Mammogram-Based Cancer Detection Using Deep Convolutional Neural Networks

Al Hussein Ahmed

Faculty of Media Engineering & Technology
German University in Cairo, Egypt
Email: elhussein.ibrahim@student.guc.edu.eg
husseinsamy32@hotmail.com

Mohammed A.-M. Salem

Faculty of Media Engineering and Technology
German University in Cairo, Egypt
Email: mohammed.salem@guc.edu.eg
Faculty of Computer and Information Sciences,
Ain Shams University, Cairo, Egypt
Email: salem@cis.asu.edu.eg

***Abstract*—In recent years, applying deep learning to medical images has experienced a surge but often comes with limitations related to the datasets: publicly available datasets have the drawback of being relatively small compared to other datasets used in image recognition tasks. We show multiple findings in our work: the immense power of Deep Convolutional Neural Networks even when applied on a small dataset such as the INbreast dataset. We also demonstrate that accuracy is not the only evaluation metric for network performance evaluation: the recall metric should be maximized. We also show the importance of using cross-validation to assure the absence of overfitting during the learning process. Results show an average classification accuracy for 5-fold cross-validation of 80.10% and an average AUC of 0.78. A graphical user interface was implemented in order to be tested by certified radiologists.**

***Keywords*—Deep Learning, Convolutional Neural Networks, Medical Images, Mammogram, Cancer Detection, Multiview NN.**

I. Introduction

Breast cancer is the most common cancer among women in 140 of 184 countries at 25%, and is the most common cancer type among women in Egypt at 34%. Screening mammography is one of the early stages in cancer detection. On multiple occasions, it cannot give a definitive assessment of the case and requires further examinations such as biopsies. These examinations being associated with their own health risks, in addition to frequent disagreement between radiologists on a mammogram's assessment, deep learning could be of huge benefit to the problem of mammogram classification. Computer Aided Detection (CAD) systems [1], [2], [3] dedicated for medical applications have shown impressive results, but the results achieved by deep learning in a vast range of fields makes the deep learning route very interesting to investigate. We implement a multiview deep convolutional neural network (DCNN) that takes six inputs: craniocaudal (CC) view, mediolateral oblique (MLO) view, with two corresponding mass and microcalcifications regions for each view and train our model with randomly initialized weights. We evaluate the model using the area under the curve (AUC), accuracy, precision and recall.

The manuscript is organized as follows: Section II contains the background to this study. The methodology that is adopted for the work in this paper is specified in Section III, while the obtained statistical results and the discussion are presented in Section IV. The conclusion and the future work are briefed in Section V.

II. Background

A. Literature Review

The datasets that were covered in previous works were DDSM [4], BCDR [5], INbreast [6] and mini-MIAS [7]. Krzysztof J. Geras et al. [8] used their own collected dataset in their work.

Preprocessing techniques that were used included cropping, data augmentation as well as local and global contrast normalization. The paper by Li Shen [9] cropped patches of size 224x224 by sampling different ROIs and background regions. J. Arevalo et al. [10] and Jose Gallego-Posada [11] applied Global and Local Contrast Normalization to the images before feeding them to the CNN. Since lighting conditions between different film images are variant, applying Global and Local Contrast Normalization could help neural networks to better identify the features, thus yielding better results.

Multiple approaches were taken by previous works. While some papers tried to use transfer learning [12], [7], [4] due to the relatively small size of the publicly available datasets, other papers investigated training from scratch. Krzysztof J. Geras et al. [8] had access to mammograms of 129,208 patients making a total of 886,437 images. This makes it the largest breast cancer screening dataset ever reported in literature, so training the network from scratch was the best option with such a huge dataset at hand. Krzysztof J. Geras et al. adopted the concept of a Multiview DCNN, i.e. multiple CNNs that are merged at a specific stage to train one single classifier. The aim of their work was to show the increase in performance when using high resolution non-downsampled images. The model takes 4 inputs: the CC view and the MLO view for each of the two breasts. No additional ROI annotations are input to the network, and this was as well due to the large size of the dataset: with a large number of data points, the network will be able to learn the specific features without the need to rely on manual segmentation. The work by Gustavo Carneiro et al. [4] on which this paper is based, used the concept of a multiview DCNN as well but doesnt use the whole case (the 4 views) as input to the DCNN. They actually made use of the annotations of breast masses and microcalcifications coming alongside the DDSM and INbreast dataset. The input to [4]'s multiview CNN was a six-image input: the two views of a single breast, in addition to two masks for each breast: one for microcalcifications and one for breast masses. Other papers such as [12] and [13] were aiming to classify the ROI regions, unlike previously mentionned works where the aim was to classify whole images. The paper by Li Shen [9] had a different and new approach. The approach was about first, training a network on patch classification, i.e. training it on classifying ROI regions only similar to [12] and [13]. Then by finetuning and adding new layers to the network, train the network on classifying whole mammogram images. The aim of this approach is to stop the reliance on ROI annotations: after training the patch classifier, the whole image classifier will train without the use of ROI annotations. The approach taken to turn the patch classifier into a whole image classifier was never encountered in previous works. It was about adding new convolution layers on top of the patch classifier. According to the paper, adding new convolutional layers on top is equivalent to applying the patch classifier on all patches in the whole image.

The best results attained by each paper are depicted in table I. Li Shen [9] achieved a per image AUC score of 0.96. For a 3 class problem, Gustavo Carneiro et al. [4] achieved an AUC score of over 0.9 with augmentation of 10 samples per image. Krzysztof J. Geras et al. [8] who trained the network without relying on ROI annotations, achieved an AUC score of over 0.8.

TABLE I: Survey on classification results of some recent work

Paper	**Evaluation metric**	**Score**
[12]	Accuracy	0.929
[13]	Accuracy	0.88
[4]	AUC score	Over 0.9
[10]	AUC score	0.860
[8]	AUC score	Over 0.8
[9]	Per-image AUC score	0.96

B. Concepts Overview

1) Mammography: Mammography is a specific type of breast imaging that is used for early breast cancer detection fo women. It uses x-rays in order to get images of different views of the breast. In general, two views of each breast (left and right) are required for a radiologist to make an analysis. These two views are cranial-caudal referred to as CC, and mediolateraloblique referred to as MLO. Radiologists look for multiple anomalies when reading a mammogram such as breast masses and calcifications. A metric called Bi-Rads score, is used to determine the severity of a case.

2) Artificial and Convolutional Neural Networks: Artificial neural networks (ANNs) fit under the umbrella of deep learning, a subfield of machine learning. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a large number of highly interconnected processing elements (neurones or simply nodes) working in unison to solve specific problems. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons, so in ANNs as well, the weights of the connections between different nodes are adjusted as the learning process is taking place (training). After the training is finished, the adjusted weights in the network will determine the most accurate prediction possible to a new input. An ANN is composed of an input layer, a number n of hidden layers, and an output layer; the connections between the nodes have weights that are adjusted with the training process. Convolutional neural networks (CNNs) is a class of ANNs, most commonly applied to analyzing visual imagery (images) [14]. It is an ANN that has specific properties since we know that we are working with images. So very simply, the input to the CNN is the image, or more specifically, the raw pixel values in each channel in the image (1 channel for grayscale and 3 channels for RGB images). There are mainly three operations in CNNs: convolution, activation layers, pooling(sub-sampling). A convolution layer has the aim of producing a feature map after being applied to a new input or an input coming from a previous layer. A filter or a set of filters (also referred to as kernels) of dimensions nxn slides through the image to produce a feature map. Activation functions are really important for a CNN to learn non-linear complex functional mappings between the inputs and the dependant variable. They introduce non-linear properties to the network. Sigmoid, tanh, softmax and RELU are examples of activation functions. Activation functions are not always visualized as layers when representing the architecture of a network. Most of the time they are visualized as functions applied in the end of a convolution or pooling layer. A pooling layer has 2 purposes, first it reduces the input volume which increases the computational efficiency, secondly, it helps generalizing the spatial properties of the features in the neural network, i.e. the exact location of a feature is not as important as its relative location to the other features. We can hence reduce overfitting to the training data. Overfitting means having a network that is tailor-made just to recognize properties of the training data but cant generalize to a broader range of examples. This will result in a high training accuracy but a low testing accuracy. These three operations could be considered as the feature extraction stage, i.e. a combination of these layers will form the feature extraction stage. After this stage, the output is transformed from a multi-dimensional vector to a 1D vector. This is called flattening. The output of the flattening stage is then input to a fully connected layer (FC layer) which could be considered as a normal ANN.

III. Methodology

The approach in this paper is based on the approach by Gustavo Carneiro et al. [4]. Our approach shows the huge power of neural networks even when applied on a small dataset such as the INbreast dataset. We highlight the importance of using cross-validation in order to ensure that the network is not overfitting to the training data.

A. Proposed Method

The proposed method in our paper follows the concept of a multiview CNN. This method proved to be effective in the works of [8], [13] and [15]. Krzystof et al. [8] used a four input CNN that take the four views RCC, RMLO, LCC, LMLO of a single patient per input without relying on the ROI annotations.

This approach could not be taken since the size of the INbreast dataset was very small compared to the dataset used in [8]. Similarily to the approach of Gustavo Carneiro et al. [4], a 6-input CNN that takes the following inputs per breast was implemented: 1- CC view. 2- Mass region(s) in CC view. 3- Microcalcifications region(s) in CC view. 4- MLO view. 5- Mass region(s) in MLO view. 6- Microcalcifications region(s) in MLO view.
The six CNNs merge to give a concatenated vector v, that in its turn trains a binary classifier (normal/benign or malignant). The architecture of the full network is depicted in Figure 1. Gustavo Carneiro et al. [4]'s inputs to networks 2,3,5,6 were the mass and microcalcifications masks however in our approach, the inputs to these networks are the the ROI masks applied on the mammograms.

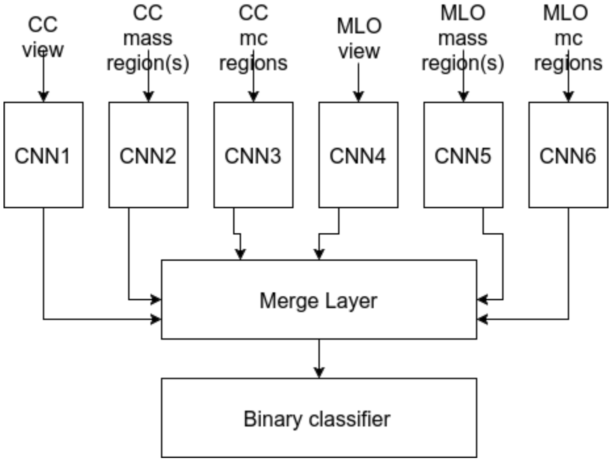

Fig. 1: Block Diagram of the Proposed Full CNN Model

The CNN architecture adapted for each CNN model is similar to the CNN-F model in [16], with modifications applied to some layers. Figure 2 shows the CNN architecture used in our approach until right before the fully connected layer and the classifier. Batch Normalization was applied after the first and second convolutional layers. After the feature extraction, the output vector is flattened into a 1D vector, then is fed to the fully connected layers. The FC part starts with 1024 neurons to which dropout of 0.5 is added, then another 1024 neurons with 0.5 dropout. The model in [16] used 4096 nodes in the fully connected layer instead of 1024, but less nodes had to be used for computational reasons. The single CNN model is depicted in Figure 2

B. Experimental Setup

1) Preprocessing: The implementation in this paper was done with the Python library Keras [17] with TensorFlow as the backend. Cases that had a single view of one breast were deemed insufficient to reflect the true state of the breast and were thus removed. After removal of missing cases, we had a total of 113 cases, that are either a case with four views for each breast (RCC, RMLO, LCC, LMLO), or a case with two views corresponding to one breast. A zero image matrix is inserted to the CNN instead of the missing breasts in the cases that contain two views for one breast only. The Bi-Rads scores provided for each view of the breast were mapped to zeros and ones since we are training a binary classifier: scores in [1,3] were mapped to a value of 0 (normal/benign), and [4,6] were mapped to a value of 1 (malignant). The next preprocessing step was using the XML files provided with the dataset to extract the ROI regions of the images. The provided contour points (x,y) in the XML files are used form a bounding box of the ROI with the corners: (x$_{min}$, y$_{min}$), (x$_{min}$, y$_{max}$), (x$_{max}$, y$_{min}$), (x$_{max}$, y$_{max}$). All images are resized to 150x150 pixels since it is the input shape of our neural network.

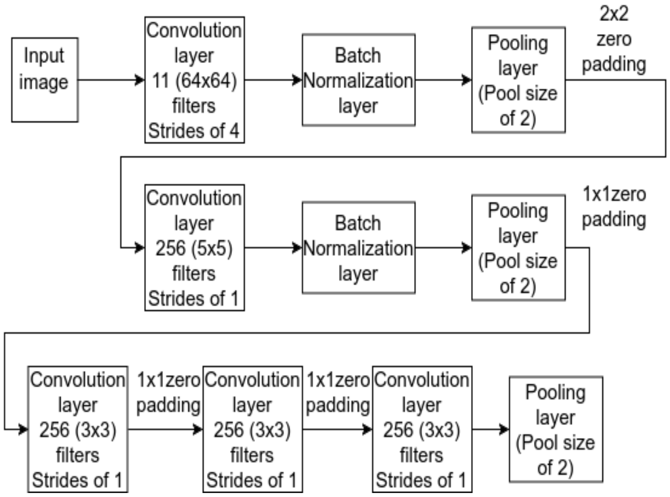
Input image
Convolution layer 11 (64x64) filters Strides of 4
Batch Normalization layer
Pooling layer (Pool size of 2)
2x2 zero padding
Convolution layer 256 (5x5) filters Strides of 1
Batch Normalization layer
Pooling layer (Pool size of 2)
1x1zero padding
Convolution layer 256 (3x3) filters Strides of 1
1x1zero padding
Convolution layer 256 (3x3) filters Strides of 1
1x1zero padding
Convolution layer 256 (3x3) filters Strides of 1
Pooling layer (Pool size of 2)

Fig. 2: Single CNN Architecture

C. Supervised Learning and Network Parameters

A single CNN takes an input image of size 150x150. After the merging layer of the six CNNs, the output layer has a sigmoid activation function for binary classification. The model is compiled using stochastic gradient descent with a learning rate of $1 * 10^{-4}$ and a momentum of 0.9. The loss is computed using the binary crossentropy loss function.
In order to make sure our network does not overfit to the training data, two approaches are taken. First, we shuffle all images in the six groups alongside their corresponding labels in a way that preserves the correlation between the six images in each input. This ensures that the results obtained are not due to a specific distribution of inputs between training and test sets. The second approach that we used is cross-validation which could be an indicator on whether the learning algorithm overfits or not. So, 5-fold cross validation was applied on our dataset using the python library sklearn [18]. All pixel values in images are rescaled from the range [0, 255] to [0, 1] for faster processing in the neural network and no data augmentation is applied. We use a batch size of 1 due to the small size of the dataset and train the multiview CNN for 50 epochs.

D. Graphical User Interface

A graphical user interface was implemented in order to test our system with real-life radiologists. This application takes the CC view and the MLO view for one breast. The radiologist then marks the mass regions and microcalcifications regions for each view and the system outputs a classification as either normal/benign or malignant. A sample from the graphical user interface is depicted in Figure 3. The red boxes are the ROI annotations provided by the radiologist. Breast masses are annotated on the left box, and microcalcifications are annotated on the right box.

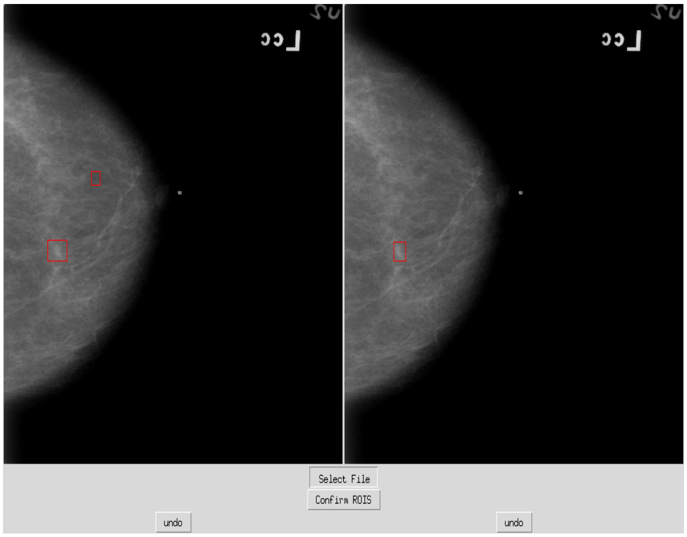

Fig. 3: Sample of the implemented graphical user interface

IV. Results & Discussion

A. Dataset

The INbreast dataset [13] was the dataset used in our work. It is available for public access at http://medicalresearch.inescporto.pt/breastresearch/. The database comprises 410 images corresponding to 115 cases. Most cases have 4 images per case: right craniocaudal (RCC), right mediolateral oblique (RMLO), left craniocaudal (LCC), left mediolateral oblique (LMLO). However, there are 25 out of 115 cases that have CC and MLO views of only one breast. Additionally, eight of the remaining 90 cases have images acquired at different timings. The two cases acquired at two different timings are treated as two separate cases in our experiments. The dataset is accompanied with pixel level annotations of the images showing the regions of interest (ROI) in the form of XML files. These files contain the regions of interest (ROI) as contour points for each type of finding. A sample of the dataset is depicted in Figure 4.

B. Evaluation Metrics

We used two metrics in our CNN evaluation: classification accuracy and AUC which stands for Area Under the Curve (the ROC curve). We also used a classification report that shows precision and recall values for each class in addition to a confusion matrix that depicts the number of true positives, false negatives, true negatives and false positives in our prediction.

1) Accuracy: Accuracy is a measure of how accurate a systems predictions were compared to the actual ground truth. It is the ratio between the number of correct predictions and the total number of predictions. Accuracy is a very important metric to evaluate the overall performance of a system however, it is not the only metric that one should be concerned about and could heavily depend on the distribution of classes in the dataset.

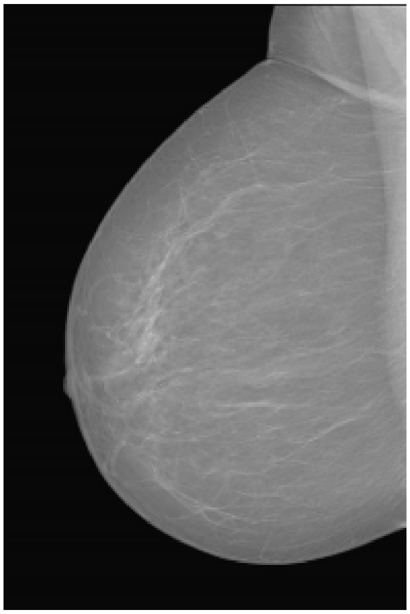

Fig. 4: Sample from the INbreast dataset

2) Precision and Recall: Precision and recall values are metrics that provide further indications than the accuracy. Precision p is defined as:

$$p = TP/TP + FP$$

with TP as true positive predictions and FP as false positives. So the higher the precision value, the less is the number of false positives.
Recall is noted as:

$$r = TP/TP + FN$$

where r represents recall value and FN represents false negative predictions. A high recall value means a less number of false negatives and in the problem of mammogram classification specifically, the number of false negatives should be minimized.

3) Area Under the ROC Curve (AUC): The ROC curve stands for Receiver Operating Characteristic curve. It is plotted on a graph where the false positive rate of predictions is plotted against the true positive rate. Using different decision threshold values for binary class probabilities, we plot points of (FP, TP). The area under this curve, the AUC, represents the probability that a randomly chosen positive sample is more likely to be classified as positive than a randomly chosen negative sample. AUC indicates how robust a system is.

4) Classification Results: Using 5-fold cross validation, a mean accuracy of 80.10% was achieved with a standard deviation of 3.5%. A classification report showing the precision and recall values after the testing phase is shown in table II. Figure 5 shows the ROC curve of the model at each fold alongside the curve that represents the mean of all 5 curves. The averaged model yields an AUC score of 0.78 which is comparable to recent papers such as Krzystof et

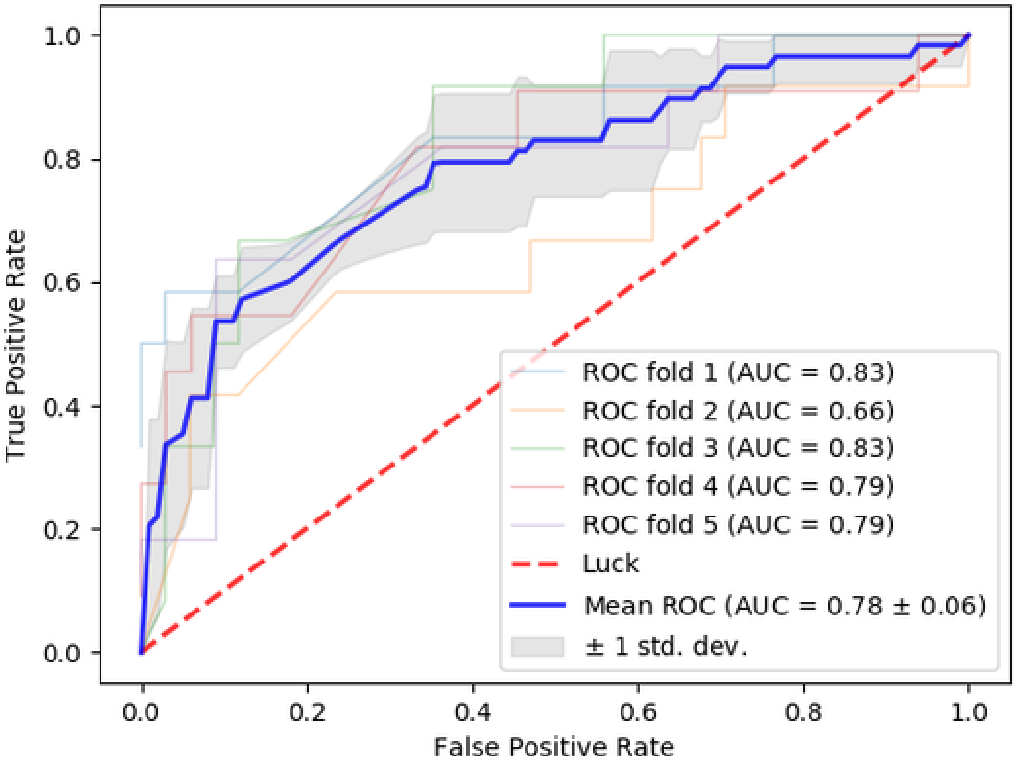

Fig. 5: ROC curves for the 5 folds of cross-validation and the mean ROC curve

al. [8]. The mean accuracy was achieved using a decision threshold of 0.5. It is noteworthy that recall value should be maximized in our system since in this specific problem concerning mammography, **false negative cases will go on to develop breast cancer**. So a tweak in the decision threshold could be made in order to achieve a higher recall value. This shows that ROC curves are a very powerful evaluation metric. Unlike the classification accuracy computed at one single threshold, the ROC curve gives a full view of the spectrum in order to achieve a system that suits our problem the most.

TABLE II: Classification Report

	Precision	**Recall**	**F1-score**	**Support**
0	0.84	0.91	0.87	34
1	0.67	0.50	0.57	12
average/total	0.79	0.80	0.79	46

5) Graphical User Interface Testing: Our system was tested by experts in the field of radiology that approved of the system as an important step towards the automation of cancer detection in mammograms. The application was also found as simple, user-friendly and easy to use.

V. Conclusion

This paper presented a multiview DCNN dedicated for mammogram classification. Due to the small size of our dataset, we had to rely on the provided ROI annotations to increase the performance of our system. We use cross-validation and shuffle the inputs in order to make sure that our model generalizes and does not overfit to the training data. We finally measure the performance of four system with two metrics: accuracy and AUC. An accuracy 80.10% was achieved and the AUC score was 0.78. We discussed how ROC curves give a bigger picture to the performance of our model in contrast to the accuracy, that is not always the only indicator to the robustness of the model.

For future work, having available large scale datasets will make a fully-automated approach to the mammography classification problem much more feasible. With such large data, no need for manual segmentation to be input to the DCNN and the network will have enough data to learn the specific features of mammograms and output accurate predictions. We would like also to explore the concept of localization of malignant lesions in a mammogram using DCNNs: this approach was never recorded in literature as far as we are concerned.

Acknowledgement

Special thanks to Dr. Menna El Sayed from the Faculty of Medicine, Cairo University, for giving us her time to try our system and give an opinion about it and what further improvements can be done.

References

[1] S. A. El-Regaily, M. A. M. Salem, M. H. A. Aziz, and M. I. Roushdy, “Lung nodule segmentation and detection in computed tomography,” pp. 72–78, Dec 2017.

[2] S. A. El-Regaily, M. A.-M. Salem, M. H. A. Aziz, and M. I. Roushdy, "Survey of computer aided detection systems for lung cancer in computed tomography," *Current Medical Imaging Reviews*, vol. 14, no. 1, pp. 3–18, 2018. [Online]. Available: http://www.eurekaselect.com/node/152893/article

[3] M. A.-M. Salem, A. Atef, A. Salah, and M. Shams, "Recent survey on medical image segmentation," pp. 129–169, 2018.

[4] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 652–660.

[5] M. A. Guevara Lopez, N. Gonzlez Posada, D. Moura, R. Ramos Polln, J. Franco-Valiente, C. Ortega, M. Del Solar, G. Daz-Herrero, I. Pereira M A Ramos, J. Pinheiro Loureiro, T. Cardoso Fernandes, and B. Ferreira M Arajo, "Bcdr: A breast cancer digital repository," pp. 1065–1066, 01 2012.

[6] I. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. Cardoso, and J. Cardoso, "Inbreast: Toward a full-field digital mammographic database," vol. 19, pp. 236–48, 11 2011.

[7] J. Suckling, "The mammographic image analysis society digital mammogram database"exerpta medica," vol. 1069, 01 1994.

[8] K. J. Geras, S. Wolfson, S. G. Kim, L. Moy, and K. Cho, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," *CoRR*, vol. abs/1703.07047, 2017. [Online]. Available: http://arxiv.org/abs/1703.07047

[9] L. Shen, "End-to-end training for whole image breast cancer diagnosis using an all convolutional design," *CoRR*, vol. abs/1708.09427, 2017. [Online]. Available: http://arxiv.org/abs/1708.09427

[10] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. Guevara Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Comput. Methods Prog. Biomed.*, vol. 127, no. C, pp. 248–257, Apr. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.cmpb.2015.12.014

[11] J. Gallego-Posada, D. A. Montoya-Zapata, and O. L. Quintero-Montoya, "Detection and diagnosis of breast tumors using deep convolutional neural networks," 2016.

[12] D. Lvy and A. Jain, "Breast mass classification from mammograms using deep convolutional neural networks," 12 2016.

[13] P. U. Hepsa, S. A. zel, and A. Yazc, "Using deep learning for mammography classification," pp. 418–423, Oct 2017.

[14] A. M. Abdelhalim and M. A. . Salem, "Intelligent organization of multiuser photo galleries using sub-event detection," in *2017 12th International Conference on Computer Engineering and Systems (ICCES)*, Dec 2017, pp. 436–440.

[15] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," *CoRR*, vol. abs/1505.00880, 2015. [Online]. Available: http://arxiv.org/abs/1505.00880

[16] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *CoRR*, vol. abs/1405.3531, 2014. [Online]. Available: http://arxiv.org/abs/1405.3531

[17] F. Chollet *et al.*, "Keras," 2015.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, and G. Louppe, "Scikit-learn: Machine learning in python," vol. 12, 01 2012.