FULLY AUTOMATED CLASSIFICATION OF MAMMOGRAMS USING DEEP RESIDUAL NEURAL NETWORKS

Neeraj Dhungel[‡] *Gustavo Carneiro*[†] *Andrew P. Bradley*[⋆] [*]

[‡]Electrical and Computer Engineering, The University of British Columbia, Canada
[†]Australian Centre for Visual Technologies, The University of Adelaide, Australia
[⋆] School of ITEE, The University of Queensland, Australia

ABSTRACT

In this paper, we propose a multi-view deep residual neural network (mResNet) for the fully automated classification of mammograms as either malignant or normal/benign. Specifically, our mResNet approach consists of an ensemble of deep residual networks (ResNet), which have six input images, including the unregistered craniocaudal (CC) and mediolateral oblique (MLO) mammogram views as well as the automatically produced binary segmentation maps of the masses and micro-calcifications in each view. We then form the mResNet by concatenating the outputs of each ResNet at the second to last layer, followed by a final, fully connected, layer. The resulting mResNet is trained in an end-to-end fashion to produce a case-based mammogram classifier that has the potential to be used in breast screening programs. We empirically show on the publicly available INbreast dataset, that the proposed mResNet classifies mammograms into malignant or normal/benign with an AUC of 0.8.

Index Terms— Mammogram, Classification, Multi-view, Residual neural network

1. INTRODUCTION

Breast cancer is the most commonly detected cancer amongst women worldwide, registering 23% of all diagnosed cancers [1]. Breast screening programs utilise mammograms to detect the initial signs of breast cancer, making the treatment process more effective and efficient [2]. Breast screening with mammograms is usually carried out using images of both breasts taken from the mediolateral oblique (MLO) and craniocaudal (CC) views. The analysis of mammograms from these views is carried out by detecting the markers of breast lesions such as masses and micro-calcifications (μCs) [3, 4], whose shape and appearance help radiologists characterise them as either normal/benign or malignant. Breast masses are typically dense and so have the characteristic of being grey to white in pixel intensity. Geometrically they can be oval, irregular or lobulated with spiculated, circumscribed, obscured or ill defined margins [5, 6]. Micro-calcifications are small round dense (bright) regions in the breast tissue [5, 6]. In general, a mass like lesion is considered to be malignant if its shape is irregular or spiculated, and clusters of μCs around certain locations in the breast can also be a sign of malignancy [5, 4].

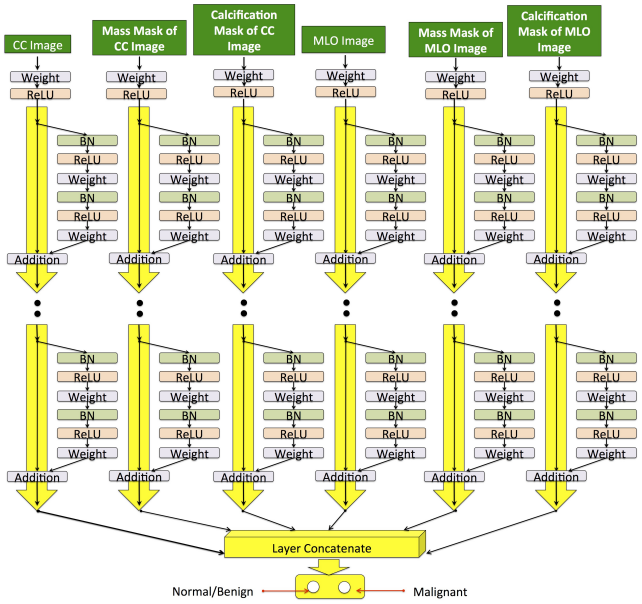

Fig. 1. The multi-view deep residual network (mResNet) for fully automated classification of mammograms from CC and MLO views and their automatically generated mass and micro-calcification segmentation masks.

The manual breast screening process is tedious and time consuming as radiologists have to examine a large volume of mammograms as all women between the ages of 45 and 65 are advised to be screened annually [7]. This volume of cases can degrade the performance of manual interpretation resulting in unnecessary secondary imaging and/or breast biopsies [7]. In fact, it has been reported that the sensitivity of manual screening programs fluctuates between 80% and 90% with a specificity of around 91% [8]. Therefore, double reading of mammograms is often recommended, which has been shown to increase the sensitivity by 9% and decrease the recall rate by 45% [9]. In this scenario, it has been shown that computer aided diagnostic (CAD) systems can also improve the perfor-

[*]Supported by the Australian Research Council Discovery Project (DP140102794). We also thank Nvidia for the TitanX provided for running the experiments in this research paper.

mance of the mammographic screening process [4].

The automated classification of mammograms is generally carried out by first detecting breast lesions (μCs and masses), followed by a second stage which classifies the lesions [5, 10, 4]. The second stage of lesion classification proceeds by extracting hand crafted intensity, morphological and texture features from each lesion and then using these as inputs to a machine learning classifier [5, 4]. One of the major drawbacks of this approach is that the design of the features and the classifier is performed separately, producing sub-optimal results. In contrast, we propose a joint learning of features and classifier using deep residual networks. Moreover, current methods focus on the classification of each individual mass and cluster of μCs rather than the classification of the mammogram as a whole. In this paper, we propose the classification of whole mammogram exams, including all views and segmentation maps of masses and μCs. In addition, results are often reported on private datasets making reproducibility and comparison difficult. Here, we use the INbreast dataset [11] which is publicly available and contains high quality full field digital mammogram (FFDM) images with accurate annotations and has previously been used as a baseline dataset [12].

Deep learning models have produced state-of-the-art results in many computer vision applications [13, 14] and also in applications related to the analysis of mammograms, such as mass segmentation [12], mass detection [15] and mammogram classification [10]. The reason behind the success of the deep learning models lies in their ability to learn and integrate low-, mid- and high-level features by stacking hidden layers in the network architecture [16]. However, networks with a larger number of layers are not easily trained because the gradients required for back propagating the errors during training either vanish (to zero) or explode (to infinity) which adversely affects convergence [17, 14]. This problem has been recently addressed with residual learning, where the layers are reformulated for learning the residual function with respect to each layer's input [14].

The aim of this paper is to present a novel approach for the fully automated classification of mammograms using deep residual neural networks [14]. This work is an extension of [10], where a multi-view mammogram classifier was developed using deep convolutional neural networks (CNN), but with manually defined mass and μC segmentation maps and pre-trained on computer vision datasets. There are three issues with this approach, 1) it is not fully automated as it requires manual detection of lesions, 2) it needs to be pre-trained on computer vision datasets, and 3) it is trained greedily for each input (images and segmentation maps), before the model is trained jointly. We address these issues with the use of automated mass [15] and μCs [18] detection methods and then using this information to train a multi-view deep residual network (mResNet) in an optimal end-to-end fashion (without pre-training). We show, on the INbreast dataset, that our proposed mResNet system classifies full mammogram exams into normal/benign or malignant with an area under the ROC curve (AUC) of 0.8. This result shows that our proposed mResNet has the potential to be used in breast screening programs.

2. METHODOLOGY

2.1. Dataset

Let $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{m}^{(i)}, \mathbf{c}^{(i)}, y^{(i)})_j\}_{j=1}^{|\mathcal{D}|}$ represent the dataset, where $i \in \{\text{left}, \text{right}\}$ indexes the patient's left and right breast denoting an individual case, $\mathbf{x} = \{\mathbf{x}_{\text{cc}}, \mathbf{x}_{\text{ml}}\}$ are two views (CC and MLO) such that $\mathbf{x}_{\text{cc}}, \mathbf{x}_{\text{ml}} : \Omega \rightarrow \mathbb{R}$ with $\Omega \in \mathbb{R}^2$, $\mathbf{m} = \{\mathbf{m}_{\text{cc}}, \mathbf{m}_{\text{ml}}\}$ represents the segmentation of masses in each view with $\mathbf{m}_{\text{cc}}, \mathbf{m}_{\text{ml}} : \Omega \rightarrow \{0, 1\}$, $\mathbf{c} = \{\mathbf{c}_{\text{cc}}, \mathbf{c}_{\text{ml}}\}$ represents the segmentation of μCs in each view such that $\mathbf{c}_{\text{cc}}, \mathbf{c}_{\text{ml}} : \Omega \rightarrow \{0, 1\}$, $y \in \{0, 1\}$ denotes the class label of the mammogram that can be either normal/benign (i.e., BI-RADS $\in \{1, 2, 3\}$) or malignant (i.e., BI-RADS $\in \{4, 5, 6\}$).

2.2. Multi-view Residual Network (mResNet)

A deep residual network (ResNet) consists of multiple stacks of residual units. Each residual unit can be expressed by [19]:

$$\mathbf{x}_{l+1} = h(\mathbf{x}_l) + f_{\text{RES}}(\mathbf{x}_l; \mathcal{W}_l), \tag{1}$$

where $\mathbf{x}_l$ is the input feature to the $l^{th} \in \{1, ..., L\}$ residual unit, $\mathcal{W}_l = \mathbf{w}_{l,k}$ is the set of weights for the l^{th} residual unit, with $k \in \{1, ..., K\}$ representing the numbers of layers in that residual unit, $f_{\text{RES}}(.)$ is called the residual function represented by a convolutional layer (weight) [13, 20], a batch normalisation (BN) [21] and a rectilinear unit (ReLU) [22], and $h(\mathbf{x}_l) = \mathbf{x}_l$ is an identity mapping [14, 19]. In general, the output at the location L within the deep residual net can be obtained recursively using (1) as:

$$\mathbf{x}_L = \mathbf{x}_l + \sum_{l=1}^{L-1} f_{\text{RES}}(\mathbf{x}_l; \mathcal{W}_l). \tag{2}$$

Our proposed multi-view residual network (mResNet), as shown the Fig. 1, can be thought of as an ensemble of individual ResNets, where we concatenate the output from the last layer of all individual ResNets, which is then followed by a final, fully connected layer that can be expressed as follows:

$$\tilde{y} = f_{\text{mRES}}(\mathbf{x}_{cc,L}, \mathbf{x}_{\text{ml},L}, \mathbf{m}_{\text{cc},L}, \mathbf{m}_{\text{ml},L}, \mathbf{c}_{\text{cc},L}, \mathbf{c}_{\text{ml},L}; \mathcal{W}_{\text{mRES}}), \tag{3}$$

where function $f_{\text{mRES}}(.)$ concatenates the outputs from an individual incoming ResNet for each view plus their segmentation masks (for both masses and μCs) which are then passed to the final fully connected layer containing two nodes, one denoting normal/benign and the other malignant. The outputs from the last layer of each individual ResNet are denoted by:

$$\begin{aligned}
\mathbf{x}_{\text{cc},L} &= \mathbf{x}_{\text{cc},l} + \sum_{l=1}^{L-1} f_{\text{RES}}(\mathbf{x}_{\text{cc},l}; \mathcal{W}_{\text{xcc},l}), \\
\mathbf{x}_{\text{ml},L} &= \mathbf{x}_{\text{ml},l} + \sum_{l=1}^{L-1} f_{\text{RES}}(\mathbf{x}_{\text{ml},l}; \mathcal{W}_{\text{xml},l}), \\
\mathbf{m}_{\text{cc},L} &= \mathbf{m}_{\text{cc},l} + \sum_{l=1}^{L-1} f_{\text{RES}}(\mathbf{m}_{\text{cc},l}; \mathcal{W}_{\text{mcc},l}),
\end{aligned} \tag{4}$$

$$\mathbf{m}_{\text{ml},L} = \mathbf{m}_{\text{ml},l} + \sum_{l=1}^{L-1} f_{\text{RES}}(\mathbf{m}_{\text{ml},l}; \mathcal{W}_{\text{mml},l}),$$
$$\mathbf{c}_{\text{cc},L} = \mathbf{c}_{\text{cc},l} + \sum_{l=1}^{L-1} f_{\text{RES}}(\mathbf{c}_{\text{cc},l}; \mathcal{W}_{\text{ccc},l}) \quad (5)$$
$$\mathbf{c}_{\text{ml},L} = \mathbf{c}_{\text{ml},l} + \sum_{l=1}^{L-1} f_{\text{RES}}(\mathbf{c}_{\text{ml},l}; \mathcal{W}_{\text{cml},l}),$$

where $\mathcal{W}_{\text{mRES}} = [\mathbf{w}_{\text{fc}}, \mathcal{W}_{\text{xcc},l}, \mathcal{W}_{\text{xml},l}, \mathcal{W}_{\text{mcc},l}, \mathcal{W}_{\text{mml},l}, \mathcal{W}_{\text{ccc},l}, \mathcal{W}_{\text{cml},l}]$ represents the weights of the mResNet, with $\mathbf{w}_{\text{fc}}$ denoting the weights of the fully connected final layer, $\mathcal{W}_{\text{xcc},l}$ the weights of the CC image, $\mathcal{W}_{\text{xml},l}$ the weights of the MLO image, $\mathcal{W}_{\text{xml,l}}$ the weights of the MLO image, $\mathcal{W}_{\text{mcc},l}$ the weights of the mass mask from the CC image, $\mathcal{W}_{\text{mml},l}$ the weights of the mass mask from the MLO image, $\mathcal{W}_{\text{ccc},l}$ the weights of the μC mask from the CC image, and $\mathcal{W}_{\text{cml},l}$ the weights of the μCs mask from the MLO image.

The training of mResNet is done in an end to end fashion using stochastic gradient decent to minimise the following cross entropy loss:

$$\ell(\mathcal{W}_{\text{mRES}}) = \sum_{j=1}^{|\mathcal{D}|} \sum_{i \in \{\text{left,right}\}} y_{(i,j)} \log \tilde{y}_{(i,j)}. \quad (6)$$

Finally, inference in a ResNet is done in a purely feed-forward direction.

2.3. Automated Lesion Detection and Segmentation

The automated mass detection method used here is based on a deep learning method proposed by Dhungel et al. [15]. The detection consists of a pixel-wise classification over an image grid using input regions of a fixed size at various scales with a multi-scale deep belief network (m-DBN) classifier [15]. This is then followed by a false positive reduction stage using a cascade of deep convolutional neural networks (CNNs) [15, 13] and random forest classifiers [23]. Similarly, our automated μC detection is based on the methodology proposed by Lu et al. [18], which uses both shape and appearance features and a cascade of boosting classifiers. We use these methods given their state-of-the-art performance in automated mass and μC detection.

3. EXPERIMENTS

We carried out experiments using the publicly available INbreast dataset [11], which comprises of 116 cases containing 410 images. Experiments were run using five fold cross-validation by randomly dividing the cases into mutually exclusive subsets, such that 60% of the cases were available for training, 20% for validation and 20% for testing. The automated set-up for mass [15] and μC [18] detection was done by selecting a fixed threshold from the free response operating characteristic (FROC) curve that limits the false positives per image (FPI) to FPI ≈ 1 on the validation set, which produces a true positive detection rate (TPR) for μCs of around 40% and for masses of around 96%. The resulting binary maps of the masses and μCs were resized to 120×120 pixels using nearest neighbour interpolation, whereas the CC and MLO images of the same breast were resized to 120×120 pixels using bi-cubic interpolation and then contrast normalised, as described in [24]. In this way, the mResNet model, shown in Fig. 1, was given six inputs: CC image, MLO image, binary maps of detected masses in CC and MLO and binary maps of detected μCs in CC and MLO. Each input was passed through the convolutional layer (weights) plus a ReLU, where the convolutional layer contains eight filters of size 3×3 followed by nine subsequent residual units. Each residual units was made up of batch normalisation (BN) plus ReLU plus weights. Each convolutional layer in the first three residual units contained the same eight filters (size 3×3), the fourth, fifth and sixth residual units contained 16 filters of size 3×3 and the seventh, eighth and ninth units had 32 filters of size 3×3. In the second to last layer, we concatenated the 32 output features from each ResNet to form 192 features (32×6), followed by a fully connected layer containing two nodes (normal/benign and malignant). For comparison, we also used an mResNet with the same network structure, but with only two inputs: the CC and MLO images. All of our experiments were performed on a computer with an Intel(R) Core(TM) i7-2600k 3.40GHz×8 CPU with 16GB RAM and graphics card NVIDIA GeForce TITANX.

4. RESULTS AND DISCUSSION

Fig.2(a-c) shows the ROC curves generated by the mResNet based on the following input images: a) CC and MLO view plus manually detected lesions, b) CC and MLO views plus automatically detected lesions and c) CC and MLO views only. The AUC values for these curves are 0.91 ± 0.03, 0.80 ± 0.04 and 0.74 ± 0.02 respectively. A paired Wilcoxon signed-rank test indicates that the mResNet using the CC, MLO views with automatically detected lesions has a significantly larger AUC than the mResNet based on only the CC and MLO views ($p \leq 0.03$). The mResNet with manually detected lesions produces an equivalent AUC of 0.91 compared to the previous (baseline) method [10] which also utilised manually detected lesions. However, the advantage of our method lies in the fact that we train the whole mResNet in a single pass, which is more robust compared to the greedy training process utilised in the baseline method. In addition, as mResNet has a deeper architecture, containing 392 layers compared to 61 layers in the baseline method [10], it has the potential to learn higher level representations of the data. However, this advantage may only come to the fore when a larger training set is available.

The fall in AUC, from 0.91 to 0.8, when automated lesion detection is performed indicates the importance that the false positives generated by the automated lesion detection algorithms have on the classification results. Here, we selected an operating threshold from the FROC curve (on the validation set) so as to maintain FPI≤ 1, with μC TPR of around 40% and mass TPR of around 96%, as mentioned above. These results indicate that precise detection and segmentation of masses and μCs is important to allow for a more precise mammogram classification. Furthermore, the AUC of

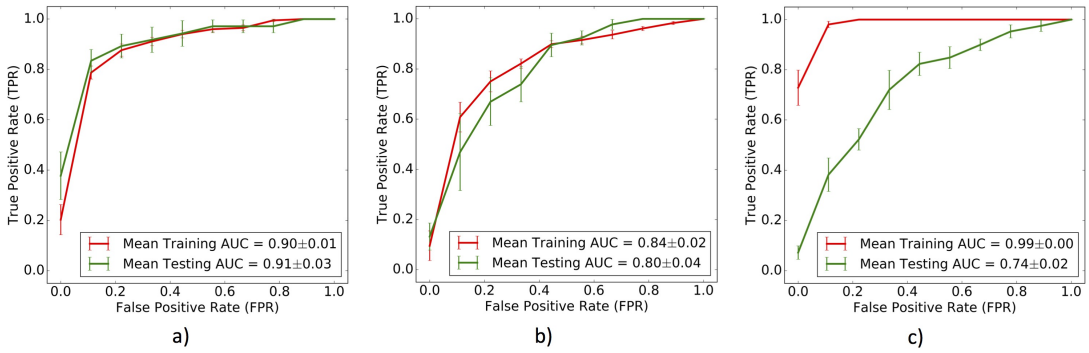

Fig. 2. ROC curves for the mResNet classifier a) CC+MLO with manual lesion detection, b) CC+MLO with automated lesion detection and c) CC+ MLO images only.

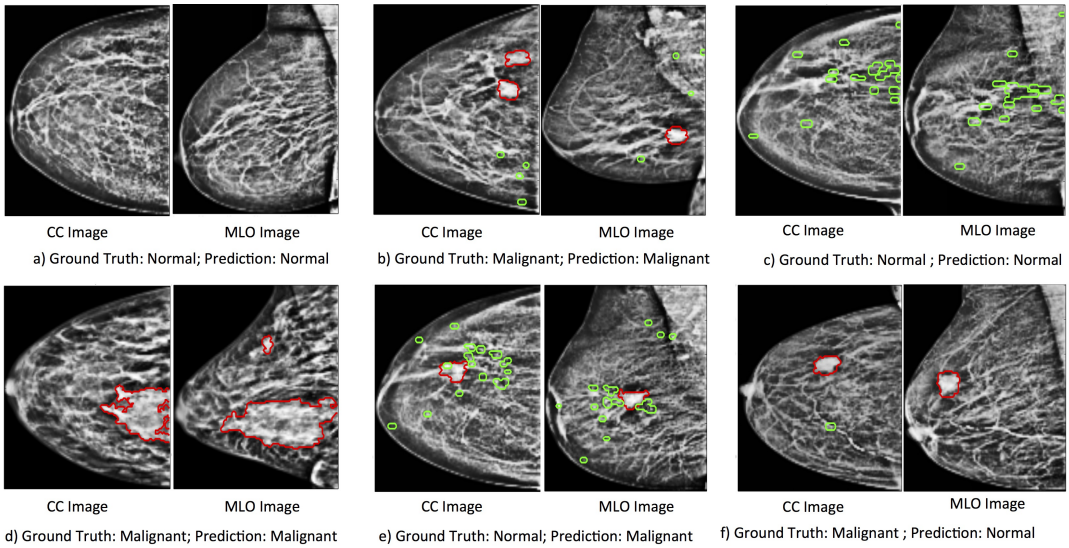

Fig. 3. Examples of classification results of mResNet on the test set. The red and green contours denote automatically detected masses and μCs respectively.

0.74 when only CC and MLO images are used suggests that the use of the masses and μC segmentation maps is important to achieve accurate mammogram classification.

Fig. 3 shows a selection of visual results from the mResNet mammogram classifier along with fully automated lesion detection and segmentation. In particular, Fig. 3(a-e) shows classifications from the system, in the presence of a) no lesions in either view, b) masses (red contour) and micro-calcifications (green contour) in both views, c) micro-calcifications (only) in both views, d) masses (only) in both views and (e-f) cases that the system fails to classify correctly.

5. CONCLUSIONS

In this paper, we have proposed a mResNet that fully automates the classification of mammograms based on information from the CC and MLO views, and associated automatically detected lesions. On the public INbreast dataset, we show that the combination of both views with the automatically generated lesion segmentation masks produces a reasonably accurate classification into malignant or normal/benign, with an AUC of 0.8. This result shows that our proposed mResNet has the potential to be used in breast screening programs.

6. REFERENCES

[1] Ahmedin Jemal, Rebecca Siegel, Elizabeth Ward, Yongping Hao, Jiaquan Xu, Taylor Murray, and Michael J Thun, "Cancer statistics, 2008," *CA: a cancer journal for clinicians*, vol. 58, no. 2, pp. 71–96, 2008.

[2] Edward A Sickles, "Breast cancer screening outcomes in women ages 40-49: clinical experience with service screening using modern mammography.," *Journal of*

the National Cancer Institute. Monographs, , no. 22, pp. 99–104, 1996.

[3] Ulrich Bick, "Mammography: How to interpret microcalcifications," in *Diseases of the Abdomen and Pelvis 2014–2017*, pp. 313–318. Springer, 2014.

[4] Maryellen L Giger, Nico Karssemeijer, and Julia A Schnabel, "Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer," *Annual review of biomedical engineering*, vol. 15, pp. 327–357, 2013.

[5] Arnau Oliver, Jordi Freixenet, Joan Marti, Elsa Perez, Josep Pont, Erika RE Denton, and Reyer Zwiggelaar, "A review of automatic mass detection and segmentation in mammographic images," *Medical Image Analysis*, vol. 14, no. 2, pp. 87–110, 2010.

[6] Jinshan Tang, et al., "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, no. 2, pp. 236–251, 2009.

[7] Silvia Bessa, Inês Domingues, Jaime S Cardoso, Pedro Passarinho, Pedro Cardoso, Vítor Rodrigues, and Fernando Lage, "Normal breast identification in screening mammography: A study on 18 000 images," in *BIBM*. IEEE, 2014, pp. 325–330.

[8] C Dromain, B Boyer, R Ferre, S Canale, S Delaloge, and C Balleyguier, "Computed-aided diagnosis (cad) in the detection of breast cancer," *European journal of radiology*, vol. 82, no. 3, pp. 417–423, 2013.

[9] I Anttinen, M Pamilo, M Soiva, and M Roiha, "Double reading of mammography screening films-one radiologist or two?," *Clinical Radiology*, vol. 48, no. 6, pp. 414–421, 1993.

[10] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *MICCAI*. Springer, 2015, pp. 652–660.

[11] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[12] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley, "Deep learning and structured prediction for the segmentation of mass in mammograms," in *MICCAI*. Springer, 2015, pp. 605–612.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks.," in *NIPS*, 2012, vol. 1, p. 4.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[15] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *DICTA*. IEEE, 2015, pp. 1–8.

[16] Christian Szegedy and et al., "Going deeper with convolutions," in *ICCV*, 2015, pp. 1–9.

[17] Yoshua Bengio, Patrice Simard, and Paolo Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[18] Zhi Lu, Gustavo Carneiro, Neeraj Dhungel, and Andrew P Bradley, "Automated detection of individual micro-calcifications from mammograms using a multi-stage cascade approach," *arXiv preprint arXiv:1610.02251*, 2016.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," *arXiv preprint arXiv:1603.05027*, 2016.

[20] Yann LeCun and Yoshua Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, pp. 1995, 1995.

[21] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[22] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML-2010*, 2010, pp. 807–814.

[23] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] John E Ball and Lori Mann Bruce, "Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation," in *EMBS*. IEEE, 2007, pp. 4973–4978.