A deep learning classifier for digital breast tomosynthesis

R. Ricciardi^{a,b,1}, G. Mettivier^{a,b,*,1,2}, M. Staffa^{a,3}, A. Sarno^{b,4}, G. Acampora^a, S. Minelli^c,
A. Santoro^c, E. Antignani^c, A. Orientale^d, I.A.M. Pilotti^d, V. Santangelo^d, P. D'Andria^d,
P. Russo^{a,b,5}

^a Università di Napoli Federico II, Dipartimento di Fisica “Ettore Pancini”, I-80126 Napoli, Italy

^b INFN Sezione di Napoli, I-80126 Napoli, Italy

^c A.O.R.N. “Antonio Cardarelli”, I-80131 Napoli, Italy

^d A.O.U. “San Giovanni di Dio Ruggi D'Aragona”, I-84131 Salerno, Italy

ARTICLE INFO

Keywords:

Digital Breast Tomosynthesis
Breast Tumor
Machine Learning
Convolution neural network
Computed Aided Diagnosis
Deep Learning

ABSTRACT

Purpose: To develop a computerized detection system for the automatic classification of the presence/absence of mass lesions in digital breast tomosynthesis (DBT) annotated exams, based on a deep convolutional neural network (DCNN).

Materials and Methods: Three DCNN architectures working at image-level (DBT slice) were compared: two state-of-the-art pre-trained DCNN architectures (AlexNet and VGG19) customized through transfer learning, and one developed from scratch (DBT-DCNN). To evaluate these DCNN-based architectures we analysed their classification performance on two different datasets provided by two hospital radiology departments. DBT slice images were processed following normalization, background correction and data augmentation procedures. The accuracy, sensitivity, and area-under-the-curve (AUC) values were evaluated on both datasets, using receiver operating characteristic curves. A Grad-CAM technique was also implemented providing an indication of the lesion position in the DBT slice.

Results: Accuracy, sensitivity and AUC for the investigated DCNN are in-line with the best performance reported in the field. The DBT-DCNN network developed in this work showed an accuracy and a sensitivity of $(90\% \pm 4\%)$ and $(96\% \pm 3\%)$, respectively, with an AUC as good as 0.89 ± 0.04 . A k -fold cross validation test (with $k = 4$) showed an accuracy of $94.0\% \pm 0.2\%$, and a F1-score test provided a value as good as 0.93 ± 0.03 . Grad-CAM maps show high activation in correspondence of pixels within the tumour regions.

Conclusions: We developed a deep learning-based framework (DBT-DCNN) to classify DBT images from clinical exams. We investigated also a possible application of the Grad-CAM technique to identify the lesion position.

1. Introduction

Breast cancer is the most common malignancy in women [1]. Breast screening with digital mammography (DM) is considered the most effective method of detecting early-stage breast cancer and reducing related mortality. However, mammography is not performing ideally as a diagnostic exam, in terms of sensitivity and specificity, particularly for dense breasts [2]. With the aim of improving its diagnostic performance

and hence also helping to reduce the rate of recalls, in the last decades research efforts have been directed towards new pseudo-3D or 3D X-ray breast imaging technologies such as digital breast tomosynthesis (DBT) [3–5] and breast computed tomography [6], respectively. These techniques allow to overcome – in part or totally, respectively – the overlap of normal and pathological tissues in the direction of the incident beam, which can decrease the visibility of malignant abnormalities or simulate the appearance of a lesion [7–9]. DBT acquires many two-dimensional

* Corresponding author at: Università di Napoli Federico II, Dipartimento di Fisica “Ettore Pancini”, I-80126 Napoli, Italy.

E-mail address: mettivier@na.infn.it (G. Mettivier).

¹ These authors contributed equally.

² ORCID: 0000-0001-6606-4304.

³ ORCID: 0000-0001-7656-8370.

⁴ ORCID: 0000-0002-3034-7166.

⁵ ORCID: 0000-0001-9409-0008.

projections from various angular positions of the X-ray tube around the compressed breast, at a comparable level of radiation dose with respect to DM [5]. This permits to reconstruct approximately the radiodensity map in different planes transverse to the direction of the beam (typically with a vertical separation of about 1 mm), thus obtaining a (pseudo) three-dimensional representation of the anatomy of the mammary tissues and a clearer localization of possible lesions (masses and microcalcifications) [10] combined with a synthetic representation of a mammography view. The interpretation of a DBT exam requires the visualization and analysis of tens of image slices in a large dataset for each exam, in the craniocaudal or in the mediolateral oblique view [11–13], hence adding complexity and increased reading time to the radiological clinical workflow with respect to a conventional DM exam.

In this context, robust computer-aided detection (CAD) systems - capable of managing the complexity of DBT lesion search space in the diagnostic interpretation task - may represent a crucial tool, also determining a reduced inter-observer and intra-observer variability in the exam reading process. In line with research in past decades for developing CAD systems for mammography exams, particularly using deep learning (DL) techniques [14–17], we developed a CAD system dedicated to classification of DBT exams, to improve radiologists' overall performance in DBT exam analysis and potentially improve diagnostic accuracy. A specific goal was an acceptable trade-off between the computational costs arising from the automatic analysis and the classification performance, in terms of increased sensitivity and reduced false positive (FP) rate.

Preliminary studies [18–35] for developing CAD systems dedicated to DBT were based on the sequence of these processes: the detection of the mass candidates; their segmentation from background; the exam classification via the extraction of relevant features. Some studies adopted hand-crafted features developed in general image processing applications or conventional mass features [18–27]. Such an approach relies on expert knowledge. System model predictions in the automated lesion detection task may be improved by exploiting recent advances in machine learning, especially related to the use of deep convolutional neural networks (DCNNs), as well as advances in graphical processing units (GPU) based technologies, availability of large datasets of annotated images and novel optimization methods [28–34]. At variance with feature-based methods, DCNN-based methods produce a decision (i.e., classify data) directly from input raw images. They do not require segmentation and hand-crafted feature extraction steps, which are necessary for traditional feature-based classifiers, such as artificial neural networks (ANNs) or support vector machines (SVMs) [23–25]. Indeed, a DCNN can automatically extract some descriptors from an image, thus avoiding developing specific image processing algorithms. On the other hand, the task of learning the complex patterns of masses requires a large set of different training samples, as well as an architecture and a regularization method. They might be less influenced by lesion-specific features than feature-based methods, resulting in a better chance of recognizing a mass in at least one of the DBT views and a significantly better breast imaging-based detection performance.

In this paper, we developed a computerized detection system for the automatic classification of the presence/absence of mass lesions in the dataset of slices of DBT annotated exams, aiming at improving over existing deep learning-based techniques in terms of sensitivity and specificity. We implemented an *ad hoc* DCNN architecture (termed DBT-DCNN) performing a binary classification of individual slices belonging to the same DBT exam dataset. We then compared the DBT-DCNN performance to MATLAB implementations [36,37] of popular architectures (AlexNet and VGG19, respectively) available in this field. The classification performance of the three neural networks was comparatively assessed on different datasets provided by two hospitals on two DBT clinical systems from two manufacturers. This process allowed to evaluate the robustness of the architecture versus the influence of different hardware instrumentations or acquisition protocols. Additionally, we implemented a technique (Grad-CAM) to highlight those pixels in all

DBT slices of a given exam, more relevant for the final classification task performed by the network. This permitted to explore the possibility of providing an indication of the position of the mass inside the slice(s) classified as abnormal, as well as for showing the location of possible network activation in zones less relevant for the diagnostic task.

2. Materials and methods

This section first describes our datasets, then briefly reviews state-of-the-art techniques that we used as benchmark (AlexNet and VGG19 networks) for comparison with our DCNN technique. Finally, it introduces the proposed DBT-DCNN architecture.

2.1. Dataset

The two DBT datasets used in this study have been made available by two Hospitals⁶ (here indicated as H1 and H2) upon approval of their institutional review board. For this research, digital images were used, in disaggregated form, not attributable in any way to any specific patient, managed in complete anonymity and in compliance with current EU legislation on the processing of sensitive data. Both sets of DBT exams were reconstructed to 1 mm slice spacing and an in-plane resolution of $90 \times 90 \mu\text{m}^2$ for reconstructed planes, using iterative and FBP reconstruction techniques for H1 and H2, respectively [38]. The reconstructed slices had a resolution of 16 bit/pixel.

To reduce the white noise, we processed all images with a denoising algorithm developed using ImageJ software. The algorithm evaluates the noise inside a selected region of interest (ROI) of 300×300 pixels in the image and the measured noise value was subtracted from the whole image. Here, the size of the ROI is a fundamental choice, since applying the algorithm on a too large matrix implies a final image that is too blurry to extract information relevant to network learning.

Fig. 1 shows an example of the application of this algorithm: as shown here, a significant noise reduction and a better edge-sharpness can be obtained.

As last step, the dimensions of the slices were reduced from their initial size (1072×2356 pixel for H1 dataset and 1996×2457 pixel for H2 dataset) to 300×300 pixels using a resizing algorithm, which averages adjacent pixels. The necessity to bin the image is due to a hardware limitation on the available GPU architecture. The maximum size of 300×300 pixels was established after trial and error and by limiting the search at the maximum value supported by our hardware (details are provided in sec. 3.1).

2.1.1. Hospital 1 dataset

Anonymized DBT images of 100 patients at H1 site were acquired with a Giotto Class 40,000 system. Each image had a matrix size of 1072×2356 pixel with $100 \times 100 \mu\text{m}^2$ pixel dimension. The Giotto Class 40,000 system uses an amorphous selenium (a-Se) flat panel digital detector. The DBT cases were acquired in craniocaudal (CC) view using a total tomographic angular range of 30° with 2.7° increments and 11 projections. The dataset consisted of 4692 slices with 137 masses confirmed after biopsy on all patients. Of these, 104 masses were malignant and 33 were benign. Details are reported in Table 1. No data augmentation technique was applied to this dataset. As indicated in Fig. 6, this dataset was used to train and validate the proposed DCNN architecture. Seventy percent of these data were used for network training, while the remaining part was used for test (see Table 1).

2.1.2. Hospital 2 dataset

The dataset from H2 site is composed of DBT images of 9 patients

⁶ Azienda Ospedaliera Cardarelli, Napoli, Italy (Hospital 1) and Azienda Ospedaliera Universitaria San Giovanni di Dio Ruggi d'Aragona, Salerno, Italy (Hospital 2)

Fig. 1. Comparison between two tomosynthesis images belonging to the same case a) before and b) after preprocessing for feeding the deep CNN. In the post-processed image (b) we can see a reduction of quantum noise and the accentuation of the edges of the structures inside the breast.

Table 1

Datasets used for DCNN training, testing and validation. Positivity (or negativity) refers to the presence (or absence) of a mass lesion in the DBT exam of each patient, as annotated by the referring radiologist (ground truth).

	No. patients	Total no. DBT slices	Positive	Negative
Hospital 1				
H1 Training dataset	70	3286	2217	1069
H1 Validation dataset	30	1406	949	457
Total	100	4692	3166	1526
Hospital 2				
H2 Augmented dataset	9	3024	1897	1127
H2 Training dataset	6	169	106	63
H2 Validation dataset	3	73	46	27
Total	9	242	152	90

imaged with a Hologic Selenia Dimension AWS8000 DBT system. The scan comprised 30 2D projections on an arc of 30°. Five cases with pathological diagnosis were confirmed by biopsy. From these data, 242 slices (1996×2457 pixel) in craniocaudal view were obtained by an automated procedure developed *ad hoc* using a macro procedure in ImageJ software. These images were processed following a background normalization and correction procedure similar to the one used to the first dataset. Due to limited number of images of this dataset we applied a data augmentation procedure. The images underwent two reflections, the first one with respect to the vertical axis of the slice and the second one with respect to the central horizontal axis, and then a rotation within an interval $[-35^\circ, +35^\circ]$ with a 10° step. The entire process took place in “On stream” mode, with images automatically generated with each workout and then discarded. From the 242 DBT slices in the original dataset, 3024 slices were obtained (see Table 1).

2.2. Benchmark DCNN architectures

2.2.1. AlexNet

AlexNet is a CNN architecture developed by Alex Krizhevsky, the winner of one of ImageNet’s biggest object recognition challenges.

AlexNet is made up of 25 levels (Fig. 2): 1 input level, 5 convolutional levels, 3 completely connected classification levels and finally 1 output level. In the first, second and last levels, the convolutional levels are followed by rectified linear units (ReLU) and local response normalization (LRN). To reduce the number of superfluous parameters for learning and thus preventing overfitting, the feature map was downsampled through max pooling. To strengthen the learning process, the first and second fully connected levels were followed by two dropout levels.

Using the transfer learning (TL) methodology, we trained the AlexNet network on our datasets, by applying a pre-processing algorithm to adapt our images to the AlexNet format. AlexNet, indeed, is trained to recognize 227×227 pixels RGB images. The images in our database being in grayscale format, all these images were transformed into false color images by means of a pre-processing algorithm that is part of the Matlab’s Data Augmentation functions for color processing. From now on we refer to this architecture as TL-AlexNet.

2.2.2. VGG19 net architecture

VGG neural networks are the natural evolution of the AlexNet network. Their architecture is based on AlexNet with the difference that they are deeper and more closely linked to the principle of information locality. The idea behind these neural networks is that within the whole image, the neighbouring pixels are more informative to analyze and interpret the image content, than pixels at the two opposite ends of the image. Within these networks, each level of convolution is always followed by a ReLU level, so increasing the non-linearity of the activation function and its discrimination power. Another important feature of these networks is that they adopt convolution levels with smaller receptive fields to recognize a greater number of parameters for each image, for an increased performance.

In this work, we used a VGG19 network composed of 19 trainable levels, divided as shown in Fig. 3. For the first and second levels two convolutional layers can be defined followed by a ReLU level and a max pooling level. The next levels were respectively composed of four convolutional layers, a ReLU level and a max pooling level. Finally, there were three Fully connected levels, interspersed with a ReLU and a dropout level, and a softmax level followed by the output. As in the case

Fig. 2. TL-AlexNet architecture is composed of 24 levels: 1 level input, 5 convolutional levels, 3 fully connected classification levels and finally 2 dropout levels immediately followed by an output level. The convolution kernel size (3×3 , 5×5 or 11×11) is indicated at each step.

Fig. 3. DCNN VGG19 architecture composed of 19 trainable convolution levels.

of AlexNet we adopted a transfer learning technique to train the VGG19 network (termed in the following as TL-VGG19).

2.3. DCNN architecture

DGNNs are a type of artificial neural methods composed of convolutional layers and fully connected layers within a deep architecture. During training, a DCNN learns patterns through the kernels in each

convolutional layer. The feature maps in one layer are generated by convolving the kernels with the feature maps in the previous layers and combining them with weights. Each feature maps is a connector to an activation function. In this work, we developed a DBT-DCNN and compared its performance with those of the two benchmark DCNN architectures described above.

Fig. 4. Deep Convolution Neural Network architecture developed for this work. It is made up of 24 convolution levels: 1 level input, 5 convolutional levels, 2 fully connected classification levels and finally 1 softmax level immediately followed by an output level.

2.3.1. DBT-DCNN architecture

The proposed DBT-DCNN architecture was developed from scratch. The images were accessed in raw format (dicom for processing) to avoid dependence on manufacture's processing methods; images from the same case were assigned to the same subset to keep the two subsets independent of each other.

The DCNN network was developed using “Matlab 2020b” development code and the Matlab “Machine and Deep Learning Matlab Toolbox”. The network architecture is shown in Fig. 4. It consisted of 24 levels: 1 level input, 5 convolutional levels, 2 fully connected classification levels and finally 1 softmax level immediately followed by an output level. All the convolutional layers used rectified linear units as an activation function given by $f(x) = \max(0, x)$: to archive rotational and translation invariance to the input patterns, the feature map was sub-sampled through max-pooling.

We analyzed the operation and the characteristics of each layer starting from the input one. In this level, the network loads all pre-processed images present in the database with the relative labels (“sick” or “healthy”), which act as the basic truth at the time of validation. Then, the network starts a sequential image analysis, so initiating the network training. Initially, each pixel of the image is associated with a neuron whose bias values and relative weights take a random value. This layer is then followed by the 5 convolution layers with 96, 128, 384, 192 and 128 filters with convolution kernels of size 11×11 and 5×5 pixels for the first two layers, and 3×3 pixels for the last three layers. The convolution step for each level is equal to 1. All the convolution kernels have been summed, so preserving the input and output dimensions of the image. Following each convolutional level, for each filter applied to the image the ReLu layers (co-aided by normalization levels) generate the maps of the characteristics of the network by connecting neurons together and updating the weights and biases at each iteration. In the second part of the network, the images are input to the two fully connected layers which, after linearization, classify them in the two possible classes. After each classification, the probability of belonging to that class is evaluated using the cross-entropy function within the softmax layer, which provides an output value of probability between 0 and 1 and finally saves it within that layer. The last layer is the output layer for displaying the result of the operations.

The analysis of the network structure (performed to check the correct

functioning of each part) was realized by means of the Matlab function “AnalyzeNetwork”. For training, we used the “ADAM” algorithm (adaptive moment estimation) as a solver [39], with an initial learning rate of 0.0004, a gradient decay factor of 0.9 and a squared gradient decay factor of 0.999. Out of 4692 slices of the H1 dataset (Table 1), 70% (3286) was used to train DCNN and the remaining 30% (1406 slices) for the testing phase. The network has been trained for 15 eras characterized by 29 interactions each, for a total of 435 maximum iterations.

To follow the progress of training and avoid network overfitting, we monitored a “live” validation point during training, at each successive 20 iterations. Fig. 5 shows the training and loss curves for the DBT-DCNN network. Both trends follow the expected behaviour. In particular, the training curve has a globally increasing trend as the number of epochs increases, while the loss curve follows globally a descending trend. The duration of the training task, running in parallel mode on two NVIDIA GPU TitanX cards, was about 130 min. Finally, the network was tested on the remaining 1406 images of the dataset stored for the test and statistical analysis of the quality of network performance, by building the confusion matrix.

2.4. Network performance evaluation

The effectiveness of the DCNN classification was calculated regarding correct/incorrect classification using performance metrics, with usual definitions:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

where TP , FN , TN and FP are the number of true positives, false negatives, true negatives and false positives, respectively, obtained from the automatic Matlab evaluation procedure in which the probability

Fig. 5. Scheme of training curves (a, in blue) and loss (b, in orange) with the corresponding validation curves (dashed in black) implemented for the DBT-DCNN network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

threshold was selected to optimize sensitivity and specificity. Furthermore, we provided the integral of the area subtended by the ROC curve (AUC) to measure the degree of separability between classes determined by the classifier. The higher the AUC value, the better the model performs in distinguishing slices with and without the disease.

2.5. Grad-CAM techniques

Grad-CAM is a generalization of the Class Activation Mapping (CAM) techniques, used to assess graphically the network functioning. Conceived by Selvaraju and co-authors [36], Grad-CAM uses the gradient of the classification score with respect to the final convolutional feature map to show which parts of the image are most important for classification [40]. The final purpose of this type of algorithm is to display the sensitivity/saliency maps of the gradients formed by heat-maps superimposed on the input image. This technique allows to visualize on a color scale the most relevant pixels for the final classification. Here we calculated these maps and compared them with the mass position marked by the radiologist with a 2D bounding box for each planar DBT slice (ground truth), to evaluate the possibility of using this algorithm in our classification task to assist the radiologist in the mass localization process.

3. Results

3.1. Effect of input image size

TL-AlexNet and TL-VGG19 adopted, respectively, an input image size of $227 \times 227 \times 3$ pixel and $224 \times 224 \times 3$ pixel. When loading the image dataset for processing by the DBT-DCNN, each input image of the network (1072×2356 pixel for H1 dataset and 1996×2457 pixel for H2 dataset) was initially binned to $100 \times 100 \times 1$, to $200 \times 200 \times 1$ or to $300 \times 300 \times 1$ pixel image. This last value represents the maximum image size which we were able to test given the computing power available for our hardware. Fig. 6 shows the values of TP, TN, FP, FN, accuracy and sensitivity of the DBT-DCNN network evaluated on H1 validation dataset as a function of the dimension of the input image, showing the improvement of all the values with the increase of the image dimension. A value of $300 \times 300 \times 1$ pixel was then considered adequate for reaching high values of accuracy and sensitivity in our DBT-DCNN network, so that all subsequent processing steps were carried out with $300 \times 300 \times 1$ pixel images input to the DBT-DCNN.

3.2. DCNN architecture evaluation

Following the scheme reported in Fig. 7, we trained the three presented DCNN architectures (DBT-DCNN, TL-AlexNet and TL-VGG19) using the two datasets (H1 and H2) to assess which architecture produced the best performance in terms of the selected evaluation metrics.

The evaluation was repeated five times and the results are reported in Tables 2 and 3. The best values were obtained by DBT-DCNN on the H1 dataset. This architecture showed an accuracy and a sensitivity of $(90\% \pm 4\%; 96\% \pm 3\%)$, respectively (Table 2, third row), compared to $(84\% \pm 1\%; 99\% \pm 1\%)$ (Table 2, first row) and $(74\% \pm 1\%; 88\% \pm 1\%)$ (Table 2, second row) provided by TL-AlexNet and TL-VGG19, respectively. However, these values were determined using grayscale images instead of RGB images (on which the TL-AlexNet has been originally pre-trained). On the other hand, the TL-VGG19 network was unaffected by this bias: though it is very similar to the TL-AlexNet network, it provides lower values than both the TL-AlexNet and the network specifically developed for this work. To validate the results of DBT-DCNN network we performed a k -fold cross validation (with $k = 4$): a F1-score test provided a value as good as 0.93 ± 0.03 (on a 0–1 scale). We note the significant lower number of FP obtained with the DBT-DCNN network ($FP = 108 \pm 28$) with respect to TL-AlexNet ($FP = 206 \pm 8$) and TL-VGG19 ($FP = 235 \pm 11$). For a fixed specificity of 80%, the corresponding sensitivity for the DBT-DCNN, TL-AlexNet and TL-VGG19 were $85\% \pm 3\%$, $67\% \pm 1\%$ and $58\% \pm 2\%$, respectively. A similar trend was observed with the second dataset (Table 3), where we obtained $(89\% \pm 1\%, 81\% \pm 5\%), (81\% \pm 3\% \text{ and } 72\% \pm 3\%), (78\% \pm 3\%, 68 \pm 8\%)$ for the (accuracy, sensitivity) of DBT-DCNN, TL-AlexNet and TL-VGG19, respectively.

3.3. DCNN performance evaluation: AUC

The ROC curves for all the investigated networks are shown in Fig. 8. For each curve is indicated the confidence interval obtained by repeating the measurement 5 times. The graph was derived from test data of the H1 database. Each image was classified by the trained network and the predicted labels was compared with the real label defined as ground truth. The algorithm then calculates the sensitivity as True Positive Rate (TPR) and the specificity from which it then derives the false positive rate (FPR) as $1 - \text{Specificity}$. The corresponding AUC value was reported in Table 1. The higher value was as high as 0.89 for the DBT-DCNN network.

3.4. Gradient map

Fig. 9a shows a DBT slice in which the DBT-DCNN indicated the presence of a mass. A Grad-CAM algorithm was applied to this image (Fig. 9b), highlighting (in red/yellow colors) the regions of input pixels that cause greater activation of the network. The two mass lesions in Fig. 9a were marked by a radiologist (of Hospital 1) with a red circle. As we can observe in Fig. 9b, the network correctly identified both lesions hence classifying the image as pathological, perfectly identifying the *retro*-areolar lesion also following its local branches, typical of this type of carcinoma. Namely, pixels corresponding to the central region of the mass had a greater impact in the classification of the image as

Fig. 6. a) Number of TP, TN, FP, and FN occurrences, and b) accuracy and sensitivity of the DBT-DCNN network as a function of the linear size of the square input image.

Fig. 7. Scheme of the training, validation and test processes implemented for all the implemented DCNN architectures using the two different datasets.

Table 2

Evaluation in terms of classification absolute numbers, accuracy, sensitivity, specificity, precision and AUC, of the DCNN architectures using dataset H1.

H1 dataset (N = 1406 slices)									
	TP (#)	TN (#)	FP (#)	FN (#)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
TL-AlexNet	935 ± 9	252 ± 8	206 ± 8	14 ± 9	84 ± 1	99 ± 1	55 ± 2	82 ± 1	0.81 ± 0.01
TL-VGG19	839 ± 5	220 ± 11	235 ± 11	112 ± 5	74 ± 1	88 ± 1	48 ± 2	78 ± 1	0.74 ± 0.02
DBT-DCNN	913 ± 21	349 ± 28	108 ± 28	37 ± 21	90 ± 4	96 ± 3	76 ± 3	89 ± 3	0.89 ± 0.04

Table 3

Evaluation in terms of classification absolute numbers, accuracy, sensitivity, specificity and precision of the DCNN architectures using dataset H2.

H2 dataset (N = 73 slices)								
	TP (#)	TN (#)	FP (#)	FN (#)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
TL-AlexNet	28 ± 13	31 ± 14	3 ± 3	11 ± 5	81 ± 3	72 ± 3	92 ± 1	91 ± 9
TL-VGG19	25 ± 11	32 ± 14	4 ± 4	12 ± 5	78 ± 3	68 ± 8	88 ± 11	87 ± 12
DBT-DCNN	32 ± 16	33 ± 16	1 ± 2	7 ± 4	89 ± 1	81 ± 5	94 ± 5	96 ± 6

Fig. 8. ROC curve for the DBT-DCNN architecture evaluated on the H1 dataset (in blue), and corresponding ROC curves for TL-AlexNet (in red) and TL-VGG19 networks (in green). For each curve a band is plotted, whose width indicates the confidence interval obtained by repeating the measurement 5 times. The AUC for DBT-DCNN was 0.89 ± 0.04 (Table 2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pathological (in red/yellow) while the branches of the tumor, less contributed to the neural activation (in light blue).

4. Discussion

Over the last five years, a plethora of methods of image analysis based on artificial intelligence (AI) have been massively introduced in the field of breast cancer detection [41], providing an important contribution in the management and interpretation of large quantities of information. Specifically, there is considerable research in the field of CAD for DM and DBT [35]; reportedly, a CAD for DBT could help reducing the exam interpretation time. Indeed, after introduction of DBT in the clinical routine after FDA approval (2011), it has been reported that the time required for reading DBT images may be twice as large as that for a DM exam, due to the substantial increase in the number of images needed to be reviewed [42]. This has a significant impact on the introduction of this method in the clinical workflow, especially in the screening activity. Thus, to help radiologists to read DBT image datasets, considerable research efforts have been devoted to the development of automatic breast cancer detection techniques and different strategies for CAD on DBT data are emerging (Table 4). For example, given the very different characteristics between soft tissue lesions and calcifications and the frequently size-limited training datasets, usually different separate detection algorithms are used for each of these types of lesions.

CAD systems can work at an image-level or a pixel-level classification. Image-level classification involves identifying an entire image (2D DBT slice or 3D DBT image) as containing a cancer or not, while pixel-level classification involves determining the image region where the

Fig. 9. Comparison between the input image (a) and the image to which the Grad-Cam technique has been applied (b) to display the saliency maps of the neural activation gradients belonging to the same patient with infiltrating ductal carcinoma. The points of greatest neural activation located in the central region of the mass are highlighted in light red/yellow, while in light blue are shown the neural activation points corresponding to the branches of the lesion that led CNN to classify the image as pathological. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4

Characteristics and performance of some DBT mass detection CAD systems reported in the literature from 2004 to 2019. The different classifiers are applied to different input: to exam DBT (3D), to a single slice from DBT (slice), to 2D ROI and 3D VOI. For this work the performance values reported refer to the H1 dataset. Values highlighted in bold represent best performance in this comparison.

Ref.	Year	Classifier	Training method	# Patients	Input type	Performance		
						AUC	Sensitivity (%)	Accuracy (%)
Feature-based classifiers								
18	2005	Feature extraction			3D	0.91	85	
22	2006	LDA		36	slice		90	
21	2008	Mutual information		100	ROI		85	
26	2008	Feature extraction		96	slice		88	
19	2008	LDA		100	slice + 3D		80	
20	2010	LDA		99	slice	0.93		
23	2013	ANN		192	slice		80	
25	2014	SVM		101	3D		90	
24	2016	SVM		160	VOI	0.847		
33	2019	SVM		24	ROI	0.798	83.87	72.54
		RF				0.757	80.65	70.59
		Naive Bayes				0.648	64.52	60.78
		Multi-layer perceptron				0.754	77.42	70.59
Deep Learning based Classifiers								
29	2016	DCNN	Transfer learning	94	ROI	0.9	90	
31	2016	DCNN	Feature Extraction	344	ROI		89	
28	2017	DCNN		185	VOI	0.92		
30	2018	DCNN	Single Transfer learning	324	ROI	0.85		
		Feature extraction	Multiple Transfer learning			0.91	83	
32	2018	DCNN Multiple-Instance RF	Training from scratch	87	slice	0.87	86.6	86.81
		DCaRBM MI-RF	Training from scratch			0.7	81.8	78.5
		Hand-crafted Featured MI-RF	Training from scratch			0.75	66.6	69.2
34	2019	ANN	Training from scratch	16	ROI			75
		VGG-19/KNN						93
This work	2020	DCNN	Training from scratch	100	slice	0.91	99.0 ± 0.5	94.0 ± 0.2

LDA = Linear Discriminant Analysis, SVM = Support Vector Machine, RF = Random Forest, ANN = Artificial Neuronal Network.

lesion is located, working on ROIs or VOIs. We decided to implement an image-level study of the identification of the presence or absence of a mass in a single slice. This classification is intended to help the DBT interpretation task by radiologists, at the screening level, when determining if a patient needs a further examination. The development of a classification network for microcalcification detection will be the goal of a future work.

In this work, we compared the performance of three different DCNN networks, one developed *ad-hoc* (DBT-DCNN) and the other two, AlexNet and VGG19, belonging to the class of networks that are commonly implemented for the classification of common scene images. The three

networks were compared on two different datasets from two hospitals with different acquisition devices. The DBT-DCNN shows favorable results for both datasets here tested with respect to benchmark architectures, though the H2 dataset is small. The accuracy and AUC values obtained by our DBT-DCNN classifier on H1 dataset (100 DBT clinical exams, for a total of 4692 DBT slices) were comparable to the ones reported in the literature (see Table 4). The goodness of the test was also confirmed by the Fscore (or F1 score) value, which was 0.93 ± 0.03 .

As shown in Table 4, all the reported classifiers featured high AUC values from 0.7 to 0.93, with an accuracy ranging from 69% to 93%. Correspondingly, the AUC value and accuracy of the DBT-DCNN

network evaluated on the H1 dataset were 0.91 and 94%, respectively, thus indicating a very high performance. Specifically, for the architectures working at image-level, we find that the AUC value obtained by our DBT-DCNN network compares favorably with the value (0.93) reported by Chan et al. [20] related to a feature-based network, at a comparable level of patient cases (99 vs. 100). At the same time, as regards classification accuracy, the value obtained by DBT-DCNN (94%) is the highest reported in the comparison offered by Table 4, and only the network reported in [34] (based on the use of a VGG19 network) reports a comparable value as high as 93%. An increase of the numerosity of the dataset including additional DBT data from a third hospital site (foreseen before year 2022) might contribute to increase the robustness of the above predicted performance of DBT-DCNN.

As an original addition to the classification task evaluation, we also reported the use of the Grad-CAM algorithm on DBT images. For each slice, this technique produces the saliency map of the neural activation gradients, showing values of maximum intensity in the image pixels related to the position of the breast lesion. This approach might provide a possible indication to the DBT observer to help in the lesion diagnosis. As a matter of example, in Fig. 8 we showed the saliency map as generated by the Grad-CAM algorithm, suggesting that this technique could correctly identify the position of the two infiltrating ductal carcinomas.

5. Conclusions

In this work, we showed that the proposed DBT-DCNN architecture outperforms state-of-the-art techniques on two benchmarking clinical DBT image datasets provided by two different Hospitals. Major result arising from this comparison is represented by increased sensitivity and a significant reduction of FPs, that is an important parameter in that the reduction of wrong evaluations can in turn reduce both patients psychological stress and the work stress of the doctors by avoiding the carrying out of further investigations.

This work also demonstrated that the performance of the proposed DBT-DCNN network in terms of AUC and accuracy are comparable with those of other DCNN networks in the literature, albeit limited to a smaller set of data due to the recent introduction of the exam diagnostic in screening programs.

An ongoing extension of this work will aim at evaluating the network performance on a larger dataset of annotated images (a few hundred DBT exams). Additionally, we are also considering applying multiclass features extraction operation of breast-specific diagnostic features such as tumor density and geometry. As an ongoing work, we are developing an algorithm - applied after DBT-DCNN evaluation of each slice belonging to a single DBT exam - which uses the z-position of the slice in the DBT exam for taking into account the 3D spatial information of the exam. This algorithm will evaluate if subsequent slices in the same DBT exam are classified as positive, indicating a section of maximum probability searching for a mammary lesion.

Finally, from our first observation of the Grad-CAM algorithm outcomes, we also conclude that the use of this technique could represent a valuable solution to select regions of interest of the slides to be adopted for training the DCNN, with the aim enhancing the mass localization process. Hence, subsequent extension of this work will consist in the localization task of masses and microcalcifications within the tomosynthesis images supported by Grad-CAM techniques.

Acknowledgements

This work was funded in part by INFN (Istituto Nazionale di Fisica Nucleare, Italy).

References

- [1] Marcom PK. Genomic and Precision Medicine: Primary Care, 3rd Edition; 2017. p. 181–94.
- [2] Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology* 2017;283(1):49–58.
- [3] Sechopoulos I. A review of breast tomosynthesis. Part I. The image acquisition process. *Med Phys* 2013;40(1):014301.
- [4] Sechopoulos I. A review of breast tomosynthesis. Part II. Image reconstruction, processing and analysis, and advanced applications. *Med Phys* 2013;40(1):014302.
- [5] Maldera A, De Marco P, Colombo PE, Origgi D, Torresin A. Digital breast tomosynthesis: dose and image quality assessment. *Phys Med* 2016;33:56–67.
- [6] Sarno A, Mettievier G, Russo P. Dedicated breast computed tomography: basic aspects. *Med Phys* 2015;42(6):2786–804. <https://doi.org/10.1118/1.4919441>.
- [7] Sage J, Fezzani KL, Fitton I, Hadid L, Moussier A, Pierrat N, et al. Experimental evaluation of seven quality control phantoms for digital breast tomosynthesis. *Phys Med* 2019;57:137–44.
- [8] Hadjipanteli A, Elangovan P, Mackenzie A, Wells K, Dance KD, Young KC. The threshold detectable mass diameter for 2D-mammography and digital breast tomosynthesis. *Phys Med* 2019;57:25–32.
- [9] Petrov D, Marshall NW, Young KC, Bosmans H. Systematic approach to a channelized Hotelling model observer implementation for a physical phantom containing mass-like lesions: Application to digital breast tomosynthesis. *Phys Med* 2019;58:8–20.
- [10] Agasthya G, Rodriguez-Ruiz A, Sechopoulos I. Digital Breast Tomosynthesis. In: Russo P, editor. *Handbook of X-ray imaging physics and technology*. CRC Press; 2018.
- [11] Astley S, et al. A comparison of image interpretation times in full field digital mammography and digital breast tomosynthesis. *Proc. SPIE* 2013;8673:1–8.
- [12] Bernardi D, Ciatto S, Pellegrini M, Anesi V, Burlon S, Cauli E, et al. Application of breast tomosynthesis in screening: Incremental effect on mammography acquisition and reading time. *Br J Radiol* 2014;85:e1174–8.
- [13] Wallis MG, Moa E, Zanca F, Leifland K, Danielsson M. Two view and single-view tomosynthesis versus full-field digital mammography: High-resolution x-ray imaging observer study. *Radiology* 2012;262:788–96.
- [14] Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R. Deep learning to improve breast cancer detection on screening mammography. *Sci Rep* 2019;9:12495.
- [15] Yu X, Pang W. Mammographic image classification with deep fusion learning. *Sci Rep* 2010;10:14361.
- [16] Wu N, Phang J, Park J, Shen Y, Huang Z, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging* 2020;39(4):1184–94.
- [17] Lehman CD, Wellman RD, Buist DS, et al. Diagnostic accuracy of digital screening mammography with and without Computer-Aided Detection. *JAMA Intern Med* 2015;175(11):1828–37.
- [18] Chan HP, Wei J, Sahiner B, Rafferty EA, Wu T, Roubidoux MA, et al. Computer-aided detection system for breast masses on digital tomosynthesis mammograms: preliminary experience. *Radiology* 2005;237(3):1075–80.
- [19] Chan HP, Wei J, Zhang Y, Helvie MA, Moore RH, Sahiner B, et al. Computer-aided detection of masses in digital tomosynthesis mammography: comparison of three approaches. *Med Phys* 2008;35:4087–95.
- [20] Chan HP, et al. Characterization of masses in digital breast tomosynthesis: Comparison of machine learning in projection views and reconstructed slices. *Med Phys* 2010;37:3576–86.
- [21] Singh S, Tourassi GD, Baker JA, Samei E, Lo JY. Automated breast mass detection in 3D reconstructed tomosynthesis volumes: a featureless approach. *Med Phys* 2008;35:3626–36.
- [22] Reiser I, Nishikawa RM, Giger ML, Wu T, Rafferty EA, Moore R, et al. Computerized mass detection for digital breast tomosynthesis directly from the projection images. *Med Phys* 2006;33:482–91.
- [23] van Schie G, Wallis MG, Leifland K, Danielsson M, Karssemeijer N. Mass detection in reconstructed digital breast tomosynthesis volumes with a computer-aided detection system trained on 2D mammograms. *Med Phys* 2013;40(4):041902.
- [24] Kim DH, et al. Latent feature representation with 3-D multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis. In: *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP*; 2016. p. 927–31.
- [25] Palma G, et al. Detection of masses and architectural distortions in digital breast tomosynthesis images using fuzzy and a contrario approaches. *Pattern Recogn* 2014;47(7):2467–80.
- [26] Park SC, Zheng B, Wang XH, Gur D. Applying a 2D based CAD scheme for detecting micro-calcification clusters using digital breast tomosynthesis images: An assessment. In: *Proc. SPIE 6915, Medical Imaging 2008: Computer Aid Diagnosis*, 691507.
- [27] Bernard S, Muller S, Onativia J. Computer-aided microcalcification detection on digital breast tomosynthesis data: A preliminary evaluation. *Digital Mammography*. Springer; 2008. p. 151–7.
- [28] Kim DH, et al. Latent feature representation with depth directional long-term recurrent learning for breast masses in digital breast tomosynthesis. *Phys Med Biol* 2017;62(3):1009–31.
- [29] Samala RK, et al. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. *Med Phys* 2016;43(12):6654–66.
- [30] Samala RK, Chan HP, Hadjiiski L, Helvie MA, Richter CD, Cha KH. Breast cancer diagnosis in Digital Breast Tomosynthesis: effects of training sample size on multi-

- stage transfer learning using deep neural nets. *IEEE Trans Med Imaging* 2019;38 (3):686–96.
- [31] Fotin SV, Yin Y, Haldankar H, Hoffmeister JW, Periaswamy S. Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. In: *Proceedings of the SPIE 9785, Medical Imaging*; 2016, 97850X.
- [32] Yousefi M, Krzyzak A, Suen CY. Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. *Comput Biol Med* 2018;96:283–93.
- [33] Sakai A, et al. A method for the automated classification of benign and malignant masses on digital breast tomosynthesis images using machine learning and radiomic features. *Radiol Phys Technol* 2020;13:27–36.
- [34] Bevilacqua V, Brunetti A, Guerriero A, Trotta GF, Telegrafo M, Moschetta M. A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. *Cogn Syst Res* 2019;53:3–19.
- [35] Geras KJ, Mann RM, Moy L. Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. *Radiology* 2019; 293:246–59.
- [36] Website: <https://it.mathworks.com/help/deeplearning/ref/alexnet.html>. Accessed on 01/20/2021.
- [37] Website: <https://it.mathworks.com/help/deeplearning/ref/vgg19.html>. Accessed on 01/20/2021.
- [38] Michell MJ, Batohi B. Role of tomosynthesis in breast imaging going forward. *Clin. Radiology* 2018;73:358–71.
- [39] Kingma DP, BaAdam JL. A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)* 2015.
- [40] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Grad-CAM Batra D. Visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision (ICCV)*, Venice 2017:618–26. <https://doi.org/10.1109/ICCV.2017.74>.
- [41] Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. *Semin Cancer Bio* 2020. <https://doi.org/10.1016/j.semcancer.2020.06.002>.
- [42] Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology* 2013;267(1):47–56.