lab) headed by Prof Ardizzone from Palermo University. Since 2012 up to 2019 he taught basics of computer science at the school of Medicine at Palermo University as a contract lecturer. In 2018 he won a fellowship at INAF (Italian National Institute for Astrophysics) and in 2019 he won a position as a postdoctoral research fellow. In 2019 He was a research visitor of the Imaging Group headed by Prof Jan-Peter Muller at MSSL from UCL (University College London). As a postdoctoral research fellow at INAF he he worked on two main topics, detecting cloud masks from remote sensing imagery and gamma-hadron separation in cosmic ray analysis using deep learning architectures. He is now working as a Research Associate at NCCA (National Centre for Computer Animation) at Bournemouth University, United Kingdom. His current main research topics are Visual Saliency, Human Computer Interaction, Biomedical Imaging, Visual Attention using both unsupervised and supervised approaches. This author is member of CVPL (already GIRPR), the Italian association for research in pattern recognition, computer vision and machine learning.

Email: abruno@bournemouth.ac.uk

Edoardo Ardizzone is a Full Professor of Computer Systems at the Department of Engineering (DI) of the University of Palermo, Italy. Currently, he teaches "Image Processing" at the graduate course of Computer Engineering of the University of Palermo. He is author or co-author of more than 180 scientific papers. Edoardo Ardizzone has been responsible of research units in Palermo involved in many research projects in his interest domains. His current research interests include image processing and analysis, medical imaging, image restoration and content-based image and video retrieval. He is a member of CVPL (already GIRPR and IAPR-IC), the association of Italian researchers in the area of pattern recognition and image analysis.

Email: edoardo.ardizzone@unipa.it

Salvatore Vitabile received the Laurea degree in electronic engineering and the Doctoral degree in computer science from the University of Palermo, Palermo, Italy, in 1994 and 1999, respectively. He is currently an Associate Professor with the Department of Biopathology and Medical Biotechnologies, University of Palermo. In 2007, he was a Visiting Professor with the Department of Radiology, Ohio State University, Columbus, OH, USA. He has co-authored over 200 scientific papers in referred journals and conferences. His current research interests include specialized architecture design and prototyping, biometric authentication systems, driver assistance systems, and medical data processing and analysis. Dr. Vitabile has chaired, organized, and served as a member of the organizing committee of several international conferences and workshops. He is currently a member of the Board of Directors of the Italian Society of Neural Networks.

Email: salvatore.vitabile@unipa.it

Massimo Midiri is a Full Professor of Radiology and the Head Director of the Section of Radiological Sciences, Department of Biopathology and Medical Biotechnologies, University of Palermo, Italy. He authored or co-authored more than 600 publications (more than 250 are also indexed in PubMed). His current research interests include magnetic resonance imaging, computed tomography, and enhanced ultrasound imaging.

Email: massimo.midiri@unipa.it