A MULTI-VIEW DEEP LEARNING ARCHITECTURE FOR CLASSIFICATION OF BREAST MICROCALCIFICATIONS

Alan Joseph Bekker[†] *Hayit Greenspan* [⋆] *Jacob Goldberger* [†]

[†] Faculty of Engineering, Bar-Ilan University, Israel
[⋆] BioMedical Engineering, Tel-Aviv University, Israel

ABSTRACT

In this paper we address the problem of differentiating between malignant and benign tumors based on their appearance in the CC and MLO mammography views. Classification of clustered breast microcalcifications into benign and malignant categories is an extremely challenging task for computerized algorithms and expert radiologists alike. We describe a deep-learning classification method that is based on two view-level decisions, implemented by two neural networks, followed by a single-neuron layer that combines the view-level decisions into a global decision that mimics the biopsy results. Our method is evaluated on a large multi-view dataset extracted from the standardized digital database for screening mammography (DDSM). Experimental results show that our network structure significantly improves on previously suggested methods.

***Index Terms*—** Mammography, Microcalcifications, multi-view analysis, deep-learning, Computer-aided diagnosis (CADx)

1. INTRODUCTION

A screening mammographic examination usually consists of four images, corresponding to each breast scanned in two views: the mediolateral oblique (MLO) view and the craniocaudal (CC) view. The MLO projection is taken in a 45° angle and shows part of the pectoral muscle. The CC projection is a top-down view of the breast. Both views are included in the diagnostic procedure. When reading mammograms, radiologists judge whether or not a malignant lesion is present by examining both views and breasts. In an expert diagnosis procedure, the expert looks at each of the views separately, and delivers one final assessment. When a radiologist does not observe a lesion in both views this can influence interpretation and decision making. Recent studies (e.g. [1] [2] [3] [4]) have demonstrated the superior performance of a multi-view CADx system over its single-view counterpart. These studies have mainly addressed the problem of mammography analysis in the presence of masses. The main focus of most previous multi-view methods was to improve the localized detection of breast cancer or to build extended feature sets using both views. In this study we address the problem of mammography analysis in the presence of micro-calcifications (MC). In this case the detection is mainly based on texture features and it is not useful to find correspondence between MC clusters in different views. A recent work [5] proposed multi-view modeling based on the EM algorithm with a logistic-regression view-level decision. We compare our algorithm to that work and show that we achieve significantly better results.

Our study is based on a large number of cases from the DDSM [6], the largest public mammogram database available. The DDSM is a labeled dataset that can be used to train an automatic system. It contains the MCs location in each of the two views marked by experts. We also have the biopsy results, whether the abnormities were benign or malignant. However, we do not have direct information as to whether there was an explicit indication of malignant MCs from each of the views. We took all the cases from the DDSM dataset with both CC and MLO views. Experiments were performed on 1410 images consisting of 705 pairs of CC+MLO views extracted from the DDSM dataset.

Many studies have examined the task of classifying MC clusters into benign or malignant. However, most studies have used different datasets [7] [8] [9], thus making it difficult to carry out a comparison. Furthermore, most studies have employed smaller datasets than shown here. Only a few studies have used a large number of MC cases from the DDSM along with texture features [10] [11] [12]. In Section 3.2 we compare the results of our algorithm to those obtained in these studies.

In this study we explicitly take into account the two-view structure of the problem by constructing a suitable neural-network architecture. Experimental results on the DDSM dataset show that this approach significantly outperforms previously suggested methods.

2. MULTI-VIEW NEURAL NETWORKS

In this study we apply the deep learning paradigm to the task of automatic classification of breast microcalcifications based

A. J. Bekker is funded in part by the Intel Collaborative Institute for Computational Intelligence (ICRI-CI).

978-1-4799-2349-6/16/$31.00 ©2016 IEEE

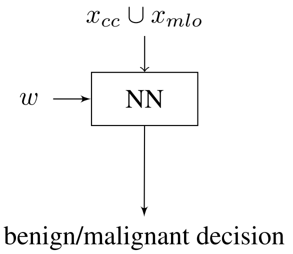

Fig. 1: A diagram of the standard NN classification model. A concatenation of CC and MLO views is used as input to a NN classifier that produces the benign/malignant decision.

on two mammography views. The DDSM labeled dataset is used to train the classifier. The straightforward way to apply a neural network (NN) is to extract features from the CC views and features from the MLO view. Then we train a neural network using the concatenated features as input and the biopsy result as binary output. This approach is illustrated in Fig. 1. There is, however, a structure specific to this problem that is not explicitly modeled by this standard neural network solution. The provided biopsy result is not always completely aligned with the image-level diagnosis. Furthermore, a biopsy-based malignant decision may be reflected in only one of these views. Given these drawbacks of standard NN, we suggest a NN architecture that is tailored to the problem of multi-view MC detection which is indirectly supervised by biopsy results. Our approach is based on allocating a separate NN for each view and then combining the view-level soft decisions in a non-linear way.

Assume that for each patient we have feature vectors x_{cc} and x_{mlo} extracted from the two mammography views, CC and MLO, respectively. The CC feature vector x_{cc} is used as input to a neural network NN$_{cc}$ with a parameter-set w_{cc}. Let p_{cc} be the probabilistic output of the classifier NN$_{cc}$ that provides a CC view-level benign/malignant decision (we use the convention 0-benign and 1-malignant). Let $y_{cc} \in \{0, 1\}$ be the binary random variable that represents the decision whether an MC cluster is benign or malignant based only on the CC view, i.e., $p(y_{cc} = 1|x_{cc}; w_{cc}) = p_{cc}$. In a similar way x_{mlo} serves as input to NN$_{mlo}$ and p_{mlo} is the soft decision based on the MLO view.

We next integrate the view-level decisions into a unified patient-level decision that mimics the biopsy test results. We take the view-level outputs p_{cc} and p_{mlo} and use them as input to another layer consisting of a single neuron with a sigmoid activation function:

$$p(z = 1|x_{cc}, x_{mlo}) = \sigma(p_{cc} + p_{mlo} - 1) \qquad (1)$$

such that $\sigma(u) = \frac{1}{1+\exp(-u)}$ is the sigmoid function. The binary r.v. z represents the biopsy-based decision. It can be easily verified from Eq. (1) that $p(z = 1|x_{cc}, x_{mlo}) > 0.5$ if and only if $(p_{cc} + p_{mlo})/2 > 0.5$. Hence, the network's final hard decision is obtained by averaging the two view-level decisions. We dub this model the Multi-View Neural Network (MV-NN). It is illustrated in Fig. 2.

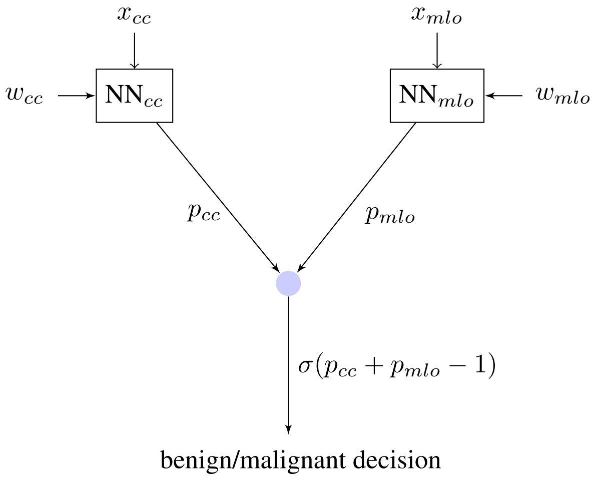

Fig. 2: A diagram of the multi-view NN (MV-NN) classification model. The CC and MLO views are used as input to NN classifiers that produce view-level probabilities p_{cc} and p_{mlo}. These are used as input to a single-neuron layer that produces the final decision.

Another way to combine the views is simply by averaging the decisions, i.e.:

$$p(z = 1|x_{cc}, x_{mlo}) = \frac{p_{cc} + p_{mlo}}{2}. \qquad (2)$$

We have found, however, that adding non-linearity to the merging process provides better results.

The MV-NN model parameters can be learned from a labeled training data. Assume we are given n (CC, MLO) pairs of feature vectors:

$$(x_{1,cc}, x_{1,mlo}), ..., (x_{n,cc}, x_{n,mlo})$$

with their corresponding binary labels $z_1, ..., z_n \in \{0, 1\}$. The log-likelihood function of the model parameters is:

$$S(w_{cc}, w_{mlo}) = \sum_{t=1}^{n} \log p(z_t|x_{t,cc}, x_{t,mlo}; w_{cc}, w_{mlo}) \quad (3)$$

Substituting Eq. (1) in Eq. (3) we obtain:

$$S(w_{cc}, w_{mlo}) = \sum_{t=1}^{n} \log \sigma((2z_t-1)(p_{t,cc}+p_{t,mlo}-1)) \quad (4)$$

such that $p_{t,cc} = p(y_{t,cc}|x_{t,cc}, w_{cc})$ is the output of the CC neural network and $p_{t,mlo}$ is defined in a similar way. The back-propagation algorithm splits the classification error between the two views as follows:

$$\begin{aligned}\frac{\partial S}{\partial w_{cc}} &= \sum_{t=1}^{n}(z_t - \hat{z}_t)\frac{\partial p_{t,cc}}{\partial w_{cc}}\\ \frac{\partial S}{\partial w_{mlo}} &= \sum_{t=1}^{n}(z_t - \hat{z}_t)\frac{\partial p_{t,mlo}}{\partial w_{mlo}}\end{aligned} \qquad (5)$$

such that $\hat{z}_t = p(z_t|x_{t,cc}, x_{t,mlo}; w_{cc}, w_{mlo})$ is computed in the feed-forward step.

Our approach which is based on learning two networks in parallel resembles the concept of Siamese nets that was first introduced to solve signature verification as an image matching problem [13] [14] and has become popular in recent years. A Siamese neural network consists of twin networks which accept distinct inputs but are joined by an energy function at the final layer. This function computes some metric between the highest-level feature representation on each side. In a Siamese network, as the name implies, the parameters between the twin networks are tied. In our network the CC and MLO views are different and, therefore, the view-level networks are not the same. Another major difference is that Siamese networks take pairs of input vectors and minimize or maximize a distance depending on whether a pair comes from the same or different classes. In our approach the cost function aims to integrate knowledge from different sources.

2.1. Dataset and features

This study is based on the DDSM dataset [6] which provides the highest available number of annotated mammograms with a biopsy-proven diagnosis. The contours of the lesions are provided by a chain code which we used to extract irregular shaped ROIs. We extracted ROIs that contained clusters of MCs for which a proven pathology was found. We chose patients in the DDSM dataset that had both CC and MLO views in order to test our model. Our dataset was comprised of 1410 clusters (705 of CC, and 705 of MLO), of which 372 were benign and 333 were malignant. Feature vectors x_{cc} and x_{mlo} were extracted from the CC and MLO views, respectively. Following [5], the features were extracted from the Curvelet coefficients at intermediate scales (in our study, two scales), and included the four features mentioned in [15] for each scale, with three additional features: entropy, skewness and kurtosis. Overall, each extracted ROI was represented by 14 features. Many other texture features that can be used for mammography analysis have been reported in the literature, e.g. GLCM [16], (GLRLM) [17] [18], Gabor filters [19] and features that are based on the wavelet transform. Using the Curvelet features we obtained the best results. Due to lack of space and since this is not the focus of this work, we do not describe here classification results based on the alternative features.

2.2. Training procedure

Using the feature described in Section 2.1, the size of the input feature set is 28 (14 features for each view). We used a two hidden layer NN comprised of 24 neurons each (12 neurons for each view). Overall, the number of parameters (linear coefficients and bias terms) for each view is $15 \times 12 + 13 \times 12 + 13$. To learn the network weights, we used the gradient descent algorithm (since the dataset size is small there is no need here for stochastic optimization based on mini-batchs). We used an adaptive learning rate combined with momentum. The learning rate was initialized to 0.01. It was then increased in each epoch by multiplying the learning rate by 1.05 if the new likelihood exceeds the old likelihood score by more than 4%. Otherwise, the learning rate is kept. If the likelihood score was less than the old likelihood, the learning rate was decreased by multiplying the learning rate by 0.7. The momentum was set to 0.5. To prevent overfitting the number of maximal training epochs was set to 100. The parameters of sub-network NN$_{cc}$ were initialized by training a NN that has CC features as input and the biopsy labels as output. The MLO sub-network was initialized in a similar way.

3. EXPERIMENTAL EVALUATION

3.1. Compared methods

We compared the proposed MV-NN method to logistic regression (LR), SVM, and neural network classifiers. We implement these classifiers on each view separately and on a concatenation of the CC and MLO features. To conduct a fair comparison we chose a neural network architecture similar to the one we used in our model (2 hidden layers each comprised of 24 neurons).

We also implemented two classifiers that explicitly take the multi-view structure of the problem into account. The first one is the EM-LR algorithm [5]. It is based on two view-level logistic regression classifiers that are combined by the EM algorithm. We also introduced an extended version of the EM-LR algorithm where the view-level logistic-regression is replaced by a view-level neural network (we used the same network architecture described above). We denote this extension EM-NN. One drawback of these methods is that both NN (used for the view-level) and EM (used for combining the views) are iterative methods and it is not apparent what the optimal scheduling for the iterations of the two methods should be. At each M-step of the EM-iteration we need to train a NN. In contrast, in the MV-NN approach the two view-level decisions and the two-view integration are done by a single NN. Hence, given the current practice of NN training, the parameter training of the MV-NN is easily done.

Table 1: Classification results (benign vs. malignant) for breast tissues.

	method	CC	MLO	CC+MLO
Accuracy	LR	61.3	61.0	61.7
	SVM	61.8	60.8	64.4
	EM-LR	-	-	69.5
	NN	76.6	77.3	77.7
	EM-NN	-	-	78.3
	MV-NN	-	-	**78.7**
AUC	LR	0.71	0.71	0.71
	SVM	0.72	0.72	0.73
	EM-LR	-	-	0.75
	NN	0.80	0.82	0.85
	EM-NN	-	-	0.87
	MV-NN	-	-	**0.89**
Sensitivity	LR	61.4	60.3	61.9
	SVM	60.0	61.2	62.7
	EM-LR	-	-	68.1
	NN	75.5	76.3	77.7
	EM-NN	-	-	78.5
	MV-NN	-	-	**78.8**
Specificity	LR	61.0	61.0	61.6
	SVM	63.0	60.3	65.0
	EM-LR	-	-	69.7
	NN	76.5	77.0	77.7
	EM-NN	-	-	78.3
	MV-NN	-	-	**78.7**

3.2. Classification Results

We evaluated algorithm performance using Receiver Operator Characteristic (ROC) curves, by calculating the area under the curve (AUC) and using the classification accuracy, sensitivity and specificity measures. The results were computed using 10-fold cross validation. In this experimental set-up there is a complete isolation of the test set from the train set. Each fold was only used for testing and never for training. We thus ensured that no bias was introduced.

Table 1 shows classification results for the six classifiers described above, and Fig. 3 shows the corresponding ROC curves. As can be seen from the classification results, the proposed MV-NN approach outperformed all other methods. Table 1 shows that for all methods taking two-views instead of a single view improved performance. Table 1 also indicates that methods based on deep architecture significantly outperformed SVM and LR. Of the NNs the standard implementation was the worst. The two multi-view NNs we introduced in this work, namely EM-NN and MV-NN, obtained the best results. Comparison of these two view-integration methods showed that MV-NN was better. We performed a t-test on the AUC values of the benchmark models and the MV-NN and EM-NN models. The input to the t-test consisted of AUC samples taken from 10-fold cross validation of both models [20]. The t-test examined the hypothesis that the two groups of AUC values came from the same distribution given the mean and standard deviation of the AUCs (taken from the 10-fold cross validation experiments). All the hypotheses were successfully rejected with a p-value < 0.01 and a confidence interval of 99%, which indicates that the AUCs of MV-NN were significantly higher than the AUCs of the benchmarks. In addition, we performed a t-test between the MV-NN accuracy, sensitivity and specificity vs. the benchmarks using the procedure described above, achieving for all a p-value < 0.01 with a confidence interval of 99%.

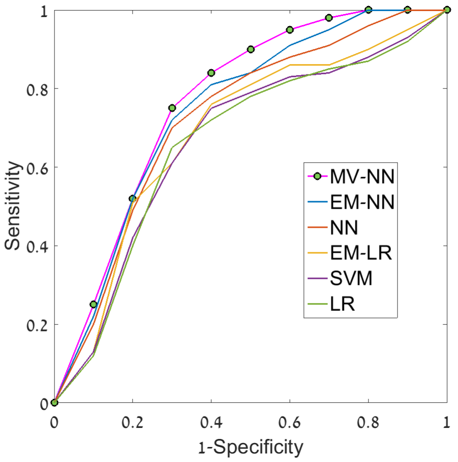

Fig. 3: ROC curves of the six classifiers. All classifiers used both CC and MLO data.

When classifying benign versus malignant clusters of MCs using the DDSM dataset, Pereira et al. [10] reported an AUC=0.607. The best feature for classifying MCs in Andreadis et al. [11] achieved ACC=70.14% (AUC=0.776) for fatty tissues and ACC=60.83% (AUC=0.636) for dense tissues, and Moura et al. [12] achieved AUC=0.776. Our MV-NN (using Curvelet rotation invariant features) achieved significantly better results (ACC=78.7%,AUC=0.89) than these studies.

To conclude, in this paper we introduced and evaluated two neural network architectures, MV-NN and EM-NN, to classify breast MCs based on the CC and MLO views. We showed that a special-purpose NN architecture yields better results than the standard NN and overall the results we obtained were significantly better than those reported in previous studies.

4. REFERENCES

[1] W. Qian, D. Song, M. Lei, R. Sankar, and E. Eikman, “Computer-aided mass detection based on ipsilateral multiview,” *mammograms Academic Radiol.*, vol. 14, pp. 530–538, 2007.

[2] M. Velikova, M. Samulski, P. J. F. Lucas, and N. Karssemeijer, “Improved mammographic CAD performance using multiview information: a Bayesian network framework,” *Phys. Med. Biol.*, vol. 54, pp. 1131–1147, 2009.

[3] M. Samulski and N. Karssemeijer, “Optimizing case-based detection performance in a multi-view CAD system for mammography,” *IEEE Trans on Med Imaging*, vol. 30, pp. 1001–1009, 2011.

[4] M. Velikova, P. Lucas, M. Smulski, and N. Karssmeijer, “A probabilistic framework for image information fusion with an application to mammographic analysis,” *Medical Image Analysis*, vol. 16, pp. 865–875, 2012.

[5] A. J. Bekker, M. Shalhon, H. Greenspan, and J. Goldberger, “Learning to combine decisions from multiple mammography views,” in *IEEE Int. Symp. on Biomedical Imaging (ISBI)*, 2015.

[6] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, “The digital database for screening mammography,” *Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed*, pp. 212–218, Medical Physics Publishing, 2001.

[7] L. Wei, Y. Yang, and R. M. Nishikawa, “Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis,” *Pattern recognition*, vol. 62, pp. 1126–1132, 2009.

[8] H. Soltanian-Zadeh, Farshid Rafiee-Rad, and Siamak Pourabdollah-Nezhad, “Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms,” *Pattern Recognition*, vol. 37, no. 10, pp. 1973–1986, 2004.

[9] A. Papadopoulos, D. I. Fotiadis, and A. Likas, “Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines,” *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 141–150, 2005.

[10] R. Pereira, P. M. Azevedo-Marques, M. O. Honda, S. K. Kinoshita, R. Engelmann, C. Muramatsu, and K. Doi, “Usefulness of texture analysis for computerized classification of breast lesions on mammograms,” *Journal of Digital Imaging*, vol. 20, no. 3, pp. 248–255, 2007.

[11] I. I. Andreadis, G. M. Spyrou, and K. S. Nikita, “A comparative study of image features for classification of breast microcalcifications,” *Meas Sci Technol*, vol. 22, no. 11, pp. 114005–114013, 2011.

[12] D. C. Moura. and M. A. Guevara-Lopez, “An evaluation of image descriptors combined with clinical data for breast cancer diagnosis,” *Int. J. Computer Assisted Radiology and Surgery*, vol. 8, pp. 561–574, 2013.

[13] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Sackinger, and R. Shah, “Signature verification using a siamese time delay neural network,” *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, pp. 669–688, 1993.

[14] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in *IEEE Comp. Vision and Patt. Recog. (CVPR)*, 2005.

[15] Y. Shang, Y. Diao, and C. Li, “Rotation invariant texture classification algorithm based on curvelet transform and SVM,” in *Machine Learning and Cybernetics, 2008 International Conference on*, 2008, vol. 5, pp. 3032–3036.

[16] R.M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[17] M. M. Galloway, “Texture analysis using gray level run lengths,” *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172–179, 1975.

[18] X. Tang, “Texture information in run-length matrices,” *Image Processing, IEEE Transactions on*, vol. 7, no. 11, pp. 1602–1609, 1998.

[19] J. G. Daugman, “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters,” *J. Opt. Soc. Am. A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[20] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve.,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.