Automated Breast Ultrasound Lesions Detection using Convolutional Neural Networks

Moi Hoon Yap, *Member, IEEE*, Gerard Pons, Joan Martí, Sergi Ganau, Melcior Sentís, Reyer Zwiggelaar, Adrian K. Davison, *Member, IEEE*, Robert Martí

***Abstract*—Breast lesion detection using ultrasound imaging is considered an important step of Computer-Aided Diagnosis systems. Over the past decade, researchers have demonstrated the possibilities to automate the initial lesion detection. However, the lack of a common dataset impedes research when comparing the performance of such algorithms. This paper proposes the use of deep learning approaches for breast ultrasound lesion detection and investigates three different methods: a Patch-based LeNet, a U-Net, and a transfer learning approach with a pretrained FCN-AlexNet. Their performance is compared against four state-of-the-art lesion detection algorithms (i.e. Radial Gradient Index, Multifractal Filtering, Rule-based Region Ranking and Deformable Part Models). In addition, this paper compares and contrasts two conventional ultrasound image datasets acquired from two different ultrasound systems. Dataset A comprises 306 (60 malignant and 246 benign) images and Dataset B comprises 163 (53 malignant and 110 benign) images. To overcome the lack of public datasets in this domain, Dataset B will be made available for research purposes. The results demonstrate an overall improvement by the deep learning approaches when assessed on both datasets in terms of True Positive Fraction, False Positives per image, and F-measure.**

***Index Terms*—Lesion detection, ultrasound imaging, breast cancer, convolutional neural networks, transfer learning**

I. Introduction

BREAST cancer is one of the leading causes of death for women worldwide and it is expected that more than 8% of women will develop breast cancer during their lifetime [1]. The most commonly used and effective technique for breast cancer detection is digital mammography (DM) [2]. However, there are some limitations to DM imaging in dense breasts, where lesions have a similar attenuation compared to the dense tissue, and as such they can be hidden by the surrounding tissue. Currently, an important alternative to DM is ultrasound (US) imaging, which is used as a complementary method for breast cancer detection due to its versatility, safety and high sensitivity [3]. However, US imaging depends more on the radiologist than other commonly used techniques such as mammography. Interpreting US images requires experienced and well-trained radiologists due to the complexity and presence of speckle noise. Thus, Computer-Aided Diagnosis (CAD) could be beneficial to help radiologists in the US-based detection of breast cancer, minimizing the effect of the operator-dependent nature of US imaging. Different studies have investigated the influence of CAD on diagnostics [4], [5] and showed that CAD is an important tool to improve the diagnostic sensitivity and specificity. The first challenge in any CAD is the ability to locate the lesion. This process should be automated to help the radiologist make a diagnosis efficiently and a high sensitivity and specificity are expected.

The lack of a public standard dataset in breast US research has limited the fair evaluation of the performance of algorithms. The quality of breast US images is highly dependent on the acquisition process and there is a vast variability between different US systems that influence the results obtained by algorithms. The appearance, location and size of the lesions also affect the results.

In this paper, we review four popular lesion detection methods [6], [7], [8], [9]. We propose the use of deep learning approaches for breast ultrasound lesion detection and investigate three different methods: a Patch-based LeNet, a U-Net, and a transfer learning approach with a pretrained FCN-AlexNet. Then the performances of deep learning approaches are compared with the state-of-the-art algorithms on two breast ultrasound datasets (Dataset A and Dataset B) and make Dataset B available for research purposes. To date, we are the first to conduct this comprehensive comparison on two common datasets and propose the use of deep learning approaches for breast US lesion detection.

II. Related Work

This section describes four state-of-the-art methodologies for lesion detection in breast US imaging. Two of the selected methodologies, Radial Gradient Index (RGI) Filtering [6] and Multifractal Filtering [7], are two of the most cited works in this area. This study also includes two recent approaches, Rule-based Region Ranking [8] and Deformable Part Models [9].

A. Radial Gradient Index (RGI) Filtering

Drukker et al. [6] developed a lesion detection and classification method as a two-stage process. The first stage

M.H. Yap is with the School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, M1 5GD, Manchester, UK. E-mail: m.yap@mmu.ac.uk.

G. Pons at the Universitat Oberta de Catalunya, Barcelona, Spain. Email: gponsro@uoc.edu.

J. Martí and R. Martí at the Department of Computer Science, University of Girona, Spain. Email: robert.marti@udg.edu.

S. Ganau and M. Sentís at Radiology Department, UDIAT-Centre Diagnòstic, Corporació Parc Taulí, Sabadell, Spain. Email: MSentis@tauli.cat.

R. Zwiggelaar at Department of Computer Science, Aberystwyth University, UK. Email: rrz@aber.ac.uk.

A.K. Davison is with the Centre for Imaging Sciences, Faculty of Biology Medicine and Health, University of Manchester. E-mail: adrian.davison@manchester.ac.uk.

Manuscript received xxxx xx, xxxx; revised xxxx xx, xxxx.

was the detection of lesion candidates using a RGI Filtering technique. The second stage was the classification of those candidates, segmenting them by maximising an average radial gradient (ARD) index for regions grown from the detected points and classifying them with a Bayesian neural network as false positives or potential lesions. Here we focus on the performance evaluation of the initial lesion detection stage, thus only the location of lesion candidates is evaluated.

Lesion candidates were identified using a filtering technique based on the calculation of the RGI of contours throughout the image [10]. For a given point (x, y) in the image, lesion-like shapes were obtained by multiplying the image with a 2D isotropic Gaussian function centred at (x, y) to construct a constrained image. Contours of the lesion candidates for a given point were obtained by grey-level thresholding the constrained image. All possible lesion contours within a specified size range were determined, and the RGI value was calculated for each contour as a measure of the likelihood that a given contour represents a lesion.

$$RGI_i(x,y) = \frac{\sum_{(x',y')\in C_i} \vec{g}(x',y') \cdot \hat{r}(x',y')}{\sum_{(x',y')\in C_i} |\vec{g}(x',y')|} \tag{1}$$

where C_i is the i-th possible lesion contour, $\vec{g}(x', y')$ is the maximum grey-value gradient vector of length $|\vec{g}(x', y')|$ and $\hat{r}(x', y')$ the unit radial vector pointing from (x, y) to (x', y').

By definition, due to normalization, RGI values are between 1 (pointing radially outward) and -1 (pointing radially inward). For a given image point (x, y), the contour with the maximum absolute RGI value was selected, and this value was assigned to the (x, y) coordinate in the RGI-filtered image. The RGI-filtered image was subsequently thresholded to determine lesion candidates. The threshold was varied iteratively until either at least one region of interest is detected, indicating a lesion candidate, or the minimum specified RGI threshold value was reached.

B. Multifractals Filtering

The main contribution of the Multifractals Filtering technique lies in the implementation of multifractals analysis in breast US. In 2008, Yap et al. [7] presented a novel initial lesion detection method based on a set of image processing operations. To ensure the homogeneity of the US images, histogram equalisation was first implemented. Then the speckle noise was reduced using a hybrid filtering approach [11]. Hybrid filtering combines the strength of nonlinear diffusion filtering [12] to produce edge-sensitive speckle reduction, followed by linear filtering (Gaussian blur) [13] to smooth the edges and to eliminate oversegmentation. Subsequent to hybrid filtering, multifractals [14] were used to further enhance the partially processed images. Multifractal analysis refers to the analysis of an image using multiple fractals (i.e. not just one as in fractal analysis). The generalized formulation for multifractal dimensions (D) of order q can be represented as:

$$D_q = \begin{cases} \frac{1}{q-1} \lim\limits_{\epsilon\to 0} \frac{\log(x_q(\epsilon))}{\log(\epsilon)} & \text{for } q \in R \text{ and } q \neq 1 \\ \lim\limits_{\epsilon\to 0} \frac{\sum\limits_i \mu_i \log \mu_i}{\log(\epsilon)} & \text{for } q = 1 \end{cases} \tag{2}$$

where ϵ is the linear size of the cells, q is the order for cell size ϵ and μ is the measure defined as the probability of the greyscale level in the images, where all the grey levels fall in the range of (0 - 1). Multifractal analysis enables improved separability of tumour regions from normal regions.

After pre-processing, images were segmented by using a grey-value thresholding segmentation method [15]. This thresholding segmentation often leads to the identification of multiple regions of interest, of which generally only one or two would be of diagnostic importance. To identify these important regions, a rule-based Region of Interest (ROI) selection, based on the size and location of the region was used as a discriminative criterion. Based on the knowledge provided by expert radiologists [16], most of the lesions are located in the upper part of the images. Hence, a reference point (x_r, y_r) where

$$x_r = \frac{\text{image height}}{3}, y_r = \frac{\text{image width}}{2} \tag{3}$$

was chosen, with x_r from the top of the image. The candidate region closest to the (x_r, y_r) location and that satisfied the size-related criterion was selected as the final detected lesion.

C. Rule-based Region Ranking (RBRR)

Shan et al. [8] proposed a lesion detection methodology that considered both texture and spatial features. They first used speckle reducing anisotropic diffusion (SRAD) [17]. The SRAD method processes the image iteratively with adaptable weighted filters to reduce noise and preserve edges. The diffusion coefficient was determined by

$$c(q) = \frac{1}{1 + [q^2(x,y;t) - q_0^2(t)]/[q_0^2(t)(1+q_0^2(t))]} \tag{4}$$

where $q(x, y; t)$ is the instantaneous coefficient of variation depending on gradient ∇I and the Laplacian $\nabla^2 I$ and determined by

$$q(x,y;t) = \sqrt{\frac{(1/2)(|\nabla I|/I)^2 - (1/4)(\nabla^2 I/I)^2}{[1 + (1/4)(\nabla^2 I/I)]^2}} \tag{5}$$

The initialisation $q_0(t)$ is given by

$$q_0(t) = \frac{\sqrt{var[z(t)]}}{z(t)} \tag{6}$$

where t is the iteration time and $z(t)$ is the most homogeneous area in the image at iteration t and $var[z(t)]$ is its variance.

Once the image was de-speckled, an iterative threshold selection algorithm was applied to segment the image. First, all local minima of the image histogram were calculated and the de-speckled image was binarised using the smallest local

minimum as the threshold value. Then, if the ratio of the number of foreground pixels and the number of background pixels was less than 0.1, the next local minimum value was set as the threshold. The process continued iteratively until the ratio was larger than 0.1. This value was chosen experimentally in the original paper [8]. Subsequently, morphological operations (dilation and erosion) were performed to remove noisy regions. If none of the regions intersected with the image centre region (a window about half the size of the entire image and located at the image centre) the threshold became the next local minimum and the process was repeated. Once some region intersects with the central window, regions connected with the boundary that do not intersect with the central window are removed. The remaining lesion region candidates were ranked using the scoring formula

$$S_n = \frac{\sqrt{Area_n}}{dis(C_n, C_0) \cdot var(C_n)}, n = 1, ..., k \tag{7}$$

where k is the number of candidate regions, $Area_n$ is the number of pixels in the region, C_n is the center of the region, C_0 is the center of the image, $dis(a, b)$ is the Euclidean distance between points a and b and $var(C_n)$ is the variance of a small circular region centered at C_n.

Finally, the location of the seed point was located in the centre of the region with a highest score. Thus, $((x_{min} + x_{max})/2, (y_{min} + y_{max})/2)$ was considered as a seed point, where $[x_{min}, y_{min}, x_{max}, y_{max}]$ defined the minimum rectangle that contained the lesion.

D. Deformable Part Models (DPM)

The DPM proposed by Felzenszwalb et al. [18] is one of the effective object detection methods in the recent literature. The work of Pons et al. [9] demonstrated the feasibility of adapting this methodology to detect lesions in breast US images and obtained accurate results. The DPM method modeled the appearance of objects based on a histogram of oriented gradients (HOG) in terms of a low resolution root filter template, which defined the detection window, along with a set of higher resolution part filter templates that captured finer details. Each part defined a set of possible placements for a part relative to the root filter and a deformation cost for each placement.

The system used a scanning window approach that searched a model over a HOG pyramid [19] to detect objects at different scales. The image was divided into a dense grid where the histogram of gradient orientations was computed in each cell and is normalised with respect to the gradient energy in the neighbourhood surrounding it. The HOG pyramid was defined by computing the HOG features of each level of an image pyramid. Hence, features at the top level captured coarse gradients as opposed to finer gradients found at lower levels.

Both root and part filters were rectangular templates F of size $w \times h$ specifying weights for subwindows of a HOG pyramid. In this case, H is a HOG pyramid and $p = (x, y, l)$ a location in the l-th level of that pyramid. The vector obtained by concatenating the HOG features in the $w \times h$ subwindow of H in p was defined as $\phi(H, p)$ and the score of F on this detection window was $F \cdot \phi(H, p)$.

The model for an object with n parts was defined by a root filter F_0 and a set of parts $P_i = (F_i, v_i, d_i)$, where F_i was a filter for the i-th part, v_i was a two-dimensional vector specifying possible locations relative to the root, and d_i was a four-dimensional vector specifying coefficients of a quadratic function that defines a deformation cost for each possible placement of the part.

The placement of the model was given by $z = (p_0, ..., p_n)$ where $p_i = (x_i, y_i, l_i)$ specifies the level and the position of the i-th filter. Note that the location of the root filter was defined when $i = 0$. The final score of a detection was the score of the root filter plus the score of the best location of the parts, placed at twice the resolution in the pyramid, minus a deformation cost that penalises undesired placements of the parts,

$$score(p_0, ..., p_n) = \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i[(\tilde{x}_i, \tilde{y}_i) + (\tilde{x}_i^2, \tilde{y}_i^2)] \tag{8}$$

where $(\tilde{x}_i, \tilde{y}_i) = (x_i, y_i) - (2(x_0, y_0) + v_i)$ gives the displacement of the i-th part relative to the root location and d_i are the deformation features.

The method took advantage of the additional information provided by the part filters. However, these part filters do not need to be labelled (they were considered as latent values). The method described a discriminative training with partially labelled data called a latent Support Vector Machine, which was an iterative training process that alternates between fixing latent values for positive examples and optimizing the latent SVM function (see Felzenszwalb et al. [18] for details).

E. Deep Learning for Breast Imaging

Overall, the state-of-the-art methods are not robust, particularly the image processing based approaches, relying on rule based approaches and specific assumptions. Without needing such strong assumption, deep learning approaches have shown a superior accuracy in object detection, which suggests that could also improve the state of the art of lesion detection in breast ultrasound. Deep learning in medical imaging is mostly represented by convolutional networks. Based on how they are trained, they can be mostly categorized in the following:

- *Patch-based CNNs approach.* This approach trains the convolutional neural networks (CNNs) with image patches for training and a sliding window approach for testing [20], [21]. However, feeding each patch to the network is time-consuming and the patch overlap produces substantial redundancy [22].
- *Fully convolutional approach.* To avoid computational redundancy, Long et al. [38] proposed a fully convolutional approach to increase the efficiency by training on whole images. It produces segmentation by pixelwise prediction rather than single probability distribution in the classification task for each image. An example of a modified version of such approach is U-Net [22].
- *Transfer learning approach.* Another approach that has been widely used recently in biomedical research is the

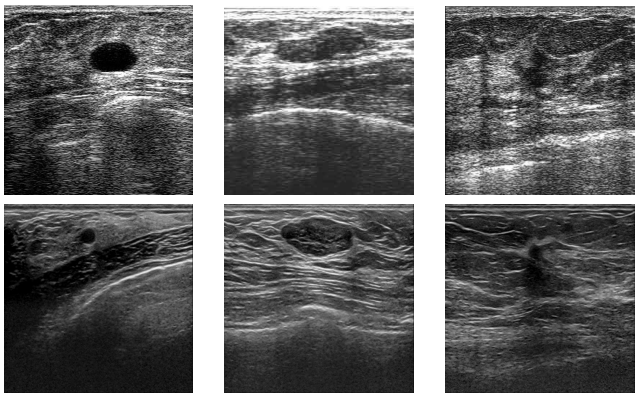
transfer learning approach [23], [24]. This method uses a pre-trained model from non-medical images to overcome the limitation of data deficiency in medical imaging research.

In breast imaging, the majority of the existing publications are focusing on using CNNs for mammography. Dhungel et al. [25] have implemented deep learning for segmentation of masses; Mordang et al. [26] proposed the use of CNNs in microcalcification detection; and more recently, Ahn et al. [27] proposed the use of CNNs in breast density estimation. In breast ultrasound imaging, Huynh et al. [23] proposed the use of a transfer learning approach for ultrasound breast images classification. This is the only work in breast ultrasound but it does not cover lesion detection. In this paper, we propose the use of deep learning approaches for automated breast ultrasound lesions detection. To show the benefits of deep learning approaches, we compare the performances with the four aforementioned (Section II A-D) state-of-the-art lesion detection algorithms.

III. Datasets

A. Overview

This study made use of two different datasets of US images. The datasets were obtained from US systems with different specifications and at different times. They are referred to as Dataset A and B.



(a) (b) (c)

Fig. 1. Examples of images in Dataset A (first row) and Dataset B (second row). (a) shows an example of cyst images, (b) images with fibroadenoma lesion and (c) examples of invasive ductal carcinoma.

Dataset A was collected in 2001 from a professional didactic media file for breast imaging specialists [16]. The images were obtained with B&K Medical Panther 2002 and B&K Medical Hawk 2102 US systems with a 8-12 MHz linear array transducer. The dataset consists of 306 images from different cases with a mean image size of 377×396 pixels. These images contained one or more lesions. Within the lesion images, 60 images presented malignant masses and 246 were benign lesions. From the malignant images, 27 were diagnosed as invasive ductal carcinomas, 4 were ductal carcinomas in situ, 6 were malignant phyllodes tumours and 23 were other unspecified malignant lesions. From the benign images, 74 were complex cysts, 89 were simple cysts, 55 were fibroadenomas and 28 were other benign lesions. To obtain Dataset A, the user needs to purchase the didactic media file from Prapavesis et al. [16].

Dataset B was collected in 2012 from the UDIAT Diagnostic Centre of the Parc Taulí Corporation, Sabadell (Spain) with a Siemens ACUSON Sequoia C512 system 17L5 HD linear array transducer (8.5 MHz). The dataset consists of 163 images from different women with a mean image size of 760×570 pixels, where each of the images presented one or more lesions. Within the 163 lesion images, 53 were images with cancerous masses and 110 with benign lesions. From the malignant images, 40 were invasive ductal carcinomas, 4 were ductal carcinomas in situ, 2 were invasive lobular carcinomas and 7 were other unspecified malignant lesions. From the benign images, 65 were unspecified cysts, 39 were fibroadenomas and 6 were of another type of benign lesions. Note that in both datasets the lesions were delineated by experienced radiologists. Dataset B and the respective delineation of the breast lesions will be available online (goo.gl/SJmoti) for research purposes.

B. Comparison

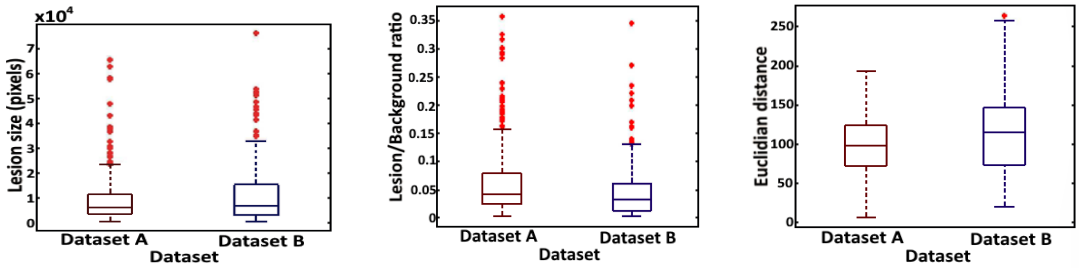
Figure 1 displays three images from each of the two datasets to represent the differences in three aspects: speckle noise, image quality and lesion appearance. In terms of speckle noise, images from Dataset A show a significant presence of this artefact but it is less obvious for images in Dataset B, where the speckle noise was partly reduced by the US acquisition system. The image quality also varies in both datasets due to the different resolutions. Note that the resolution for the recent US device to produce Dataset B is better than in the older US device (Dataset A). Consequently, the defined structures (such as ribs, pectoral muscle or parenchymal tissue) are more visible in Dataset B. The lesion appearance also varies in both datasets. In Dataset B the appearance of tissue is better defined than in Dataset A, as is illustrated in Figure 1(b) where even the inner structures in the fibroadenoma lesion are visible.

To further evaluate the datasets, we compare the lesion size, the ratio between the area of the lesion and the area of the image, and the distance from the image centre and the lesion centroid. Figure 2 shows the box plot charts for these comparisons where differences between both datasets are noticeable: the average size of the lesions in Dataset A is smaller than in Dataset B (Figure 2(a)) but the ratio between lesion pixels and total image pixels is higher (Figure 2(b)). Regarding the spatial distribution of the lesions in the image, lesions in Dataset A are more centred than in Dataset B (Figure 2(c)). However, none of these differences are significant. Furthermore, other characteristics such as the quality of the image may affect the performance of the lesion detection results.

IV. Methodology

A. Convolutional Neural Networks

Deep learning is a representation learning method [28] that will automatically discover features suited for a particular task from the raw data. The feature extractors are task-specific, in



(a) (b) (c)

Fig. 2. Dataset feature comparison. Box plot chart comparing (a) the lesion size, (b) the ratio between the area of the lesion and the area of the image and (c) the distance from the image centre to the lesion centroid.

that they are not fixed to a set of specific rules each time [29]. Each network contains multiple layers that lead to hierarchical features used in the learning process [30], [28].

Convolutional Neural Networks (CNNs) [31] have become an important technique in image analysis, particularly in detection or recognition of faces [32], text [30], human bodies [33] and biological images [34]. However, it has not been used in breast ultrasound lesion detection. For these reasons, we study the performance of CNNs in breast ultrasound lesion detection.

CNNs consist of convolutional layers and pooling layers [31], where the role of the former is to extract local features from a set of learnable filters and the role of the latter is to merge neighbouring patterns, reducing the spatial size of the previous representation and adds spatial invariance to translation [28]. CNNs are hierarchical neural networks and their accuracy is dependent on the design of the layers and training methods [35].

Some popular CNNs available in the Caffe framework [29] are LeNet [30], AlexNet [36] and GoogleNet [37]. We investigated the use of three types of deep learning for breast lesion detection: a patch-based approach using LeNet [30], U-Net [22] and a transfer learning approach using Fully Convolutional Networks [38].

1) Patch-based LeNet: As the ultrasound breast images in the datasets are grayscale and the size of the breast lesions is relatively small, LeNet [30] was chosen as a suitable architecture to solve the two-class classification problem. The training and validation images are input as patches from areas of the images containing abnormal breast lesions and normal tissue. These input patches are sized at 28×28, which correlates to the input size of LeNet. The LeNet architecture is simple and was originally created for digit classification [30]. Breast lesions contain similar gradients that can be exposed through CNNs. The overall architecture can be seen in Figure 3, with the inputs consisting of image patches of breast lesions and normal tissue. The inputs are fed into the first convolution layer and max pooling layer, which is repeated once and finalised with two fully connected layers. The final number of outputs are 2 neurons, which are the activations generated for the two classes: lesion and non-lesion. The final part of the CNN is the output of class probabilities to measure how close the final fully connected parameters are with respect to the ground truth labels of the training and validation data. The loss was calculated using multinomial logistic loss with a softmax classifier.

The output of our network is a prediction of whether the patch is a lesion or healthy breast tissue. It is formed by two fully connected layers with the softmax function defined as

$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \tag{9}$$

where f_j is the j-th element of the vector of class scores f and z is a vector of arbitrary real-valued scores that are squashed to a vector of values between zero and one that sum to one. The loss function is defined so that having good predictions during training is equivalent to having a small loss.

A Rectified Linear Unit (ReLU) layer is included at the first fully connected layer. This element-wise operation is calculated in-place for the Caffe framework [29], and so saves on some memory. It is defined as

$$f(x) = \max(0, x) \tag{10}$$

where the function f thresholds the activations at zero.

Using a sliding window of 28×28 pixels with a stride of 1 for the test images, the predicted lesion patches were segmented. The unconnected regions with an area of less than 10 pixels were removed from the segmented images to reduce False Positives (FPs) through empirical experimentation. The centre points of the segmented regions were recorded as seed points.

2) U-Net: U-Net is a modified and extended version of a fully convolutional network [22], which can overcome the need of large-scale dataset in biomedical imaging research. It is an encoder-decoder based CNN with skip connections. Ronneberger et al. [22] proposed U-Net to enable the use of data augmentation, including the use of non-rigid deformations, to make full use of the available annotated sample images to train the model. These aspects suggest that the U-Net could potentially provide satisfactory results with the size of the available datasets currently used.

3) Transfer Learning: Transfer Learning is a procedure where a CNN is trained to learn features for a broad domain after which the classification function is changed to optimize the network to learn features of a more specific domain. Under this setting, the features and the network parameters are transferred from the broad domain to the specific one. Our proposed transfer learning approach is based on fully

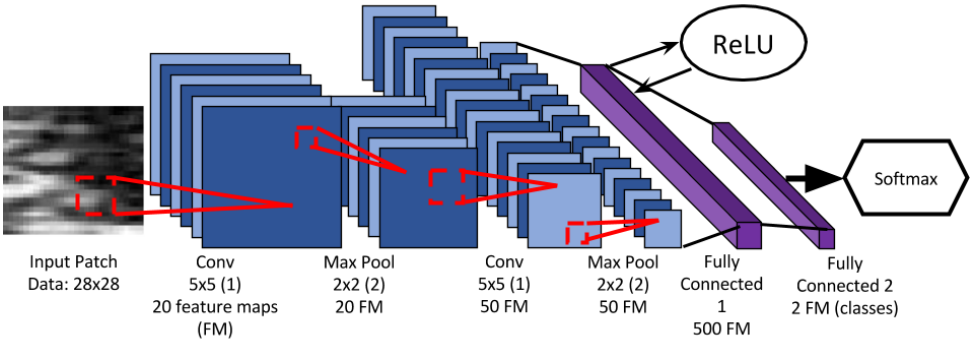


Fig. 3. The overall LeNet architecture. The numbers at the convolution and pooling layers indicate kernel size, stride (in brackets) and the total amount of neurons present at each layer.

convolutional networks (FCN-AlexNet) [38] for semantic segmentation. FCN-AlexNet is a fully convolutional network version of the original AlexNet classification model with a few adjustments of the network layers for segmentation [38]. This network was originally used for the classification of 1000 different objects of classes on the ImageNet dataset [36].

B. Performance Metric

Lesion detection is an initial stage of CAD, which most of the time, uses the detected lesion location as a seed point to subsequently initialise a segmentation algorithm. Most of the breast US lesion detection methodologies in the literature evaluate their algorithms using the seed point detection as a criterion. In current practice, a radiologist annotated a rectangular ROI with four crosses. Based on these four extreme points (top, bottom, left and right), we generated a bounding box as illustrated in Figure 5. Detection is considered as a True Positive (TP) if the detection point (centre of the segmented region) is placed within the bounding box of an expert radiologist. Otherwise, it was considered to be a False Positive (FP).

In this paper, we compare the performance of lesion detection techniques in breast US research by using True Positive Fraction (TPF) and False Positives per image (FPs/image) [6], [7], [8]:

$$TPF = \frac{\text{number of TPs}}{\text{number of actual lesions}} \tag{11}$$

$$FPs/image = \frac{\text{number of FPs}}{\text{number of images}} \tag{12}$$

TPF measures the sensitivity of the method. Some of the algorithms are capable of detecting multiple lesions while some are only capable of detecting a single lesion. The TPF allows a fair measurement as it is measuring the total detected lesions to the total number of actual lesions. Thus, if a method can detect only one lesion in an image with multiple lesions, the TPF of this methodology will be lower than the method that is capable of detecting multiple lesions.

In addition to TPF and FPs/image, the F-measure (the weighted harmonic mean of recall and precision) [39], is computed as:

$$F\text{-}measure = \frac{2 \times \text{TP}}{(2 \times \text{TP}) + \text{FP} + \text{FN}} \tag{13}$$

C. Implementation

It is worth mentioning that the implementation of DPM [9] and Multifractal Filtering [7] were provided by the original authors, while the implementation of the RGI Filtering [4] and RBRR [8] were accurately re-implemented following the description in their respective papers.

To obtain the best performance for the state-of-the art methods on the datasets, we have defined some parameters. For Rule-based Region Ranking, since most of the lesions in [8] appear in the top region of the image, the central window was initialised in the centre-top part of the image. In addition, the iteration time t was set to 50 in the speckle reducing anisotropic diffusion (SRAD) process. In Multifractal Filtering [7], the order was specified as $q = -1$ for the cell size $\epsilon = 3$.

The DPM approach [9] has been trained with a mixture model of 3 components and 8 parts for each root filter. These parameters were chosen in a previous study [40] where different configurations of DPM parameters were assessed in order to obtain the best results in breast US images. For the number of available images, we have configured the training and testing processes as a 10-fold cross-validation. This methodology vastly increases the computation costs in the training stage but allows a more accurate assessment of the methods.

The proposed Patch-based CNNs approach for this study is the LeNet framework [30]. The breast ultrasound images are in grayscale and are split into 28×28 patches. The network is trained by using Root Mean Square Propagation (RMSProp) with a learning rate of 0.01, 60 epochs with the dropout rate of 0.33. The experiment is run using 10-fold cross validation.

For the U-Net implementation, the training data includes the original ultrasound breast images and ground truth training label as shown in Figure 4. We assessed the performance of the model using 10-fold cross validation. The network is trained by using the Adam optimizer [41], with a learning rate of 0.0001 and 300 epochs. The training data for the proposed transfer learning approach for this study was breast ultrasound images and ground truth training label (as illustrated in Figure

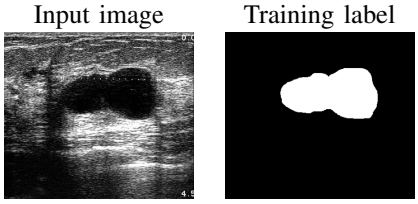


Fig. 4. An example input image for training the network, with the training label used for U-Net and FCN-AlexNet.

4). We used the Caffe [29] framework to implement FCN-AlexNet. We have evaluated the model using 10-fold cross validation. We train the model using stochastic gradient descent with a learning rate of 0.001, 60 epochs with a dropout rate of 33%. The number of epochs was kept at 60 as in [42] as which convergence has already happened when we performed empirical experiment.

V. Results and Discussion

Figure 5 shows the results of breast lesion detection where Row 1 present an image from Dataset A, with a well-defined lesion boundary and a distinct appearance to the normal tissue (intensity values and texture). This is the best case scenario where all the detection methods identified the lesion correctly. Row 2 presents a case from Dataset B where the lesion's appearance is close to the normal tissue and the location where the lesion is close to the top. In this case, only DPM and CNNs detected the lesion correctly. The methodologies that depend on the lesion location have failed to detect the lesion. Row 3 depicts a case from Dataset A where there is a complex shadow in the image. None of the state-of-the-art methods were able to detect the lesion apart from the proposed CNNs. Finally, Row 4 shows a case where none of the methods were able to detect the lesion due to the small lesion size.

Quantitative results are presented in Table I. These are provided in terms of True Positive Fraction (TPF), False Positives per image (FPs/image) and F-measure. When training and testing on a single dataset, the Transfer Learning FCN-AlexNet out-performed other methods for lesion detection, with TPF of 0.98, FPs/image of 0.16 and F-measure of 0.91 for Dataset A; and TPF of 0.92, FPs/image of 0.17 and F-measure of 0.89 for Dataset B. It is observed that the performance of U-Net is lower than Patch-based LeNet. DPM achieved good results in TPF, with 0.80 for Dataset A and 0.79 for Dataset B and with a comparable F-measure to CNNs. Deep learning approaches and DPM achieved low FPs/image. The Multifractal Filtering [7] and RBRR [8] obtained good results for the images in Dataset A, with TPF of 0.76 and 0.75 respectively, but not for the images in Dataset B (with TPF of 0.59 and 0.60, respectively). The average FPs/image for Multifractal Filtering is lower than the RBRR. Finally, the RGI Filtering [6] showed a good performance in terms of TPF in both datasets (0.76 and 0.72) but with a high FPs/image and poor F-measure.

Methods based on image processing (RGI Filtering [6], Multifractal Filtering [7] and Rule-based Region Ranking [8]) were inconsistent and obtained poor results when dealing with images acquired from two different US systems. One explanation is that most of the approaches take the characteristics of their datasets into consideration, such as the lesion location, the influence of the speckle noise or the appearance of the lesions. These characteristics may differ in another dataset, which reduce the accuracy of the algorithms.

TABLE I
Comparison of performance for different methods when training and testing on single dataset. The method LeNet represents the Patch-based LeNet and FCN-AlexNet represents Transfer Learning FCN-AlexNet. Bold indicates the best results when training and testing on a single dataset.

Method	Dataset	TPF	FPs/image	F-measure
RGI [6]	A	0.76	1.57	0.46
	B	0.72	2.47	0.34
Multifractal [7]	A	0.76	0.31	0.74
	B	0.59	0.51	0.56
RBRR [8]	A	0.75	0.50	0.67
	B	0.60	0.54	0.56
DPM [9]	A	0.80	0.20	0.80
	B	0.79	0.21	0.79
LeNet	A	0.89	**0.10**	0.88
	B	0.85	**0.14**	0.86
U-Net	A	0.91	0.21	0.86
	B	0.77	0.28	0.75
FCN-AlexNet	A	**0.98**	0.16	**0.91**
	B	**0.92**	0.17	**0.89**

Dataset B was acquired from a modern US system, which introduces new challenges for the existing techniques in lesion detection. These US systems acquire high-resolution images which may include other structures such as ribs, pectoral muscle or the air in the lungs making the lesion detection more difficult. Dataset A was obtained from an older US system. The nature of the images is normally of a lower resolution and with a higher noise level. For a better visualisation, the radiologist tends to place the suspected lesion at the centre of the image. Nowadays, with high quality US systems this is no longer necessary due to the fact that one image can capture larger regions of the breast. Hence, methodologies that assume that the lesion is centred in the image fail in more cases when using the modern US systems.

The techniques with better results in breast lesion detection are the machine learning and deep learning approaches, where the Transfer Learning FCN-AlexNet performed the best overall. This is due to the fact that these approaches adopt a training process, which helps the method to build a particular model of each dataset. The training stage mimics an adaptation process for different datasets. Thus, it is not as dataset-dependent as other methodologies. However, this methodology contains some drawbacks. The main drawback is the training process, which is time consuming and requires a representative set of normal images. The acquisition of these images in an ultrasonic examination is not common practice in clinical environments.

To investigate the robustness of deep learning approaches on different datasets, we conducted an experiment by combining the two datasets (A+B) - this formed a total of 356 benign lesions and 113 malignant lesions. By using the similar settings as outlined in the methodology, the results are shown

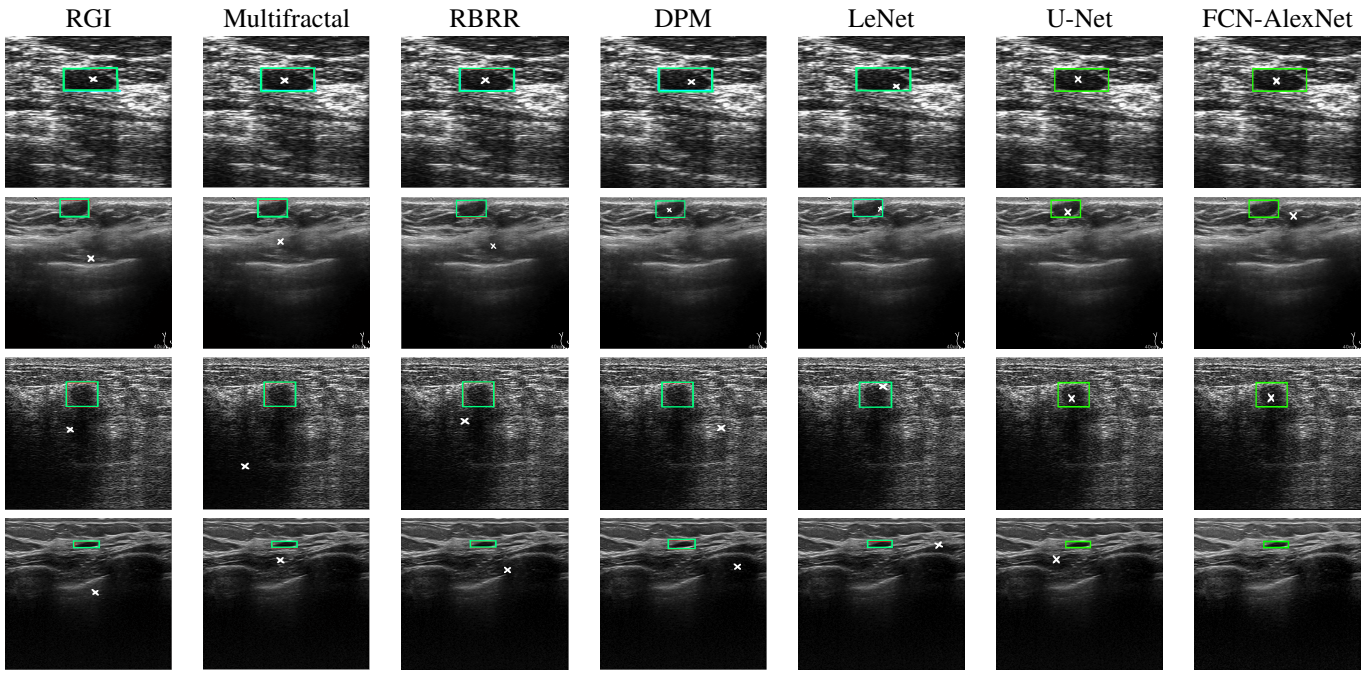


Fig. 5. Examples cases from Dataset A and B to illustrate the performance of the lesion detection algorithms. The rectangle indicates the ground truth and the crosses are the detected abnormality. The first row (image from Dataset A) shows an easy case where all methods detected the lesion. The second row (image from Dataset B) illustrate a case where the lesion is located close to the top and only DPM, Patch-based LeNet and U-Net detected the lesion. The third row (image from Dataset A) shows an image with complex shadow and only the proposed deep learning approaches detected the lesion. The fourth row (image from Dataset B) shows an image with a very small region where none of the methods detect the lesion, and only the FCN-AlexNet has no false positive.

TABLE II
COMPARISON OF THE PERFORMANCE OF THE PROPOSED DEEP LEARNING APPROACHES ON THE COMBINED DATASET. THE METHOD LENET REPRESENTS THE PATCH-BASED LENET AND FCN-ALEXNET REPRESENTS TRANSFER LEARNING FCN-ALEXNET. BOLD INDICATES THE BEST RESULTS TRAINING AND TESTING ON THE COMBINED DATASETS.

Method	Dataset	TPF	FPs/image	F-measure
LeNet (A+B)	A	0.92	**0.07**	0.91
	B	0.91	**0.09**	**0.91**
U-Net (A+B)	A	0.94	0.18	0.89
	B	0.80	0.27	0.78
FCN-AlexNet (A+B)	A	**0.99**	0.16	**0.92**
	B	**0.93**	0.18	0.88

in the final three rows of Table II - with (A+B). Overall, Transfer Learning FCN-AlexNet performed best for Dataset A with a slight improvement on TPF of 0.99, FPs/image of 0.16 (unchanged) and F-measure of 0.92. For Dataset B, the best TPF was 0.93, achieved by Transfer Learning FCN-AlexNet, but the overall best result was Patch-based LeNet with FPs/image of 0.09 and F-measure of 0.91. These results indicated that the supervised deep learning approaches were data-driven and the performance improved with more training data. For many deep learning applications, there is a requirement for large amounts of representative training and testing data to be collected to achieve high accuracies [43].

We have explored the possibility to train on one dataset and test on the other. When training on Dataset B and testing on Dataset A using U-Net, the result dropped to a TPR of 0.83, FP/Image of 0.08 and F-measure of 0.87. When training on Dataset A and test on Dataset B, the result was 0.70 TPR, 0.66 FP/image and 0.59 F-measure. This experiment shows that it is not ideal to train on one dataset different from the testing set. Combining the datasets provides improved training for the framework.

VI. CONCLUSION

This paper investigated the use of three deep learning approaches (Patch-based LeNet, U-Net, Transfer Learning FCN-AlexNet) and a comprehensive evaluation of the most representative lesion detection methodologies for breast ultrasound lesion detection. The performances were evaluated on two datasets in terms of TPF, FPs/image and F-measure.

Amongst the different methodologies discussed in this paper, the Transfer Learning FCN-AlexNet achieved the best results for Dataset A and the proposed Patch-based LeNet obtained the best results for Dataset B in terms of FPs/image and F-measure. DPM and deep learning methods are adaptable to the specific characteristics of any dataset, since these are machine-learning based and a particular model is constructed for each dataset. However, the limitation of such methods is that they require a training process and negative images in the experiment. For further research, it is our assertion that deep learning approaches could be adapted to other medical imaging techniques such as 3 dimensional ultrasound or elastography.

Lesion detection is the initial step of a CAD system. Hence, future work will focus on increasing the accuracy by adding more training data, extending our works to breast ultrasound lesion segmentation and classification, and evaluate the performance of the complete CAD framework.

REFERENCES

[1] H. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognition*, vol. 43, no. 1, pp. 299 – 317, 2010.

[2] O. Akin, S. Brennan, D. Dershaw, M. Ginsberg, M. Gollub, H. Schöder, D. Panicek, and H. Hricak, "Advances in oncologic imaging: Update on 5 common cancers," *CA Cancer Journal for Clinicians*, vol. 62, no. 6, pp. 364–393, 2012.

[3] A. Stavros, D. Thickman, C. Rapp, M. Dennis, S. Parker, and G. Sisney, "Solid breast nodules: Use of sonography to distinguish between benign and malignant lesions," *Radiology*, vol. 196, no. 1, pp. 123–134, 1995.

[4] K. Drukker, N. P. Gruszauskas, C. A. Sennett, and M. L. Giger, "Breast US computer-aided diagnosis workstation: Performance with a large clinical diagnostic population," *Radiology*, vol. 248, no. 2, pp. 392–397, 2008.

[5] M. H. Yap, E. Edirisinghe, and H. Bez, "Processed images in human perception: A case study in ultrasound breast imaging," *European Journal of Radiology*, vol. 73, no. 3, pp. 682–687, 2010.

[6] K. Drukker, M. L. Giger, K. Horsch, M. A. Kupinski, C. J. Vyborny, and E. B. Mendelson, "Computerized lesion detection on breast ultrasound," *Medical Physics*, vol. 29, no. 7, pp. 1438–1446, 2002.

[7] M. H. Yap, E. A. Edirisinghe, and H. E. Bez, "A novel algorithm for initial lesion detection in ultrasound breast images," *Journal of Applied Clinical Medical Physics*, vol. 9, no. 4, pp. 181–199, 2008.

[8] J. Shan, H. Cheng, and Y. Wang, "Completely automated segmentation approach for breast ultrasound images using multiple-domain features," *Ultrasound in Medicine and Biology*, vol. 38, no. 2, pp. 262–275, 2012.

[9] G. Pons, R. Martí, S. Ganau, M. Sentis, and J. Martí, "A feasibility study of lesion detection using deformable part model in breast ultrasound images," in *Iberian Conference on Pattern Recognition and Image Analysis*, vol. 7887, 2013, pp. 269–276.

[10] M. Kupinski, M. Giger, and A. Baehr, "Computerized detection of mass lesions in digital mammography using radial gradient index filtering," in *Radiology*, vol. 213, 1999, pp. 229–229.

[11] M. H. Yap, E. A. Edirisinghe, and H. E. Bez, "Object boundary detection in ultrasound images," in *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*. IEEE, 2006, pp. 53–53.

[12] J. Weickert, "Nonlinear diffusion filtering." in *Handbook on Computer Vision and Applications*, vol. 2, 1999, pp. 423–450.

[13] D. R. Chen, R. F. Chang, W. J. Wu, W. K. Moon, and W. L. Wu, "3-d breast ultrasound segmentation using active contour model," *Ultrasound in Medicine and Biology*, vol. 29, no. 7, pp. 1017–1026, 2003.

[14] D. Gan and Y. Soon, "A multifractal approach for auto-segmentation of SAR images," in *Geoscience and Remote Sensing Symposium, 2001. IGARSS '01. IEEE 2001 International*, vol. 5, 2001, pp. 2301–2303.

[15] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[16] S. Prapavesis, B. Fornage, A. Palko, C. Weismann, and P. Zoumpoulis, *Breast Ultrasound and US-Guided Interventional Techniques: A Multimedia Teaching File*. Thessaloniki, Greece, 2003.

[17] Y. Yu and S. Acton, "Speckle reducing anisotropic diffusion," *IEEE Transactions on Image Processing*, vol. 11, no. 11, pp. 1260–1270, 2002.

[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 886–893.

[20] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems*, 2012, pp. 2843–2851.

[21] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical image analysis*, vol. 35, pp. 303–312, 2017.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[23] B. Huynh, K. Drukker, and M. Giger, "Mo-de-207b-06: Computer-aided diagnosis of breast ultrasound images using transfer learning from deep convolutional neural networks," *Medical Physics*, vol. 43, no. 6, pp. 3705–3705, 2016.

[24] H. Ravishankar, P. Sudhakar, R. Venkataramani, S. Thiruvenkadam, P. Annangi, N. Babu, and V. Vaidya, "Understanding the mechanisms of deep transfer learning for medical images," in *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2016, pp. 188–196.

[25] N. Dhungel, G. Carneiro, and A. P. Bradley, "Deep learning and structured prediction for the segmentation of mass in mammograms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 605–612.

[26] J.-J. Mordang, T. Janssen, A. Bria, T. Kooi, A. Gubern-Mérida, and N. Karssemeijer, "Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks," in *International Workshop on Digital Mammography*. Springer, 2016, pp. 35–42.

[27] C. K. Ahn, C. Heo, H. Jin, and J. H. Kim, "A novel deep learning-based approach to high accuracy breast density estimation in digital mammography," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2017, pp. 101 342O–101 342O.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[29] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[31] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The Handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10, p. 1995, 1995.

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[33] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3626–3633.

[34] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. E. Barbano, "Toward automatic phenotyping of developing embryos from videos," *IEEE Transactions on Image Processing*, vol. 14, no. 9, pp. 1360–1371, 2005.

[35] D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 1. Barcelona, Spain, 2011, p. 1237.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[39] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, pp. 37–63, 2011.

[40] G. Pons, R. Martí, S. Ganau, M. Sentís, and J. Martí, "Computerized detection of breast lesions using deformable part models in ultrasound images," *Ultrasound in Medicine & Biology*, vol. 40, no. 9, pp. 2252–2264, 2014.

[41] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[42] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[43] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3-4, pp. 197–387, 2014.