Multi-label Transfer Learning for the Early Diagnosis of Breast Cancer

Article *in* Neurocomputing · April 2019

DOI: 10.1016/j.neucom.2019.01.112

3 authors, including:

Hiba Chougrad
Université Chouaib Doukkali
7 PUBLICATIONS **117** CITATIONS

SEE PROFILE

Multi-label Transfer Learning for the Early Diagnosis of Breast Cancer

Hiba Chougrad,[a,*] Hamid Zouaki,[a] Omar Alheyane [b]

[a]*Laboratory of Computer Science and Mathematics and their Applications (LIMA), Faculty of science, University Chouaib Doukkali, El Jadida 24000, Morocco*

[b]*Laboratory of Fundamental Mathematics (LMF), Faculty of science, University Chouaib Doukkali, El Jadida 24000, Morocco*

Abstract

Early diagnosis of breast cancer, when it is small and has not spread, can make the disease easier to treat which increases the patient's chances of survival. The recent proposed methods for the early diagnosis of breast cancer, and while showing great success in achieving this goal, rely on one of the indicators in the mammogram to diagnose the patient's condition. Whether it is identifying differences in shapes and patterns of the findings (i.e. masses, calcifications...etc.) or assessing the breast density as a risk indicator, these Computer-aided Diagnosis (CAD) systems by using single-label classification, fail to exploit the intrinsic useful correlation information among data from correlated domains.

Rather than learning to identify the disease based on one of the indicators, we propose the joint learning of the tasks using multi-label image classification. Furthermore, we introduce a new fine-tuning strategy for using transfer learning, that takes advantage of the end-to-end image representation learning when adapting the pre-trained Convolutional Neural Network (CNN) to the new task. We also propose a customized label decision scheme, adapted to this problem, which estimates the optimal confidence for each visual concept. We demonstrate the effectiveness of our approach on four benchmark datasets, CBIS-DDSM, BCDR, INBreast and MIAS, obtaining better results compared to other commonly used baselines.

Keywords: *Computer-aided diagnosis; Breast cancer; Multi-label classification; Convolutional neural network; Transfer learning; Fine-tuning*.

1. Introduction

Breast cancer is the most common cancer in women, and the second most common cause of death by cancer after lung cancer. [1]. About 40,920 women in the U.S. are expected to die in 2018 from breast cancer, though death rates have been decreasing since 1989 [1]. These decreases are thought to be the result of treatment advances, earlier detection through screening, and increased awareness [1]. Early diagnosis of breast cancer continues to be the best way to save lives and decrease healthcare costs over time. Technologies to detect and diagnose breast cancer continue to advance for the purpose of giving patients less invasive options and better diagnoses [2].

Mammography is the primary factor in breast cancer mortality reduction, despite the potential drawbacks to the procedure [3]. In fact, reading a mammogram accurately is challenging for most radiologists. In some recent surveys [4], error in diagnosis was the most common cause of malpractice suits against radiologists. The majority of such cases arose from failure to diagnose breast cancer on mammography [4]. To reduce

* Corresponding author
E-mail address: chougrad.h@ucd.ac.ma
Manuscript accepted in Neurocomputing, published manuscript can be found at https://doi.org/10.1016/j.neucom.2019.01.112

the rate of false-negative diagnoses, lesions with a 2% chance being malignant are recommended for a biopsy [5]. However, only 15 to 30% of the biopsies are found to be malignant [5]. Benign biopsies cause many negative consequences which include fear, pain, anxiety, direct financial expenses, indirect costs related to work missed, and risk of complications [6–8].

One way researchers have sought to improve the performance of mammography and increase the accuracy of the diagnoses was through a better estimation of breast cancer risk on the basis of mammography findings [9,10].

Mammograms are normally subject to multiple annotations. The labels commonly attempt to describe the density of the breast according to the BI-RADS categories, the type of findings (masses, calcifications, etc.) and the pathology [11]. Initially, the radiologist will inspect the images looking for abnormalities in the form of masses, calcifications or other. Masses are defined as three-dimensional and occupy space with completely or partially convex-outward borders. When a new mass is identified, a diagnostic evaluation maybe warranted. Calcifications are deposits of calcium salts in the breast. Sometimes the calcifications can be associated with cancer. Certain characteristics of the calcifications help the radiologist decide if further action is needed. The Breast Imaging Reporting and Data System (BI-RADS) [12] was developed in part to improve the predictive capability of mammography. Radiologists classify density for each mammographic examination into one of four categories, as defined in the BI-RADS lexicon, fourth edition [13]. Approximately 9% of women have almost entirely fatty breasts (BI-RADS I), 40% have scattered fibro-glandular densities (BI-RADS II), 45% have heterogeneously dense breasts (BI-RADS III), and 6% have extremely dense breasts (BI-RADS IV) [13]. Dense breasts are defined as BI-RADS density categories III or IV. Thus, approximately 50% of the population who undergo mammography, have been categorized as having dense breasts [14]. The risk of breast cancer is higher for women with higher breast densities. It has been reported that women with a high breast density compared to women with a low breast density have a four to six fold increased risk of developing the disease [15,16]. After careful annotations, the radiologist is then compelled to provide an accurate, specific, and sufficiently comprehensive diagnosis from the mammogram, to enable the clinician to estimate the prognosis and develop an optimal plan of treatment.

Computer-aided diagnosis (CAD) systems were proven very efficient in recognizing patterns in mammogram images that might suggest malignancy [16–28]. Automated screening of mammograms or computer-aided diagnosis (CAD) of breast cancer is a vast field of research. In [29,30], authors provide an extensive review on different stages of a CAD methodology for breast cancer. To the best of our knowledge, most existing works focus on single-label classification problems, where each mammogram is assumed to have only one class label from the aforementioned mammogram characteristics. On the one hand, we have those who focused on the types of findings in a mammogram to give a diagnosis. For example in some works [17–19], authors used micro-calcification to identify breast cancer. While in others [20–25] authors focused on the morphology of the contour and shape of the breast mass lesion in mammography images, considering them the most discriminating criterions between benign and malignant masses. In other works [16,26–28] , authors relied on the mammographic density and parenchymal patterns for breast cancer risk assessment. These techniques, although working well, fail to exploit the dependencies that exist between the different annotations to be able to provide a full diagnosis. Multi-label classification is the extension of single-label classification in which the goal is to predict the set of relevant labels for a given input. The inputs used to train such a model for this type of problems have several labels. It differs with multi-class classification in terms of output label space, labels are not assumed to be mutually exclusive and multiple labels may be associated with a single training example [31,32]. Historically, multi-label classification has primarily been applied to

text categorisation [33] and medical diagnosis [34]. More recently, the use of multi-label classification has become increasingly popular for a variety of problem types [35], such as image and video annotation [36,37], genomics [38–40] and sentiment classification [41,42].

There are two prevailing categories of algorithms in multi-label learning, namely algorithm adaptation and problem transformation [34,35]. Algorithm adaptation methods extend specific learning models to handle the multi-labelled data. On the other hand, problem transformation algorithms transform the multi-label learning task into either several binary classifications or one multi-class classification problem. Another category is multi-label ensemble methods [31], where models are defined as meta-algorithms based on the top of common multi-label learners. Many other methods have been proposed in the literature, some of which exploit graphical models to capture the label dependencies and conduct structured classification, including those using Bayesian networks [43–45] and conditional random fields [46].

Work on CAD for mammography [47–49] has been done since the early nineties. However, most of the proposed methods were developed on private data sets [50–52] which are not always shared and algorithms which are difficult to compare [49]. Moreover, the proposed methods for early diagnosis of breast cancer relied on only one of the indicators in the mammogram to diagnose the patient's condition. Whether it is identifying differences in shapes and patterns of the findings (i.e. masses, calcifications...etc.) or the assessment the breast density. Based on one of these indicators, separate CAD systems are then developed to identify the disease. A drawback of such methods is that they completely ignore the interdependencies among the multiple labels, we therefore felt motivated to improve upon the state-of-the art and propose the joint learning of the tasks using multi-label image classification.

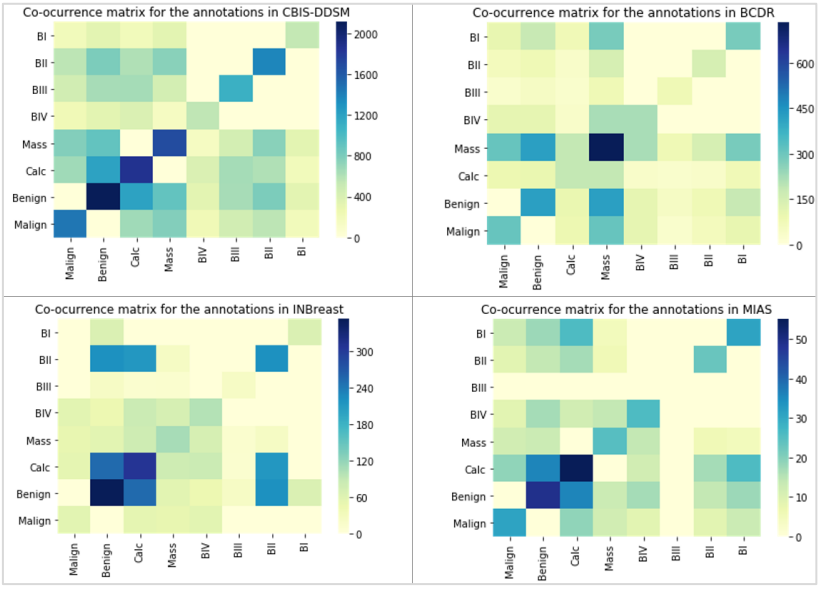

Fig. 1 – The co-occurrence matrices for the 8 labels in the four benchmark datasets (CBIS-DDSM, BCDR, INBreast and MIAS), to show the possible correlations and interdependencies that may exist between labels. For instance the masses or calcifications found in a BI or BII density tissue are most likely to benign than malignant, this is illustrated in all four matrices with a darker blue cell for the benign label compared to the malignant label in both columns of density BI and BII.

Figure 1 demonstrates that a strong correlation exists between some of the labels. For instance, we notice a strong dependency between the findings and the BI-RADS density class IV in most datasets, which is noticeable in the bottom left corner of the correlation matrices. We can

also see that the masses and calcifications found in the BI-RADS I or II density tissue are most likely to be benign and this in all datasets. The matrices also indicate that most of the malignant cases are highly linked to the finding of a mass. To that end, modelling the rich semantic information contained in a mammography image and the possible label dependencies that may exit, is essential for understanding the image as a whole. Therefore, it is a practical and important problem to be able to design a framework that can accurately assign multiple labels to a suspected region.

On a separate note, transfer learning [53,54] is an attractive technique when dealing with little data, which is the case in the medical domain. Our last work [21] demonstrated that using transfer learning, from natural images with fine-tuning, we could efficiently learn from mammography image datasets and achieve better results than when learning from scratch. Some of the limitations of the work were due to the texture of some of the images, i.e. when examining these images, we noticed that the texture of some of the benign and malignant mass lesion images was similar and this resulted in a misclassification of the suspected region. Most of the misclassified mass lesion images were also labelled as highly dense. Accordingly, research showed that cancer is more difficult to detect, in mammograms of women with radiographically dense breasts [55]. Breasts are composed of lobules, ducts, fatty and fibrous connective tissue. The breasts are dense in the presence of a lot of glandular tissue and not much fat. On mammograms, dense breast tissue looks white, while breast masses or tumours also look white. Therefore, the dense tissue hides the potential findings. On the other hand, fatty tissue looks almost black, and on a black background, it is clearly easier to identify a tumour that looks white (see Figure 2). Therefore, mammograms can be less accurate in women with dense breasts.

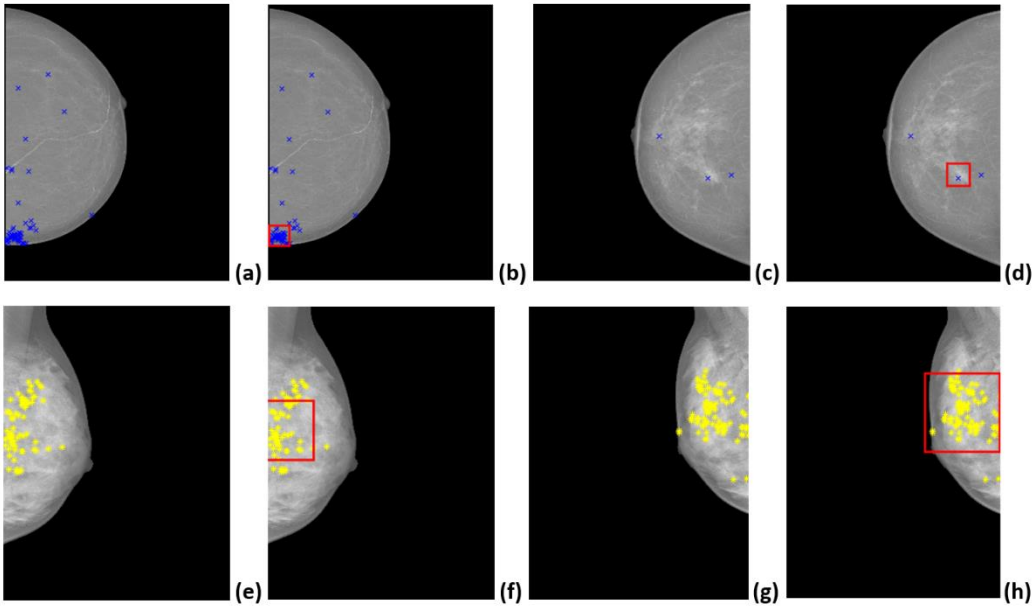

Fig. 2– Comparison between the markings of suspected regions in dense and non-dense mammograms; First row of the figure gives the example of a non-dense mammogram showing more "grey to black" fatty tissue and the second row illustrates a dense mammogram showing more white glandular tissue. (a) ,(c), (e) and (g) highlight the suspected regions marked by imaging specialists while (b) ,(d), (f) and (h) show the delimited region of interest to be cropped. The regions of interest in (b) and (d) the non-dense mammograms are well-targeted and easier to find compared to (f) and (h) the dense mammograms.

Inspired by the success of Convolutional Neural Networks (CNNs) in single-label mammography classification [17,18,21] , we seek to build an end-to-end deep learning framework for multi-label breast lesion classification. We want to take advantage of the very expressive

convolutional neural network architecture (CNN) [56] to build an automatic multi-labelling framework able to help assist the radiologist in giving a full report and more accurate diagnoses to his patients. This is a follow-up and improvement of our last work [21], which extends the image classification from a single-label to a multi-label problem. In this work, we compare the performance of the CNN while using different initialization and optimization procedures. When fine-tuning we propose to use the new strategy we previously presented as a preliminary proposal in [57], but this work goes in much more depth to give detailed explanations and extensive experimentations that underline the superiority of the proposed approach. The method is a new training procedure for fine-tuning when using transfer learning, the idea behind it is that when fine-tuning we don't want all the weights to change in the same manner, we want some of the layers to be more or less receptive to change, depending on their nature. Accordingly, the proposed fine-tuning strategy optimizes the model using SGD momentum with an exponentially decaying learning rate to customize all the pre-trained weights and make them more suited to our type of data. The per-layer decaying learning rate helps control the rate at which weights change in each part of the network i.e. the change will be small to non-existent as we go backwards in the network towards the first layers. The results obtained from using this approach show robustness and efficiency when predicting labels for new breast lesions, whether trained on a small or a slightly larger dataset. We also adopt a final decision labelling mechanism adapted to this task and we evaluate the proposed approach on four benchmark datasets and using many evaluation criteria.

The remainder of this paper is organized as follows: section 2 formulates the problem and describes the methodology proposed to solve the issue at hand. In section 3 we provide details of the experimentations lead to evaluate the proposed approach, and we give the results along with a discussion to analyze them in section 4. Finally section 5 concludes the paper.

2. Materials and Methods

2.1 Problem formulation

Let $\mathcal{D} = \{(x_i,\ y_i)\}_{i=1}^{N}$ be our dataset, the task is a multi-label classification problem, where the input is a Region-Of-Interest (ROI) image $x \in \mathbb{R}^d$ with $d = r\ x\ r$ and the output is the set of labels $y \in \mathcal{Y} = \{0,1\}^L$.

Each instance ROI is associated with multiple class labels $\mathcal{L} = \{1, \dots, L\}$, in this case ($L = 8$) and the predefined set of class labels is {1: BI-RADS I, 2: BI-RADS II, 3: BI-RADS III, 4: BI-RADS IV, 5: Mass, 6: Calcification, 7: Benign, 8: Malignant}.

For an image x the target is the label set $y = [y_1, \dots, y_L] \in \boldsymbol{\mathcal{Y}}_j$ where $y_j = 1$ if the j[th] label is relevant to x; else $y_j = 0$.

The goal is to learn a multi-label prediction model $f(x)$ that maps a given unlabelled image to the L-dimensional label space $\mathcal{Y}$ representing the confidence scores: $f(x)$: $\mathbb{R}^d \rightarrow \mathbb{R}^L$

We adopt a label power set transformation where each label combination becomes a single-class label, the method takes into consideration the correlations that may exist between labels. The multi-label problem is transformed to a single-label multi-class classification problem, where the possible values for the transformed class attribute is the set of distinct unique labels present in the original training data.

While this methods tends to be expensive for highly complex models, where the output space dimensionality is an M-dimensional class-label vector ($M = 2^{L)}$. In our case, the number of possible combinations is smaller, because of the nature of the problem. We consider a collection

of C distinct subsets of labels s of S, the output space is $m = \prod_{s \subset S} \mathcal{P}_s$ where $m \leq M$ and $s_i \subset S$ with $i \in \{1: BI-RADS\ category, 2: findings, 3: Pathology\}$, such as; for s_1 we have $c = 4$ labels will annotate the breast density according to the ACR BI-RADS categories i.e. an annotated breast is associated with one and only one category, which means we'll a have a one-hot encoded vector and $\mathcal{P}_{s_1} = A_c^p = \frac{c!}{(c-p)!}$ possibilities, c for the number of BI-RADS categories and p = 1.

For the second subset of labels s_2, we have $c = 2$ annotates the findings in the mammogram as masses and calcifications. The annotation can be none, one or both of them which gives $\mathcal{P}_{s_2} = 2^c$.

While for s_3 same as s_1, the $c = 2$ labels will annotate the pathology as either "Benign" or "Malignant" which gives $\mathcal{P}_{s_2} = A_c^p = \frac{c!}{(c-p)!}$ possibilities, c for the number of possible pathologies and p = 1.

We learn a label prediction model $f(x) \in \mathbb{R}^L$ that produces a set of labels $\hat{y} \in \mathcal{Y} = \{0,1\}^L$ from the confidence scores. To get the final label decision, we combine two simple heuristics, for $s_1$and s_3 we choose the top-k (i.e., choose top k results from a ranked list) where in both we have $k = 1$. And for s_2 we use global thresholding (i.e., choose the labels for which the confidence score is greater than a single threshold σ).

We optimize the threshold σ for s_3 automatically so that we can convert the outputted probabilities into binary predictions. We compute Matthews Correlation Coefficient (MCC) [58,59] between the ground-truth and the predicted label at many different thresholds and then selects the best σ to apply to the testing data.

We finally measure the similarity between y and $\hat{y}$ using different evaluation criteria adapted to multi-label image classification, which will be given in details in section 3.2. We note that our metrics are different from the standard metrics used for evaluating the robustness of a model on natural images in computer vision. This is because we are dealing with medical data, and our approach focuses on the practical setting of multi-label image classification where the expected output of the system is a set of labels with confidence scores, rather than a ranked list of the most relevant labels describing the most prominent visual concepts in an image.

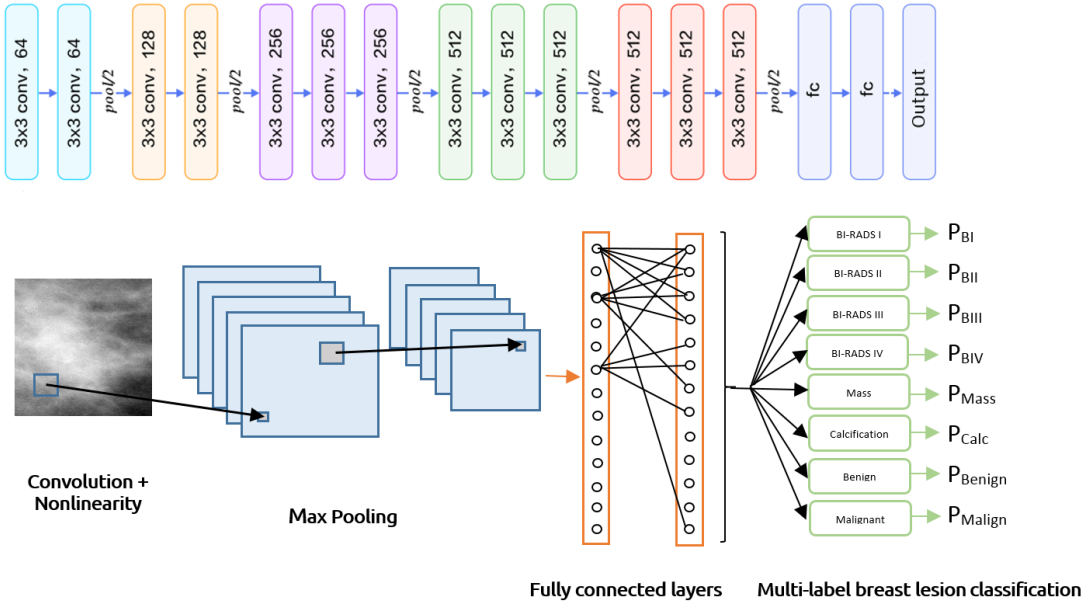

Fig. 3–The architecture of the CNN model for multi-label breast lesion classification. In row 1 we give details about the size and number of layers composing the model. Row 2 highlights the major stages of the classification process. There are five convolutional stages and two fully-connected layers interspersed by max-pooling layers. Probabilities for the output labels are computed in an end-to-end fashion depending on the image representation.

2.2 Model architecture and proposed approach

To take advantage of the end-to-end image representation learning, we use VGG16 [60] pre-trained on the ImageNet ILSVRC challenge dataset [61] as our CNN model. We modify the network to make it output a vector of binary labels indicating the possible annotations for the ROI. We replace the final fully connected layer with a fully connected layer producing an 8-dimensional output, after which we apply a sigmoid nonlinearity. The final output is the predicted probability distribution over the L class labels (see figure 3).

We also modify the loss function to optimize the sum of binary cross entropy losses: $\mathfrak{L}(x,y) = -\frac{1}{N}\sum_{j=1}^{L}\left[y_j\, log\, \hat{y}_j + (1-y_j)\, log(1-\hat{y}_j)\right]$, where $\hat{y}_j = \frac{1}{1+e^{-\theta x_j}}$

Given our training set of tuples (x, y), x representing an image and y the corresponding image labels. We learn a prediction model $f(x;\ \theta) \in \mathbb{R}^L$ with parameters θ which are the set of weights on each layer that we are trying to learn and adapt to our task by solving the following optimization problem:

$$J(\theta) = \min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathfrak{L}(f(x_i;\ \theta),\ y_i) + R(\theta) \qquad (1)$$

With $\mathfrak{L}(f(x_i;\ \theta),\ y_i)$ the loss function and $R(\theta)$ a regularization term.

Stochastic Gradient Descent [62] optimizes stochastically the objective function $J(\theta^{(t)})$ at iteration t with respect to the model's parameters θ. This is done by updating the parameters in the opposite direction of the gradient of the objective function $g^{(t)} = \nabla_\theta J(\theta^{(t)})$. The learning rate $\eta^{(t)}$ determines the size of the steps we take to reach the minimum.

$$\Delta\,\theta^{(t)} = \eta^{(t)} g^{(t)} = \eta^{(t)}\,\nabla_\theta\, J(\theta^{(t)}) \qquad (2)$$

SGD tends to oscillate across the slopes of areas where the surface is not evenly steeply [63]. The momentum technique helps accelerate SGD in the relevant direction by adding a fraction μ of the update vector of the past time step to the current update vector (μ is the momentum):

$$\Delta\,\theta^{(t)} = \mu\,\Delta\,\theta^{(t-1)} - \eta^{(t)}\, g^{(t)} \quad (3)$$

This way, it increases the learning rate for parameters, for which, the gradient persistently points in the same direction, while in the meantime, it is decreasing it for those for which the gradient is changing fast [64,65]. Additionally, and before for-looping, the dataset is randomly shuffled and batches are generated, by collecting subsets of the training set, which helps stabilize the SGD optimization to give better results [66–68].

Over the years, many variants of the gradient descent optimizer have been proposed. In practice, it has been shown that adaptive optimizers [69–71] converge faster than SGD, but their final performance tends to be worse. SGD usually achieves a better minimum but it can take much longer. It is also much more reliant on a robust initialization which can be quite challenging to tune in practice. Recently, it has been shown that it was possible to combine both worlds [72]. Starting off with an adaptive optimizer then switching to SGD with momentum to achieve that peak performance. Initializing the weights with pre-trained and already accurate values will almost always give good convergence as it was shown by many [23,73,74].When using transfer learning, we won't be worrying about initialization. We can start-off using an adaptive optimizer like ADAM [75] to train the model after modifying it for our number of outputs, then proceed with a fine-tuning strategy to get a better performance.

According to [54], the first layers of a CNN learn generic features, while the last layers tend to be more specific to the data. When freezing the top layers and fine-tuning only the last ones, we help the model learn more data-specific features [21]. As such, the weights of the last convolutional layers need to be tweaked as much as possible, while the ones from the top layers need to remain nearly untouched. In accordance with this, we propose a per-layer exponentially decaying learning rate for fine-tuning all of the network's layers while controlling the amount of change occurring in the weights depending on which layer where at. Let $\eta_l^{(t)}$ be the learning rate at the t^{th} iteration and for the l^{th} layer. We propose to modify $\eta_l^{(t)}$ as follows: $\eta_l^{(t)} = \eta_0 e^{-\lambda l}$ with λ a parameter to fine-tune.

The update equation for SGD momentum with the exponentially decaying learning rate is then:

$$\Delta\,\theta_l^{(t)} = \mu\,\Delta\,\theta_l^{(t-1)} - \eta_l^{(t)}\, g^{(t)} = \mu\,\Delta\,\theta_l^{(t-1)} - \eta_l^{(t)}\,\nabla_\theta\, J(\theta_l^{(t)}) \qquad (4)$$

3. Experimentations

3.1 Data

We used four publicly available datasets, CBIS-DDSM [76] , BCDR [77], INBreast[78] and MIAS[79], and adopted the same pre-processing steps described in [21] for cropping, normalizing and augmenting the images. First we cropped fixed sized regions of interest using the ground truth provided with the datasets and resized them to 224 x 224. Then, we used global contrast normalization where every image was normalized by subtracting the mean and dividing by the standard deviation of its elements. Next, we used data augmentation by applying series of random

transformations to the images i.e. width and height shifts by a fraction of 0.25, random rotation that ranges from 0 to 40 degrees and random horizontal flip to generate batches of tensor image data with real-time data augmentation. At each learning epoch, transformations with randomly selected parameters, within the specified range, are applied to all original images in the training set. After an epoch is completed, training data is once again augmented by applying transformations to the original training data for the next learning epoch. This way, the number of times each image is augmented is equal to the number of learning epochs. Thus, the learning algorithm almost never sees the exact same training example twice, because at each epoch training examples are randomly transformed. Each image patch holds its corresponding set of 8 labels, 4 labels to annotate the density, 2 others to annotate the types of findings, and the last 2 labels for the pathology. Figure 4 illustrates the distribution of the three subsets of class labels in each dataset.

CBIS-DDSM: (Curated Breast Imaging Subset of DDSM) [76] is an updated modernized version of the Digital Database for Screening Mammography (DDSM) [80]. The images have been decompressed and converted to DICOM format. Updated ROI segmentation and pathologic diagnosis are given alongside the mammograms, which wasn't the case for DDSM [80] where annotations for the abnormalities weren't very precise. The database contains pixel-level annotations of the ROIs including pathology labelling: benign or malignant, types of findings: masses or calcifications and ACR class density labelling. We used 3564 ROI lesions extracted from the mammograms of 1566 patients by combining the training and test sets for both mass and calcification classes proposed in [76].

BCDR: BCDR-F03 dataset [77] from the Breast Cancer Digital Repository (BCDR-FM) contains 736 biopsy-proven lesions from 344 patients. We used all of the 736 ROIs contained in the dataset, which were all provided with proper annotations to describe their density, pathology and findings.

INBreast: The database [81] contains full-field digital images as opposed to digitized images from films. It contains 410 images from 115 patients. The database is fully-annotated, and therefore is another excellent source to test the multi-label classification scheme. We used all of the 410 ROIs with their density, pathology and types of findings annotations.

MIAS: The Mammographic Image Analysis Society [79] contains 322 digitized film mammograms, we used a subset of 80 images of lesions from 40 patients. All of the mammogram images included the radiologist's ground-truth markings on the locations of the suspected lesion while also providing annotations for the density, findings and severity of the abnormality.

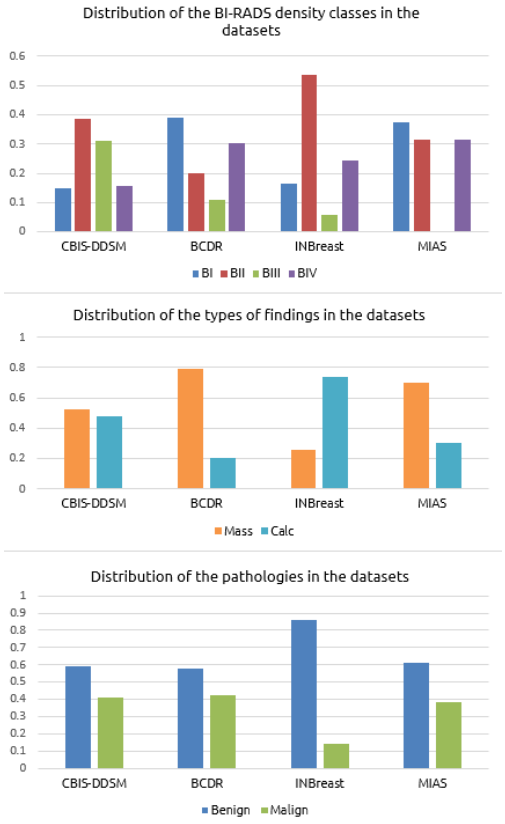

Fig. 4 –The number of images with respect to each of the 8 different labels composing the 3 subsets categories (density, findings and pathology) in all the datasets (CBIS-DDSM, BCDR, INBreast and MIAS).

3.2 Performance evaluation

We measure the performance of our models by comparing the labels predicted with the true labels given by expert radiologists. We used six evaluation criteria named: F1 score, Mean average precision, Hamming loss, Ranking loss, Coverage and Exact match.

Given a test set $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ with N samples, y_N is the set of true labels for the image x_N. A label set y_N is composed of $\{c_1, \dots, c_L\}$ with L possible class labels.

y_N is the set of true labels for the Nth sample and $\hat{y}_N$ the predicted set of labels for x_N. Let $f(x_N, y)$ be a function that returns a value which indicates the confidence that y is in $\hat{y}_N$, and $r(x_N, y)$ is the rank of y in the sorted list of predicted labels with respect to $f(x_N, y)$ where the rank of the label with the highest confidence $(arg \max_{y \in \{c_1, \dots, c_L\}} f(x_N, y))$ is 1. We computed the following evaluation criteria that are commonly used for multi-label classification [82,83]:

- $F1\_Score(f) = \frac{1}{N} \sum_{n=1}^{N} \frac{2\,|\hat{y}_n \cap y_n|}{|\hat{y}_n| + |y_n|}$ is the harmonic mean of precision and recall. It is a balanced measure between the positive predictive value (precision) and the ratio of correctly predicted positive values to the actual positive values (recall). As in single label multi-class classification, the higher the value of the F1 score, the better the performance of the learning algorithm.

- *Mean_Average_Precision(f)* $= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|y_n|}\sum_{y\in y_n}\frac{|\{y' \mid r(x_n,y')\leq r(x_n,y),\ \ y'\in y_n\}|}{r(x_n,y)}$ mAP is the average fraction of reference labels ranked above a particular label. The performance is better as the average precision gets closer to 1.
- *Hamming_Loss(f)* $= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}|\hat{y}_n \Delta\, y_n|$ Where Δ is an operator used to compute the symmetric difference between two label sets. The Hamming loss is the fraction of the wrong labels (false positives or false negatives) to the total number of labels. Thus, a smaller value indicates better performance.
- *Ranking_Loss(f)* $= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|y_n||\overline{y_n}|}|\{(y_1,y_2) \mid r(x_n,y_1) \geq r(x_n,y_2),(y_1,y_2)\in y_n\times\overline{y_n}\}|$, Where $\overline{y_n}$ denotes the complement of the set y_n. Ranking loss counts the number of times the rank of a wrong label is above the rank of a reference label. Hence, a smaller value indicates better performance.
- *Coverage(f)* $= \frac{1}{N}\sum_{n=1}^{N} max_{y\in y_n} r(x_n,y) - 1$. Coverage is defined as the average number of labels to investigate on the ordered list of predicted numbers (by their ranks) to cover all reference labels. Since fewer number of visits on the list is desired, a smaller value indicates a better performance.
- *Exact_Match(f)* $= \sum_{n=1}^{N} I[\hat{y}_n == y_n]$. The Exact match is a severe metric for multi-label classification, since it does not include the additional notion of being partially correct. $\hat{y}_n$ and y_n are the predicted and the ground-truth labels for the n^{th} example in a dataset, and $I[\cdot]$ is an indicator function. This considers the prediction to be correct only if it is the same as the ground-truth labels.

Differently, we use the *Receiver Operating Characteristic (ROC)* analysis and *Area Under the curve (AUC)* for reporting the malignant and benign classification results in a binary classification setting. ROC is a great way to visualize the performance of the built classifier, and the AUC can summarize this performance by assessing the ranking regarding the separation of the two classes. The higher the number, the better are the results. The ROC curve is defined by: $TPR = TP/(TP + FN)$ and $FPR = FP/(FP + TN)$. Where TP stands for the true positive cases in the classification results, and TN denotes the true negative cases. In addition, FP reports the false positive cases and FN the false negative cases. Typically, the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the results are. Therefore, a larger Area Under the Curve stands for a better classification.

3.3 Implementation details

We trained four variants of the VGG model, the first we call (VGG-Sc) was trained from scratch with weights initialized to small random numbers generated from a Gaussian distribution, while we initialized the other three models with pre-trained weights from the ImageNet ILSVRC challenge dataset [61]. All the models have the same architecture composed of 16 layers as proposed in [60] but we replace the number of outputs to 8 probability predictions (see figure 3). For the second model VGG-2FT we froze the 10 first layers and fine-tuned only the 6 last layers, while for the third model VGG-AllFT , and for the purpose of the experiment, we fine-tuned all the 16 layers. The last model

VGG-FTED was trained using our proposed fine-tuning strategy where all of the 16 layers were fine-tuned using an exponentially decaying learning rate.

The choice of hyper-parameter configuration directly influences the performance of the gradient descent algorithm in the training phase of the neural network. Especially, given the fact that, contrary to training shallow models, a deep learning architecture requires the initialization of more hyper-parameters before training. We use grid search to optimize the learning rate, number of epochs, dropout rate, weight decay, learning rate decay parameter and momentum. This way we could construct and evaluate one model for each combination of parameters and retain the combination of parameters that achieved the best score during the optimization procedure. We used appropriate search ranges for each hype-parameter, to ensure that each searched parameter is reasonable. Moreover, the search ranges also help in reducing the computational time of the search procedure, so that the algorithm performs only few search iterations to determine the optimal solution. The values of these search ranges are dependent on the properties of the parameter itself, the model, the dataset and the optimization task to be performed.

Along these lines, we report the parameters values that achieved the best results. VGG-Sc was trained for 50 epochs with SGD with momentum of 0.9 and a learning rate of 1e-2. We optimized all the three other models using a hybrid approach where we started of training the fully-connected layers on top with ADAM for 15 epochs, then we fine-tuned using SGD for 90 epochs. We employed a momentum rate of 0.9, dropout [84] rate 0.2, an initial learning rate of 1e-3 decayed by a factor of 0.1 if it plateaus for 10 epochs, and an early stopping strategy with patience of 15 epochs. We use these same parameters for VGG-FTED but we optimize with the modified SGD (Eqn. 4). For the learning decay rate λ, we tried values ranging from 0 to 0.3 in steps of 0.025, with regard to the 16 layers network architecture used, to ensure that the learning rate gets lower, in an exponentially decaying manner, as we go backwards to the top layers. The best results were achieved with λ set to 0.225. The regularization term in (Eqn. 1) was defined as an L2 norm with a weight decay of 1e-2.

We first train the model's variants on the training set and we fine-tune the hyper-parameters on the validation set. To ensure that no overfitting is happening and that the model performs well on the training data as well as on validation data, we monitored the performance on both sets as illustrated in figure 5.

Once the models were optimized, we fixed all their parameters, then the test set was used as a hold-out set to measure the models' performance on never-seen data. This way we can get less biased results, as the test set was kept unseen. Each dataset was split as follows 70% for training, 10% validation, 20% test. We carry out the same procedure during 5 fold cross-validation for each of training, validation and test partitions. The obtained test results are then reported (see table X).

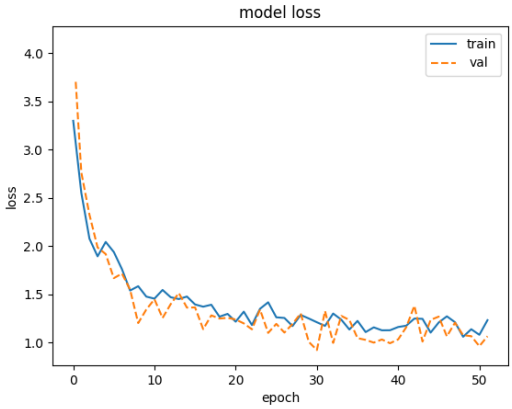

Fig. 5 – The parallel plots of loss over the epochs for the training and validation sets for real-time monitoring (VGG-FTED trained on Inbreast). The model shows comparable performance on the training and validation data.

4. Results and discussion

Table 1 compares the performance of the 4 models on the CBIS-DDSM [76], BCDR [77], INBreast [78] and the MIAS [79] datasets. We report results with respect to the metrics introduced above. The results indicate that the proposed fine-tuning strategy achieves a substantial improvement over existing methods, in terms of all of the used metrics and over all the datasets.

When comparing the models' performance on the different datasets, we can see that the under-achieving model is either VGG-Sc or VGG-AllFT and this in terms of all metrics. On the one hand, VGG-Sc is randomly initialized and is trained on a small dataset while being a deep network which make it very susceptive to overfitting. On the other hand, all the layers of VGG-AllFT were fine-tuned using the same learning rate, causing thereby a change in the pre-trained weights of all the layers while trying to adapt them to the new small dataset.

The baseline model VGG-2FT uses the common recommended fine-tuning strategy [23,54,74] which adapts the last layers of the network to the new task while keeping the first ones untouched. Our proposed fine-tuning method embodied by the VGG-FTED model, is a compromise between the two models VGG-2FT and VGG-AllFT. The layers will all change as in VGG-AllFT, but not in the same manner. Most of the change will occur in the last layers as in the VGG-2FT, but instead of manually controlling which of the layers will receive the updates and which will not (freezing), this would be done automatically. The amount of change gets lower as we go backwards to the top layers, with the exponentially decaying learning rate, ensuring this way that the main changes happen is the last layers where the model learns task specific features, while the rest of the weights in the top layers remain almost the same. The VGG-FTED method outperforms the VGG-2FT baseline by 4 to 8% in terms of Exact Match accuracy.

Experiments showed that transfer learning is advantageous to our task of interest, even if it is between two unrelated tasks. The use of pre-trained weights for initialization is a good way to start the learning process. The next step is to gradually fine-tune the loaded weights to adapt the network to the new dataset. This is achieved by resuming backpropagation on the layers with a small learning rate.

Table 1 – Comparison summary of the different model's implementations and their performance on the four benchmark datasets.

	Metric / Method	F1	mAP	HL	RL	COV	EM
DDSM	VGG-Sc	0.668 ±0.025	0.590 ±0.018	0.260 ±0.019	0.397 ±0.043	6.498 ±0.250	0.277 ±0.033
	VGG-2FT	0.912 ±0.022	0.877 ±0.012	0.079 ±0.011	0.137 ±0.016	4.174 ±0.190	0.754 ±0.028
	VGG-AllFT	0.685 ±0.015	0.583 ±0.010	0.227 ±0.013	0.391 ±0.014	6.487 ±0.150	0.312 ±0.020
	VGG-FTED	**0.935** ±0.019	**0.895** ±0.017	**0.047** ±0.022	**0.087** ±0.025	**3.895** ±0.320	**0.822** ±0.041
BCDR	VGG-Sc	0.718 ±0.020	0.574 ±0.033	0.267 ±0.028	0.375 ±0.052	6.579 ±0.191	0.307 ±0.059
	VGG-2FT	0.887 ±0.086	0.862 ±0.201	0.089 ±0.035	0.172 ±0.098	4.580 ±1.133	0.728 ±0.127
	VGG-AllFT	0.731 ±0.038	0.623 ±0.046	0.212 ±0.075	0.332 ±0.059	6.189 ±0.216	0.358 ±0.133
	VGG-FTED	**0.915** ±0.055	**0.889** ±0.097	**0.064** ±0.065	**0.108** ±0.098	**4.203** ±1.148	**0.802** ±0.139
INBREAST	VGG-Sc	0.672 ±0.016	0.465 ±0.021	0.266 ±0.037	0.386 ±0.022	5.764 ±0.079	0.076 ±0.020
	VGG-2FT	0.930 ±0.110	0.788 ±0.134	0.057 ±0.110	0.097 ±0.152	4.103 ±0.225	0.776 ±0.118
	VGG-AllFT	0.687 ±0.028	0.492 ±0.079	0.173 ±0.045	0.305 ±0.037	5.325 ±0.072	0.460 ±0.033
	VGG-FTED	**0.942** ±0.102	**0.887** ±0.140	**0.042** ±0.092	**0.082** ±0.125	**3.723** ±0.147	**0.827** ±0.092
MIAS	VGG-Sc	0.602 ±0.013	0.486 ±0.030	0.481 ±0.055	0.728 ±0.075	6.788 ±0.137	0.000 ±0.000
	VGG-2FT	0.850 ±0.132	0.886 ±0.133	0.096 ±0.135	0.163 ±0.220	4.112 ±1.406	0.730 ±0.120
	VGG-AllFT	0.617 ±0.114	0.673 ±0.115	0.305 ±0.089	0.412 ±0.171	6.004 ±0.741	0.357 ±0.129
	VGG-FTED	**0.907** ±0.150	**0.895** ±0.167	**0.075** ±0.123	**0.120** ±0.098	**3.875** ±1.140	**0.782** ±0.206

As it was proved in our last work [21], too much fine-tuning leads to overfitting and therefore gives worse results. The accustomed way for fine-tuning freezes the top layers of the network and back-propagate through few of the last convolutional layers. This comes from the assumption that the first layers of a CNN learn generic features, while the last layers tend to be more specific to the data. Therefore, the goal is to make the last convolutional layers learn more data-specific features, but there is no need to change the weights of the first convolutional layers as much; they're already well-tuned to learn generic features, especially when we are lacking data to train on.

Fine-tuning with the per-exponentially decaying learning rate (VGG-FTED) follows this idea and improves on it, to make the adaptation happen more naturally and automatically compared to its counterpart. The proposed approach helps us control the rate at which weights change in each part of the network. In other words, when fine-tuning, we want some of the layers to be more or less receptive to change. The last convolutional layers will be highly targeted, while the first layers are the least affected by the change. The middle convolutional layers might be responsible of learning somewhat more complex features compared to the first layers and thus they need to be slightly modified. The exponentially decaying learning rate is best suited to help us in this regard.

Table 2 – Comparison summary of our approach with global thresholding for inferring the final label decision.

	Metric / Method	F1	mAP	HL	RL	COV	EM
DDSM	*Threshold*	0.933 ±0.017	0.893 ±0.020	0.049 ±0.027	0.095 ±0.030	3.933 ±0.240	0.790 ±0.045
	Ours	**0.935** ±0.019	**0.895** ±0.017	**0.047** ±0.022	**0.087** ±0.025	**3.895** ±0.320	**0.822** ±0.041
BCDR	*Threshold*	0.915 ±0.075	0.888 ±0.104	**0.060** ±0.085	**0.105** ±0.108	4.185 ±1.270	0.790 ±0.154
	Ours	0.915 ±0.055	**0.889** ±0.097	0.064 ±0.065	0.108 ±0.098	**4.203** ±1.148	**0.802** ±0.139
INBrst	*Threshold*	0.937 ±0.102	**0.892** ±0.113	0.048 ±0.085	0.084 ±0.164	3.810 ±1.273	0.805 ±0.159
	Ours	**0.942** ±0.102	0.887 ±0.140	**0.042** ±0.092	**0.082** ±0.125	**3.723** ±0.147	**0.827** ±0.092
MIAS	*Threshold*	0.905 ±0.147	0.895 ±0.170	0.079 ±0.149	0.123 ±0.108	3.895 ±1.259	0.779 ±0.144
	Ours	**0.907** ±0.150	0.895 ±0.167	**0.075** ±0.123	**0.120** ±0.098	**3.875** ±1.140	**0.782** ±0.206

Our final label decision output contains the probability predictions of the three subsets of annotations; regarding the class density, the types of findings and the pathology decision. For the class density and the pathology only one label will be relevant, and therefore the top-1 ranked probability seems an evident choice. However, for the types of findings, it seemed better fitted to use global thresholding, for it is less restrictive than top-k; in case we have more than one abnormality in the suspected region. The optimal per-class threshold estimation for the classes "mass" and "calcification" is obtained after careful selection over a range of threshold values.

Table 2 gives a comparison between the use of global thresholding and the proposed combined approach to get the final label decision. We notice that the proposed method outperforms global thresholding for most cases, even though most of the works on multi-label classification recommend using global thresholding to get the final binary output [85–88].

Because of the nature of the problem, the combination of top-k ranking and thresholding outperforms its counterpart over all datasets. This suggests that the current common practice of choosing top-k labels or the global thresholding mechanisms in multi-label classification can be improved depending on the nature of the task.

Additionally, we report the results of the binary classification of the benign and malignant lesions in each dataset. Figure 6 shows the Receiver Operating Curve (ROC) and gives the Area Under the Curve (AUC) results achieved by the proposed VGG-FTED model on the four datasets. We give the ROC response of different datasets, created from 5-fold cross-validation. Taking all of these curves, it is possible to calculate the mean area under curve, and see the variance of the curve when the training set is split into different subsets. This roughly shows how the classifier output is affected by changes in the training data, and how different the splits generated by the cross-validation are from one another. Overall, the proposed model yielded good AUC values, 0.89 for CBIS-DDSM, 0.94 for BCDR, 0.93 for INBreast and 0.86 for MIAS.

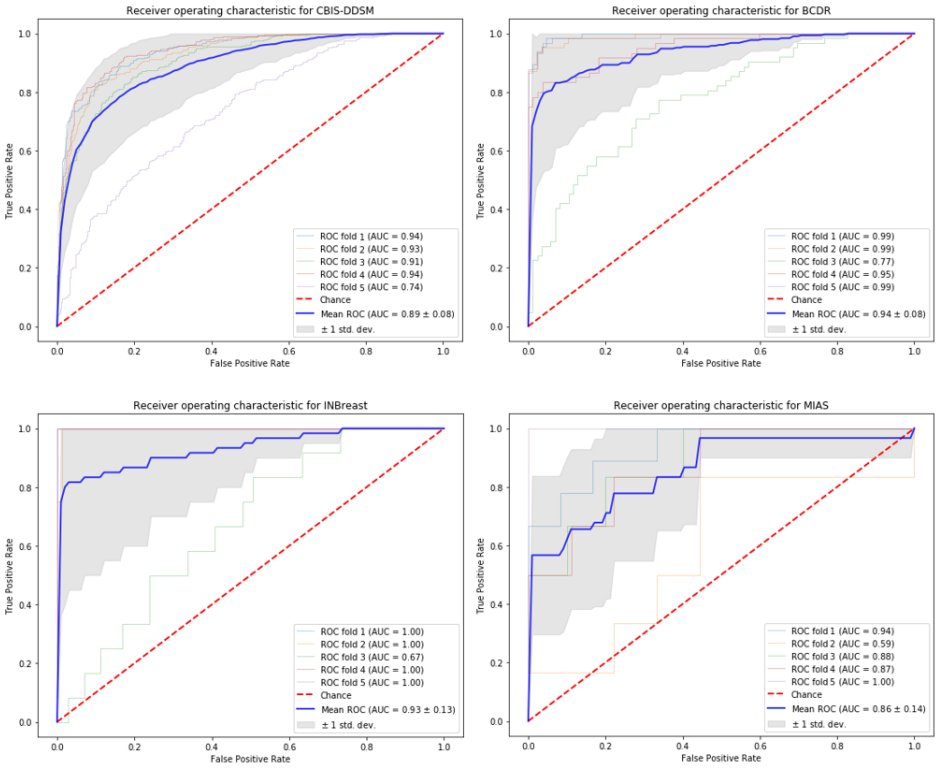

Fig. 6–Receiver Operating Characteristic (ROC) curves and AUC values on 5-fold cross-validation for the binary classification of the four datasets using the VGG-FTED model.

(NEW) Due to some concerns about the data interoperability and the possible differences in clinical acquisition, we could not test the methodology across different datasets, i.e, training on one dataset and testing on another. There can be discrepancies between annotations obtained from different clinicians. Plus, the nature of data can differ from hardware to hardware, which means there may exist large variations in images due to sensors and other factors leading to biased results.

We finally give some image annotation examples in Figure 7. For instance, we show examples of the difference in shape between the masses and calcifications (row 1 and 2 in first column), while (row 3 and 4 in first column) and (row 1 and 2 in the second column) give the difference in shape between malignant and benign masses, and malignant and benign calcifications respectively. (Row 3 and 4 in the second column) illustrates the two types of findings occurring in the same image (i.e. masses and calcifications) for both the benign and malignant cases. We can see from the probability predictions that the proposed method was able to successfully annotate some of the most challenging images. As an illustration, the third column of images gives an example of highly dense breasts that are more difficult to diagnose compared to the ones in the first column which are more likely to be correctly annotated.

Rather than only identifying the type of the pathology as it is commonly done in computer-aided diagnosis systems based on single-label classification models, we proposed a CAD that annotates the region of interest from the correlated sub-domains, as it would be normally done by expert radiologists. This way, the model ends up building a higher level of understanding of the different patterns and features that could be found in an image. For instance, in highly dense breast where the texture of the images makes it harder for the radiologist to identify the

shapes and detect the abnormalities, the proposed CAD can use its acquired knowledge of the label dependencies from training to give higher predictive values to relevant labels regardless of what can be seen in the image.

We would also like to note that the proposed fine-tuning strategy combined with the multi-label learning helped alleviate the risk of overfitting due to the little data problem. Empirically, the performance of the model was alike for small datasets (INBreast and MIAS) as for relatively larger ones (CBIS-DDSM and BCDR). We assume that fine-tuning with the per-layer decaying learning rate, while taking into account label correlations during the classification process, alongside other tricks such as: data augmentation, regularization and dropout, helped us avoid overfitting.

Being capable of providing full annotations, the proposed CAD system gives the radiologist a holistic perspective of the region of interest, which can provide assistance in the decision making and will eventually help him increase the accuracy of his diagnoses.

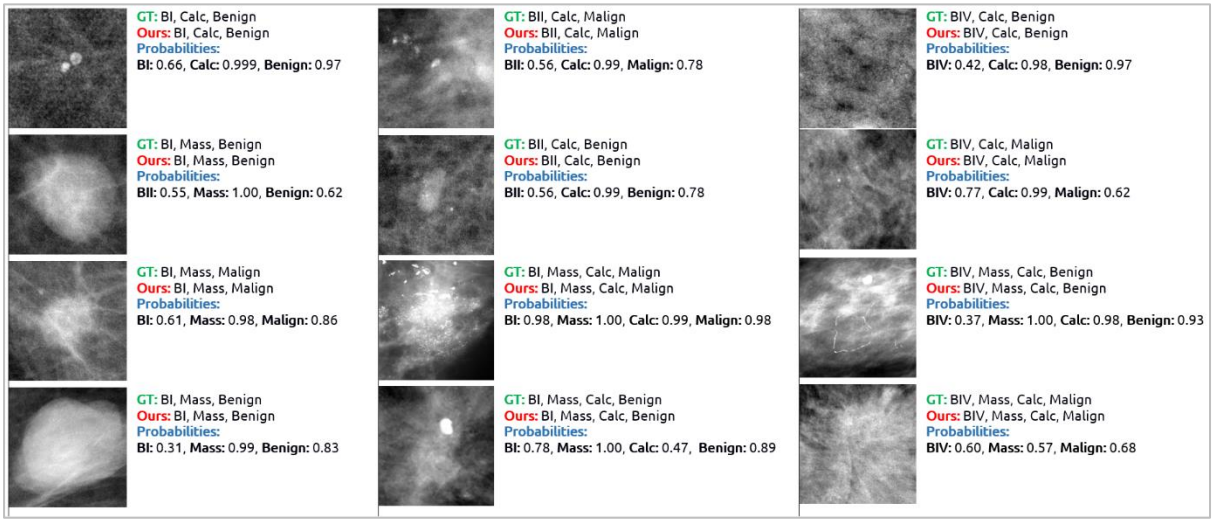

Fig. 7–Examples of predictions on test images from the different datasets. In the 1st column, the images in row 1 and 2 give an example of the difference between masses and calcifications. And row 3 and 4 illustrate the difference between a benign and malignant mass. In the 2nd column, images in row 1 and 2 shows the difference between benign and malignant calcifications. Row 3 and row 4 illustrate the case where masses and calcifications occur in the same time. Column three shows cases of more challenging images because of the class density (BI-RADS IV) unlike column one and two where the findings are easier to identify on a non-dense background (BI-RADS I and BI-RADS II).

5. Conclusions

In this paper, we propose a computer-aided diagnosis system that aims to make use of all the annotations in a mammography image which are normally neglected in the diagnostic building process of other proposed CADs, and this in order to build a CAD system capable of providing a full diagnosis.

The CAD is a multi-label image classification system that aims to capture spontaneous label correlation relationships while taking advantage of an end-to-end image representation learning architecture. The system uses a pre-trained CNN to profit from the attractive properties of transfer learning. This work improved upon existing baselines for fine-tuning, by introducing a new optimization method which uses SGD with an exponentially decaying learning rate. The assumption is that with a learning rate specific to each layer in the network we can focus on

learning more task specific features in the most relevant layers. The method helps improve the domain adaptation, as it maximizes the learning on the new domain when controlling the amount of change happening throughout the convolutional layers. The proposed CAD also adopted a personalized labeling decision scheme adapted to the problem at hand. Our approach for diagnosing breast cancer outperforms all baselines by a large margin and over four benchmark datasets; and this in terms of multiple metrics including the exact match score which is the strictest measure of all for multi-label classification, and this highlights the superiority of our method in the practical setting.

In the future, we would like to explore ways to leverage the distinctive properties of the multi-label problem, to learn a joint image-label embedding, characterizing the semantic label dependency as well as the image-label relevance. We would also like to incorporate imaging modalities other than mammography in the learning process to take advantage of other existing rich representations (ultrasound, DCE-MRI,…etc.) to build an efficient, consistent and powerful computer-aided diagnosis system for the early diagnosis of breast cancer.

Acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The BCDR database used in this work was a courtesy of MA Guevara Lopez and coauthors, Breast Cancer Digital Repository Consortium.

The INbreast database used in this work was a courtesy of the Breast Research Group, INESC Porto, Portugal.

Declarations of interest

None.

REFERENCES

[1] Siegel Rebecca L., Miller Kimberly D., Jemal Ahmedin, Cancer statistics, 2018, CA. Cancer J. Clin. 68 (2018) 7–30. doi:10.3322/caac.21442.

[2] S.A. Narod, J. Iqbal, A.B. Miller, Why have breast cancer mortality rates declined?, J. Cancer Policy. 5 (2015) 8–17. doi:10.1016/j.jcpo.2015.03.002.

[3] S.H. Heywang-Köbrunner, A. Hacker, S. Sedlacek, Advantages and Disadvantages of Mammography Screening, Breast Care. 6 (2011) 199–207. doi:10.1159/000329005.

[4] J.S. Whang, S.R. Baker, R. Patel, L. Luk, A. Castro, The causes of medical malpractice suits against radiologists in the United States, Radiology. 266 (2013) 548–554. doi:10.1148/radiol.12111119.

[5] E.A. Sickles, Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases., Radiology. 179 (1991) 463–468. doi:10.1148/radiology.179.2.2014293.

[6] N.T. Brewer, T. Salz, S.E. Lillie, Systematic review: the long-term effects of false-positive mammograms, Ann. Intern. Med. 146 (2007) 502–510.

[7] B. Yazici, A.R. Sever, P. Mills, D. Fish, S.E. Jones, P.A. Jones, Scar formation after stereotactic vacuum-assisted core biopsy of benign breast lesions, Clin. Radiol. 61 (2006) 619–624. doi:10.1016/j.crad.2006.03.008.

[8] F. Zagouri, T.N. Sergentanis, A. Gounaris, D. Koulocheri, A. Nonni, P. Domeyer, C. Fotiadis, J. Bramis, G.C. Zografos, Pain in different methods of breast biopsy: emphasis on vacuum-assisted breast biopsy, Breast Edinb. Scotl. 17 (2008) 71–75. doi:10.1016/j.breast.2007.07.039.

[9] J.A. Baker, P.J. Kornguth, C.E. Floyd, Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description., Am. J. Roentgenol. 166 (1996) 773–778. doi:10.2214/ajr.166.4.8610547.

[10] C.J. D'Orsi, The American College of Radiology mammography lexicon: an initial attempt to standardize terminology., Am. J. Roentgenol. 166 (1996) 779–780. doi:10.2214/ajr.166.4.8610548.

[11] K.H. Allison, L.A. Abraham, D.L. Weaver, A.N. Tosteson, H.D. Nelson, T. Onega, B.M. Geller, K. Kerlikowske, P.A. Carney, L.E. Ichikawa, D.S.M. Buist, J.G. Elmore, Trends in Breast Tissue Sampling and Pathology Diagnoses among Women Undergoing Mammography in the U.S.: A Report from the Breast Cancer Surveillance Consortium, Cancer. 121 (2015) 1369–1378. doi:10.1002/cncr.29199.

[12] W.A. Berg, C. Campassi, P. Langenberg, M.J. Sexton, Breast Imaging Reporting and Data System, Am. J. Roentgenol. 174 (2000) 1769–1777. doi:10.2214/ajr.174.6.1741769.
[13] American College of Radiology, The American College of Radiology Breast Imaging Reporting and Data System (BI-RADS), in: 2003.
[14] B.L. Sprague, R.E. Gangnon, V. Burt, A. Trentham-Dietz, J.M. Hampton, R.D. Wellman, K. Kerlikowske, D.L. Miglioretti, Prevalence of mammographically dense breasts in the United States, J. Natl. Cancer Inst. 106 (2014). doi:10.1093/jnci/dju255.
[15] G. Torres-Mejía, B. De Stavola, D.S. Allen, J.J. Pérez-Gavilán, J.M. Ferreira, I.S. Fentiman, I. Dos Santos Silva, Mammographic features and subsequent risk of breast cancer: a comparison of qualitative and quantitative evaluations in the Guernsey prospective studies, Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. 14 (2005) 1052–1059. doi:10.1158/1055-9965.EPI-04-0717.
[16] C.M. Vachon, C.H. van Gils, T.A. Sellers, K. Ghosh, S. Pruthi, K.R. Brandt, V.S. Pankratz, Mammographic density, breast cancer risk and risk prediction, Breast Cancer Res. BCR. 9 (2007) 217. doi:10.1186/bcr1829.
[17] A.N. Karahaliou, I.S. Boniatis, S.G. Skiadopoulos, F.N. Sakellaropoulos, N.S. Arikidis, E.A. Likaki, G.S. Panayiotakis, L.I. Costaridou, Breast Cancer Diagnosis: Analyzing Texture of Tissue Surrounding Microcalcifications, IEEE Trans. Inf. Technol. Biomed. 12 (2008) 731–738. doi:10.1109/TITB.2008.920634.
[18] I.I. Andreadis, G.M. Spyrou, K.S. Nikita, A CAD$_{\bf x}$Scheme for Mammography Empowered With Topological Information From Clustered Microcalcifications' Atlases, IEEE J. Biomed. Health Inform. 19 (2015) 166–173. doi:10.1109/JBHI.2014.2334491.
[19] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, L. Li, Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning, Sci. Rep. 6 (2016). doi:10.1038/srep27327.
[20] Z. Jiao, X. Gao, Y. Wang, J. Li, A deep feature based framework for breast masses classification, Neurocomputing. 197 (2016) 221–231. doi:10.1016/j.neucom.2016.02.060.
[21] H. Li, X. Meng, T. Wang, Y. Tang, Y. Yin, Breast masses in mammography classification with local contour features, Biomed. Eng. Online. 16 (2017) 44. doi:10.1186/s12938-017-0332-0.
[22] H. Berment, V. Becette, M. Mohallem, F. Ferreira, P. Chérel, Masses in mammography: What are the underlying anatomopathological lesions?, Diagn. Interv. Imaging. 95 (2014) 124–133. doi:10.1016/j.diii.2013.12.010.
[23] H. Chougrad, H. Zouaki, O. Alheyane, Deep Convolutional Neural Networks for breast cancer screening, Comput. Methods Programs Biomed. 157 (2018) 19–30. doi:10.1016/j.cmpb.2018.01.011.
[24] M. Jiang, S. Zhang, J. Liu, T. Shen, D.N. Metaxas, Computer-aided diagnosis of mammographic masses using vocabulary tree-based image retrieval, in: Biomed. Imaging ISBI 2014 IEEE 11th Int. Symp. On, IEEE, 2014: pp. 1123–1126.
[25] J. Arevalo, F.A. González, R. Ramos-Pollán, J.L. Oliveira, M.A. Guevara Lopez, Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Programs Biomed. 127 (2016) 248–257. doi:10.1016/j.cmpb.2015.12.014.
[26] R.R. Winkel, M. von Euler-Chelpin, M. Nielsen, K. Petersen, M. Lillholm, M.B. Nielsen, E. Lynge, W.Y. Uldall, I. Vejborg, Mammographic density and structural features can individually and jointly contribute to breast cancer risk assessment in mammography screening: a case–control study, BMC Cancer. 16 (2016). doi:10.1186/s12885-016-2450-7.
[27] K. Bovis, S. Singh, Classification of mammographic breast density using a combined classifier paradigm, in: 4th Int. Workshop Digit. Mammogr., 2002: pp. 177–180.
[28] A. Gastounioti, E.F. Conant, D. Kontos, Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment, Breast Cancer Res. 18 (2016). doi:10.1186/s13058-016-0755-8.
[29] M.P. Sampat, M.K. Markey, A.C. Bovik, others, Computer-aided detection and diagnosis in mammography, Handb. Image Video Process. 2 (2005) 1195–1217.
[30] R.M. Rangayyan, F.J. Ayres, J.L. Desautels, A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs, J. Frankl. Inst. 344 (2007) 312–348.
[31] G. Madjarov, D. Kocev, D. Gjorgjevikj, S. Džeroski, An extensive experimental comparison of methods for multi-label learning, Pattern Recognit. 45 (2012) 3084–3104.
[32] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (2014) 1819–1837.
[33] A. McCallum, Multi-label text classification with a mixture model trained by EM, in: AAAI Workshop Text Learn., 1999: pp. 1–7.
[34] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, Int. J. Data Warehous. Min. IJDWM. 3 (2007) 1–13.
[35] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, in: Data Min. Knowl. Discov. Handb., Springer, 2009: pp. 667–685.
[36] M.R. Boutell, J. Luo, X. Shen, C.M. Brown, Learning multi-label scene classification, Pattern Recognit. 37 (2004) 1757–1771. doi:10.1016/j.patcog.2004.03.009.
[37] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, H.-J. Zhang, Correlative multi-label video annotation, in: Proc. 15th ACM Int. Conf. Multimed., ACM, 2007: pp. 17–26.
[38] M.-L. Zhang, Z.-H. Zhou, Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization, IEEE Trans Knowl Data Eng. 18 (2006) 1338–1351. doi:10.1109/TKDE.2006.162.
[39] Z. Barutcuoglu, R.E. Schapire, O.G. Troyanskaya, Hierarchical multi-label prediction of gene function, Bioinformatics. 22 (2006) 830–836.
[40] E.A. Tanaka, S.R. Nozawa, A.A. Macedo, J.A. Baranauskas, A multi-label approach using binary relevance and decision trees applied to functional genomics, J. Biomed. Inform. 54 (2015) 85–95. doi:10.1016/j.jbi.2014.12.011.

[41] S.M. Liu, J.-H. Chen, A multi-label classification based approach for sentiment classification, Expert Syst. Appl. 42 (2015) 1083–1093. doi:10.1016/j.eswa.2014.08.036.
[42] F. Bravo-Marquez, E. Frank, S.M. Mohammad, B. Pfahringer, Determining Word-Emotion Associations from Tweets by Multi-label Classification, in: 2016 IEEEWICACM Int. Conf. Web Intell. WI, 2016: pp. 536–539. doi:10.1109/WI.2016.0091.
[43] P.R. De Waal, L.C. Van Der Gaag, Inference and learning in multi-dimensional Bayesian network classifiers, in: Eur. Conf. Symb. Quant. Approaches Reason. Uncertain., Springer, 2007: pp. 501–511.
[44] J.D. Rodríguez, J.A. Lozano, Multi-objective learning of multi-dimensional Bayesian classifiers, in: Hybrid Intell. Syst. 2008 HIS08 Eighth Int. Conf. On, IEEE, 2008: pp. 501–506.
[45] C. Bielza, G. Li, P. Larranaga, Multi-dimensional classification with Bayesian networks, Int. J. Approx. Reason. 52 (2011) 705–727.
[46] N. Ghamrawi, A. McCallum, Collective multi-label classification, in: Proc. 14th ACM Int. Conf. Inf. Knowl. Manag., ACM, 2005: pp. 195–200.
[47] S.M. Astley, F.J. Gilbert, Computer-aided detection in mammography, Clin. Radiol. 59 (2004) 390–399. doi:10.1016/j.crad.2003.11.017.
[48] R.M. Nishikawa, Current status and future directions of computer-aided diagnosis in mammography, Comput. Med. Imaging Graph. 31 (2007) 224–235.
[49] M. Elter, A. Horsch, CADx of mammographic masses and clustered microcalcifications: a review, Med. Phys. 36 (2009) 2052–2068.
[50] N.R. Mudigonda, R. Rangayyan, J.L. Desautels, Gradient and texture analysis for the classification of mammographic masses, IEEE Trans. Med. Imaging. 19 (2000) 1032–1043.
[51] B. Zheng, X. Wang, D. Lederman, J. Tan, D. Gur, Computer-aided detection: the effect of training databases on detection of subtle breast masses, Acad. Radiol. 17 (2010) 1401–1408.
[52] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C.I. Sánchez, R. Mann, A. den Heeten, N. Karssemeijer, Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.
[53] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Comput. Vis. Pattern Recognit. CVPR 2014 IEEE Conf. On, IEEE, 2014: pp. 1717–1724.
[54] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Adv. Neural Inf. Process. Syst. 27, Curran Associates, Inc., 2014: pp. 3320–3328. http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf.
[55] J.E. Joy, E.E. Penhoet, D.B. Petitti, I. of M. (US) and N.R.C. (US) C. on N.A. to E.D. and D. of B. Cancer, Benefits and Limitations of Mammography, National Academies Press (US), 2005. https://www.ncbi.nlm.nih.gov/books/NBK22311/ (accessed April 11, 2018).
[56] Y. LeCun, K. Kavukcuoglu, C. Farabet, Convolutional networks and applications in vision, in: ISCAS 2010 - 2010 IEEE Int. Symp. Circuits Syst. Nano-Bio Circuit Fabr. Syst., 2010. doi:10.1109/ISCAS.2010.5537907.
[57] H. Chougrad, H. Zouaki, O. Alheyane, Convolutional Neural Networks for Breast Cancer Screening: Transfer Learning with Exponential Decay, NIPS-Mach. Learn. Health Workshop. (2017). http://arxiv.org/abs/1711.10752.
[58] D.M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, (2011).
[59] Comparison of the predicted and observed secondary structure of T4 phage lysozyme - ScienceDirect, (n.d.). https://www.sciencedirect.com/science/article/pii/0005279575901099 (accessed April 16, 2018).
[60] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, ArXiv Prepr. ArXiv14091556. (2014).
[61] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.
[62] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat. (1951) 400–407.
[63] Sutton, R. S., Two problems with backpropagation and other steepest-descent learning procedures for networks, in: Proc. Eighth Annu. Conf. Cogn. Sci. Soc., Hillsdale, NJ: Erlbaum, 1986. https://www.bibsonomy.org/bibtex/2c6102b6afaca7b482f471c196364c2be/schaul.
[64] N. Qian, On the momentum term in gradient descent learning algorithms, Neural Netw. 12 (1999) 145–151. doi:10.1016/S0893-6080(98)00116-6.
[65] I. Sutskever, J. Martens, G. Dahl, G. Hinton, On the Importance of Initialization and Momentum in Deep Learning, in: Proc. 30th Int. Conf. Int. Conf. Mach. Learn. - Vol. 28, JMLR.org, Atlanta, GA, USA, 2013: pp. III–1139–III–1147. http://dl.acm.org/citation.cfm?id=3042817.3043064.
[66] L. Bottou, Stochastic gradient learning in neural networks, Proc. Neuro-Nımes. 91 (1991) 0.
[67] L. Bottou, Curiously fast convergence of some stochastic gradient descent algorithms, in: Proc. Symp. Learn. Data Sci. Paris, 2009.
[68] M. Gürbüzbalaban, A. Ozdaglar, P. Parrilo, Why random reshuffling beats stochastic gradient descent, ArXiv Prepr. ArXiv151008560. (2015).
[69] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (2011) 2121–2159.
[70] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, Int. Conf. Learn. Represent. ICLR 2015. (2015). http://arxiv.org/abs/1412.6980.

[71] T. Tieleman, G. Hinton, Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA Neural Netw. Mach. Learn. 4 (2012) 26–31.
[72] N.S. Keskar, R. Socher, Improving Generalization Performance by Switching from Adam to SGD, ArXiv171207628 Cs Math. (2017). http://arxiv.org/abs/1712.07628.
[73] W. Ge, Y. Yu, Borrowing Treasures from the Wealthy: Deep Transfer Learning through Selective Joint Fine-tuning, ArXiv170208690 Cs Stat. (2017). http://arxiv.org/abs/1702.08690.
[74] H.-C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning, ArXiv160203409 Cs. (2016). http://arxiv.org/abs/1602.03409.
[75] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, ArXiv14126980 Cs. (2014). http://arxiv.org/abs/1412.6980.
[76] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Sci. Data. 4 (2017) 170177. doi:10.1038/sdata.2017.177.
[77] M.G. Lopez, N.G. Posada, D.C. Moura, R.R. Pollán, J.M.F. Valiente, C.S. Ortega, M. Solar, G. Diaz-Herrero, I. Ramos, J. Loureiro, BCDR: a breast cancer digital repository, in: 15th Int. Conf. Exp. Mech., 2012.
[78] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, INbreast: Toward a Full-field Digital Mammographic Database, Acad. Radiol. 19 (2012) 236–248. doi:10.1016/j.acra.2011.09.014.
[79] Suckling J. et al., The Mammographic Image Analysis Society Digital Mammogram Database Exerpta Medica., in: 1994: pp. pp375-378.
[80] M. Heath, K. Bowyer, D. Kopans, R. Moore, P. Kegelmeyer, The digital database for screening mammography, Digit. Mammogr. (2000) 431–434.
[81] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, INbreast, Acad. Radiol. 19 (2012) 236–248. doi:10.1016/j.acra.2011.09.014.
[82] Multi-label classification by exploiting local positive and negative pairwise label correlation - ScienceDirect, (n.d.). https://www.sciencedirect.com/science/article/pii/S0925231217301571 (accessed May 11, 2018).
[83] Multi-label learning based on label-specific features and local pairwise label correlation - ScienceDirect, (n.d.). https://www.sciencedirect.com/science/article/pii/S0925231217313462 (accessed May 11, 2018).
[84] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[85] R.-E. Fan, C.-J. Lin, A study on threshold selection for multi-label classification, Dep. Comput. Sci. Natl. Taiwan Univ. (2007) 1–23.
[86] Y. Li, Y. Song, J. Luo, Improving Pairwise Ranking for Multi-label Image Classification, ArXiv170403135 Cs. (2017). http://arxiv.org/abs/1704.03135.
[87] I. Triguero, C. Vens, Labelling strategies for hierarchical multi-label classification techniques, Pattern Recognit. 56 (2016) 170–183. doi:10.1016/j.patcog.2016.02.017.
[88] Y. Yang, A Study of Thresholding Strategies for Text Categorization, in: Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., ACM, New York, NY, USA, 2001: pp. 137–145. doi:10.1145/383952.383975.