Breast Mass Classification in Mammograms using Ensemble Convolutional Neural Networks

Andrik Rampun, Bryan W. Scotney
and Philip J. Morrow
School of Computing, Ulster University,
Coleraine BT52 1SA, UK
Email: {y.rampun,bw.scotney,pj.morrow}@ulster.ac.uk

Hui Wang
School of Computing, Ulster University,
Jordanstown, Newtownabbey BT37 0QB, UK
Email: h.wang@ulster.ac.uk

***Abstract*—The paper presents quantitative results of a preliminary study undertaken as part of Decision Support and Information Management System for Breast Cancer (DESIREE). DESIREE is a European-funded project to improve the management of primary breast cancer through image-based, guideline-based, experience-based, and case-based information systems. In this study we explore the use of ensemble deep learning for breast mass classification in mammograms. The proposed method is based on AlexNet with some modifications in order to adapt it to our classification problem. Subsequently, model selection is performed to select the best three results based on the highest validation accuracies during the validation phase. Finally, the prediction is based on the average probability of the models. Experimental evaluation shows that accuracy from individual models ranges between 75% and 77%, but combining the best models (ensemble networks) results in over 80% classification accuracy and aura under the curve.**

I. Introduction

A recent report in [1] estimated that approximately 252,000 new cases of invasive breast cancer would be diagnosed in the United States in 2017, and every year around 11,400 women die from breast cancer in the United Kingdom [2]. Mammography is the most common imaging technology used for screening breast cancer to find early signs of abnormality. In current clinical practice, radiologists have to examine the whole mammogram of a patient and doctors require biopsy test to decide whether a tumour is benign or malignant. Although the overall current clinical methods have improved greatly in the last two decades, there are still a number of deficiencies such as variability among radiologists and that procedures are time consuming and invasive.

Computer-aided diagnosis (CAD) systems can assist clinicians in terms of efficiency, effectiveness and consistency. CAD systems can assess lesions uninvasively and make predictions as to whether a lesion is benign or malignant. However, developing a CAD system that is able to replicate radiologists' knowledge requires a significant amount of time and effort. Machine learning is a branch of Artificial Intelligence which enable machines to learn from experience and make predictions for future occurrences. In the last few years, the success of deep learning in many classification problems has attracted computer scientists, particularly in the medical imaging domain. As a result, hundreds of papers about deep learning for medical image analysis have been published according to the studies by Litjens *et al.* [3] and Hamidinekoo *et al.* [4].

In this paper, we present our preliminary results using ensemble Convolutional Neural Networks (CNNs) for breast mass classification in mammograms taken from the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [5]. This work is part of the European funded project Decision Support and Information Management System for Breast Cancer (DESIREE) which applies computer vision techniques in several problem domains such as breast segmentation [6], breast density classification [7], [8], breast mass [9] and micro calcification cluster [10] classification.

II. Literature Review

This section reviews some of the existing methods in the literature, which we divide into two categories: (a) Based on conventional machine learning methods which use hand crafted features and feature selection phase and (b) Based on deep learning methods bypassing feature extraction and selection phases.

A. Conventional machine learning methods

Muramatsu *et al.* [11] extended local ternary patterns (LTP) into radial LTP, which takes account of the pattern orientation with respect to the lesion centre making the local patterns robust for differentiating between circumscribed and spiculated margins. Several classifiers and patch dimensions were investigated and the Artificial Neural Network (ANN) produced the highest area under the curve (AUC) value of 0.90 at patch size 200×200. Narváez and Romero [12] used continuos (Zernike) and discrete (Krawtchouk) orthogonal moments to characterise breast masses. A k-nearest neighbours strategy was employed as a classification approach and they reported an accuracy of up to 90% and $AUC = 0.93$. Reyad *et al.* [13] made a quantitative comparison between first-order statistical, local binary patterns (LBP) and multi-resolution wavelet-based features using a support vector machine (SVM) classifier and reported above 98% accuracy when fusing all features. Choi and Ro *et al.* [14] used a multiresolutioin LBP approach and proposed a variable selection technique to select a subset of discriminant features to maximise the separation

between breast masses and normal tissues. The proposed method was evaluated on false positive reduction and over 90% accuracy was achieved. Despite high accuracies reported in the literature, recent studies of Rampun *et al.* [9], [10] showed that conventional machine learning methods may not be useful enough when it comes to it's actual implementation in a real clinical environment. This is due to many cases classified correctly (high accuracy) but most of them are classified with probability outputs between 0.5 and 0.7 (low confidence). In a real clinical environment, CAD systems is used as a 'second reader' opinion, hence a system with high confidence measure and high accuracy is more useful to assist radiologists in diagnostic decision making.

B. Deep learning methods

For deep learning based methods, CNN been widely applied to breast mass classification. Levy and Jain compared the performance of GoogleNet [16] and AlexNet [17] and reported 89% and 92.9% accuracy, respectively. Jiao *et al.* [19] developed a CNN architecture to classify benign and malignant masses of breast cancer by utilising the combination of low and high level deep features from two different layers. Arevalo *et al.* [20] evaluated their CNN framework against the histogram of oriented gradient and the gradient divergence methods, which extracted the features from the histogram. In 2015, Dhungel *et al.* [21] developed an algorithm using a cascade of deep learning and used a random forest classifier to detect suspicious regions in mammograms. Their algorithm consisted of a multi-scale deep belief network (DBN) to detect all potential suspicious masses, a CNN to keep the correct candidates of those regions, and a random forest classifier to reduce false positives. Shen [22] develop an end-to-end training algorithm for whole-image breast cancer diagnosis which has the advantage of training a deep learning model without relying on cancer lesion annotations. A model averaging technique was used to make a final prediction, producing AUC scores of 0.91 and 0.96 on two different datasets.

III. Methodology

Figure 1 shows an overview of the work flow in our study. Firstly, our CNN model is trained with 50 epochs. Once the training is completed, we select the best three CNN models based on top three highest validation accuracies. Subsequently, each selected CNN model was evaluated based on the testing set. The final prediction is based on the average prediction of the three models.

A. Materials and Dataset

The dataset used in this study is taken from the CBIS-DDSM [5]. In total it contains 1593 masses (829 benign and 764 malignant) of both mediolateral oblique (MLO) and craniocaudal (CC) views from 838 patients. Each case is a biopsy proven 'benign' versus 'malignant' annotation by expert radiologists. Each mass contour is provided by an expert radiologist. The dataset is randomly split by patient into training (70%), validation (20%) and testing (10%) sets. Figure 2 shows examples of malignant and benign breast masses taken from CBIS-DDSM.

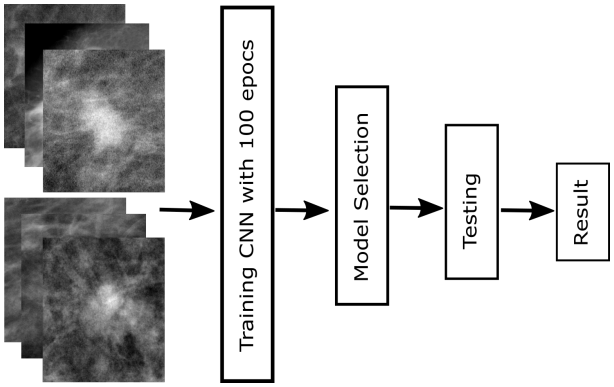

Fig. 1. Work flow overview of our study.

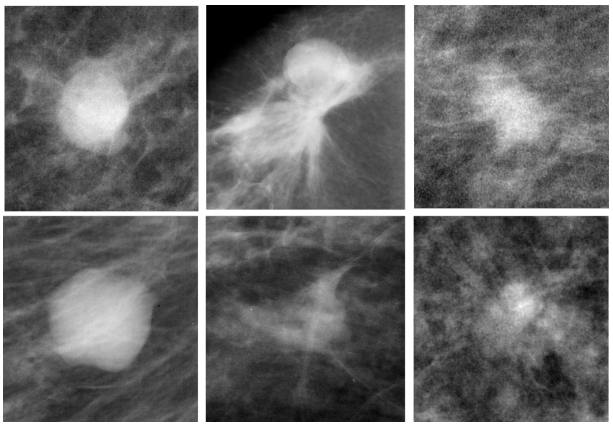

Fig. 2. Top and bottom rows represent samples of malignant and benign masses, respectively.

For the implementation, the proposed CNN was trained, validated and tested on an Intel Xeon E5-1620 v3 processor, using Nvidia Corporation's Deep Learning GPU Training System (DIGITS) based on Caffe, with Nvidia's GeForce GTX1080 8GB GPU on Intel Core i7-4790 Processor with Ubuntu 16.04 operating system.

B. Data Augmentation

Since most deep learning based methods require a large number of images to learn the characteristics of different classes, a data augmentation process is an essential step in this study. To address the need for training deep learning with large sets of labelled data, for each bounding box mass (P_1) we performed the following data augmentation techniques:

- Double the dimension of P_1 (P_2). This captures the surrounding texture information around the mass.
- Take the mass region only in P_1, removing the surrounding background (P_3). This enables the network to learn the shape and margin of the mass.
- Take only 50% the dimension of P_1 for a zoom-in effect of the mass (P_4).

- Take only 70% the dimension of P_1 for a zoom-in effect of the mass (P_5).
- Double the dimension of P_2 to capture the texture information around the mass (P_6).
- For each mass patch (P_1, ... , P_6), we perform five random rotations $0 < \theta < 360$.

Hence, this phase generates 30 images in total for each breast mass patch. Figure 3 shows results of data augmentation of P_1. It can be observed that there are six main patches. These patches are further augmented via random rotation $0 < \theta < 360$. Finally, all images are resized to 224×224.

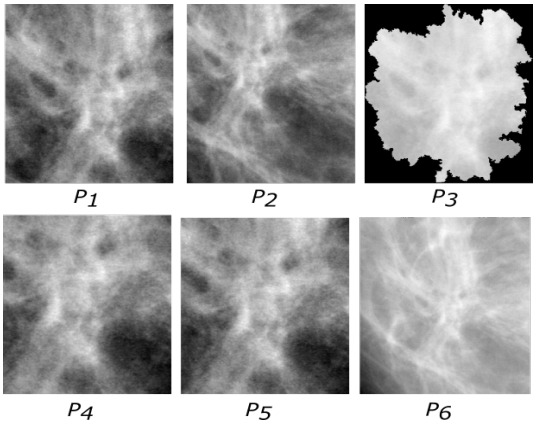

Fig. 3. Examples of breast mass after data augmentation. Note that each of these main patch will be randomly rotated five times.

C. Network Architecture

The proposed network is similar to AlexNet [17] and we have changed the last fully connected layer into two outputs representing malignant and benign classes. Figure 4 shows a visualisation of the proposed network architecture. The batch sizes are 128 and 64 for training and validation, respectively, and 50 epochs. Furthermore, we employed the Adaptive Movement Estimation (Adam) [23] as an optimisation scheme with base learning rate$= 0.0001$ and a step-down learning policy, momentum$= 0.9$, weight decay$= 9.9 \times 10^{-7}$ and drop out rate is 0.5. We also performed mean image subtraction during the training of the network and used a small learning rate of 0.1 in each convolution layer.

To adapt the AlexNet architecture to our problem, we made the following modifications:

- AlexNet uses Local Response Normalisation (LRN) whereas in our network we use the Batch Normalisation (BNorm) technique.
- To avoid over-fitting we applied BNorm for every convolutional layer whereas in AlexNet LRN is applied only the first two convolution layers.
- Our network uses the Parametric Rectified Linear Unit (PReLU) [24] activation function, whereas AlexNet uses the Rectified Linear Unit (ReLu) [25].
- We applied dropout and PReLu at every fully connected (FC) layer, whereas AlexNet avoids over-fitting only at the last two FC layers.

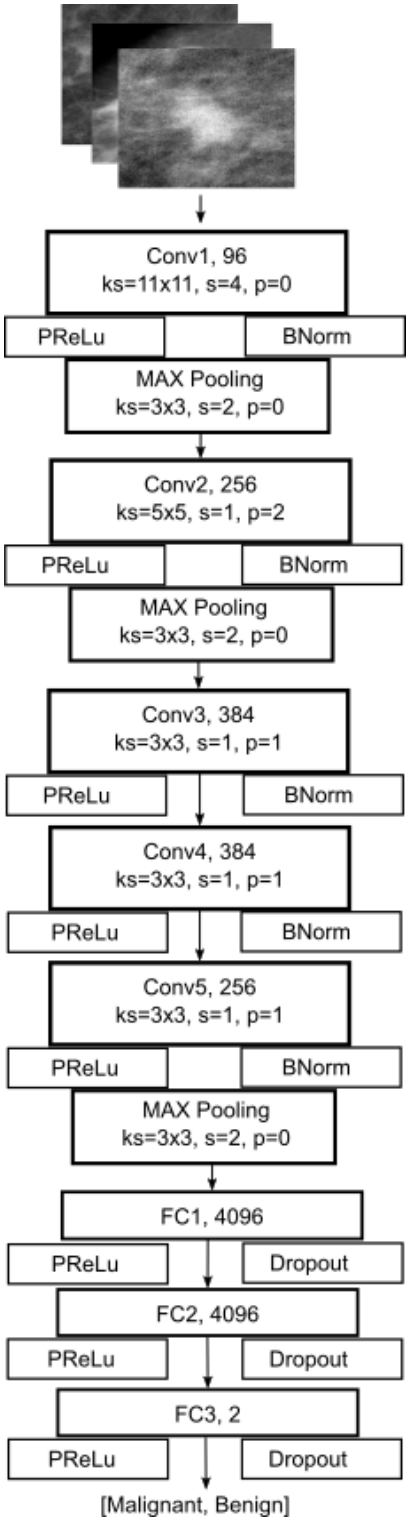

Fig. 4. The architecture of the proposed network.Note that ks, s, p represent kernel size, stride and padding, respectively.

AlexNet's weights are initialised based on the ImageNet dataset, and we fine-tuned it with our training dataset consisting of over 36,000 images. In this study we made an assumption that by using the knowledge of ImageNet dataset

features, the network is expected to classify breast masses with fewer samples and shorter training time. Once the training is completed, we perform model selection by taking the models which have the best three validation accuracies. Subsequently, each model is tested with unseen images and the final decision result is based on the average prediction probability of the three chosen models.

IV. Experimental Results

In this section we present experimental results based on the classification accuracy of the original AlexNet, the accuracy for each of the best three selected models and the accuracy for the ensemble model. Table I shows the experimental results of our study based on classification accuracy (ACC) and area under the curve (AUC).

TABLE I
QUANTITATIVE EXPERIMENTAL RESULTS. NOTE THAT MODEL A, B AND C ARE THE BEST THREE MODELS WITH THE HIGHEST VALIDATION ACCURACY DURING TRAINING.

Networks	ACC (%)	AUC
Original AlexNet [17]	65.5	0.71
Best model A	77.8	0.80
Best model B	76.5	0.78
Best model C	75.9	0.78
Ensemble model A, B and C	80.4	0.84

It can be observed that our proposed network (a modification of AlexNet) performed significantly better the original AlexNet by at least 10% in terms of classification accuracy and 0.07 in area under the curve. The original AlexNet produced $ACC = 65.5\%$ and $AUC = 0.71$. The proposed ensemble network produced $ACC = 80.4\%$ and $AUC = 0.84$, which is over 2% better than the accuracy of the best performing single model. Testing the performance of each best model resulted in accuracies of 77.8%, 76.5% and 75.9% for model A, B and C, respectively. In terms of AUC, model A produced 0.80, and model B and C each produced 0.78. The proposed network took approximately two hours to complete the training phase, covering over 36,000 images.

V. Discussion

Combining prediction results of several classifiers is a common approach in machine learning [26], [27]. The main advantage of this approach is that it prevents a single model from being exploited by outliers/noise/complicated cases. For example, the model might be prone to false positive/negative in some cases such as masses with lobulated shape which tend to be benign but in some cases are found to be malignant. Moreover, this approach prevents biased decisions from a single model, hence giving the opportunity of the system to consider other decisions from the other models. Therefore, the final decision boundary is based on a collection of models which is more noise tolerant. Figure 5 shows validation accuracies of the proposed network at different epochs (ϵ) during the training phase. The best three models A, B and C produced validation accuracies 83.5% ($\epsilon = 49$), 82.2% ($\epsilon = 40$) and 81.8% ($\epsilon = 46$), respectively.

In comparison to the other studies in the literature, several studies achieved around 90% accuracies. For example, Levy and Jain [15] reported that they achieved 89% and 91% classification accuracy for AlexNet and GoogleNet, respectively. Using their own netweork (LevyNet), they achieved a maximum of 60% accuracy. In another study, Hamidinekoo *et al.* [28] reported accuracy of 89% using different parameter settings on AlexNet and evaluated the effects of different data augmentation techniques. Shen [22] and Dhungel *et al.* [21] reported over 0.90 area under the curve value whereas the proposed method produced $AUC > 0.84$. Although the proposed method produced lower results, the purpose of our study is to initiate the use of deep learning approach in breast mass classification problem in DESIREE project. In medical image analysis, many studies [28], [3] have successfully used deep learning approach in disease classification such as prostate cancer, breast cancer and lung cancer. Therefore, with the following directions we are optimistic that we will achieve similar performance to radiologists.

- Modifying other existing networks such as VGGNet [29], GoogleNet [16] and ResNet [30]. As mentioned earlier, this paper presents our preliminary results in the DESIREE project using a deep learning based approach for breast mass classification which we started with Alexnet [17]. Many studies have reported that VGGNet, GoogleNet and ResNet outperformed Alexnet. This is due to deeper and more robust architecture which enables the network to learn more details of the image, resulting more discriminant features.
- Using a more robust ensemble network such as combining networks which are fine tuned based on different training datasets or combining different network architectures. For example a combination of modified AlexNet, GoogleNet and ResNet. Generally, the behavior of deep learning methods depend on their network architectures. This means, its architecture plays an important role in learning image representation and extracting important features of the object that can maximise the class boundary. Ensambling different network architectures such as AlexNet, GoogleNet, ResNet and VGGNet gives the opportunity to the system to extract more robust and diverse information from the breast mass, hence a greater opportunity to improve the performance of the classification results.
- Using different data augmentation techniques to increase the amount of data, hence exposing the network to learn more class characteristics. For example the study of [28] found that different augmentation techniques have significant effects on deep learning performance. Data augmentation increases the amount the data hence enriches obvious features that distinguishes one class from another. Moreover, it helps our network to be invariant to translation, viewpoint, size or illumination

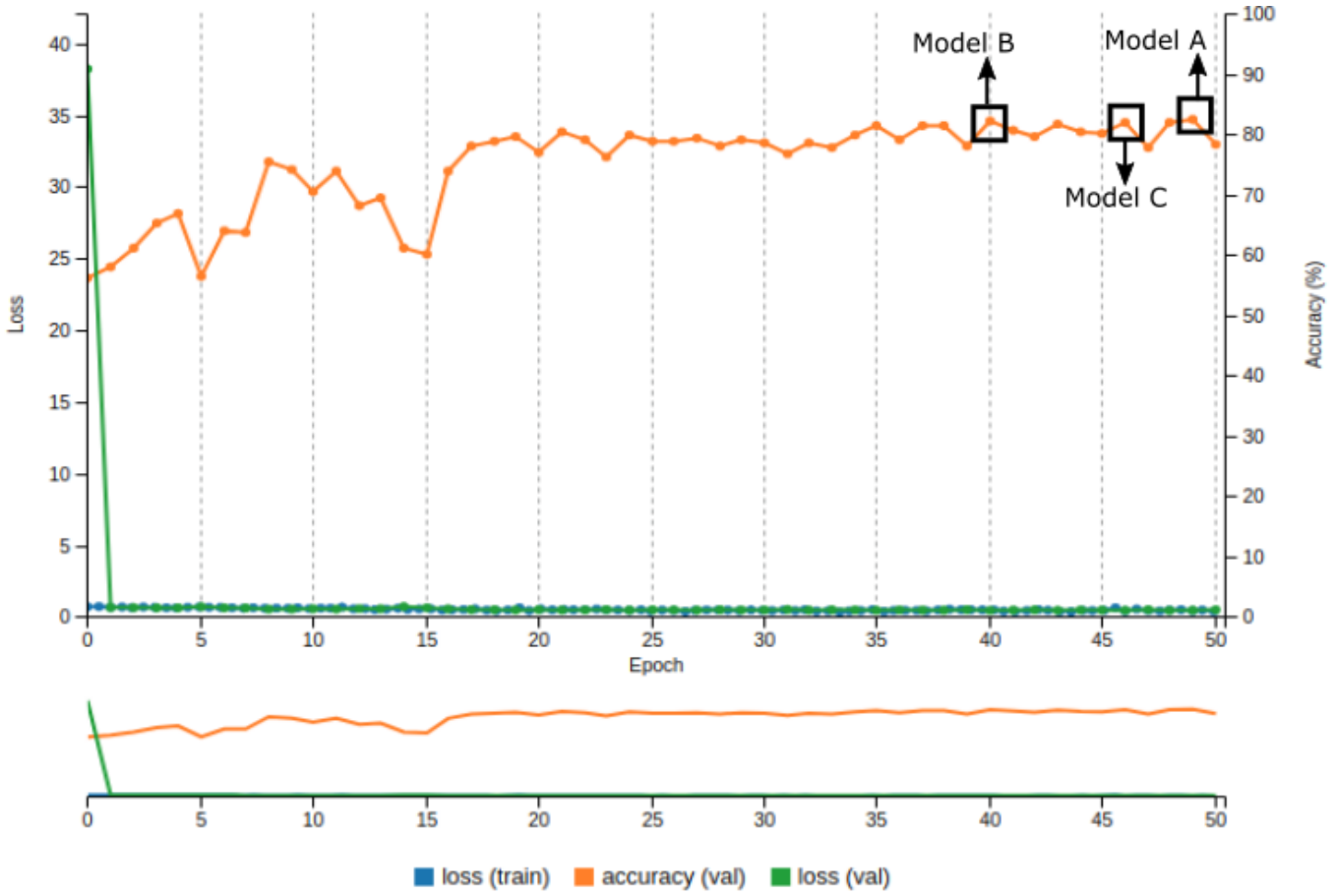

Fig. 5. Loss (green) and accuracy (amber) in the validation phase and loss (blue) in the training phase taken from DIGITS.

(or a combination of these). This means, the network is robust in recognising the same/similar objects that are taken at different translations or viewpoints, which could help the system to reduce false positives/negatives. Furthermore, performing data augmentation can also prevent our network from learning irrelevant patterns, essentially boosting overall performance.

- Our preliminary studys show that we achieved good results by performing simple modifications on AlexNet. We plan to increase the number of datasets using other publicly available datasets such as MIAS and InBreast as well as our own dataset from our clinical partners in DESIREE. This will provide diversity of the data in terms of feature representation and characteristics which can improve the overall architecture and performance of the network.

VI. Conclusion

In conclusion, we have presented our network which is a modification of AlexNet. Experimental results suggest that good results could be achieved with simple modifications. In this study we did not make significant architecture modifications to AlexNet but used different parameters with more sophisticated functions such as PReLu rather than ReLu. Although the performance of our proposed network does not yet outperform radiologists, we are optimistic that with further development the proposed model will achieve similar performance to radiologists. For example, the study of Levy and Jain [15] claimed that with few modifications on GoogleNet, they achieved recall rate 0.92 which is similar to radiologist performance.

Acknowledgment

This research was undertaken as part of the Decision Support and Information Management System for Breast Cancer (DESIREE) project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690238.

References

[1] C. E. DeSantis, J. Ma, A. G. Sauer, L. A. Newman and A. Jemal. Breast cancer statistics, 2017, racial disparity in mortality by state. CA: A Cancer Journal for Clinicians, 67, pp. 439–448, 2017.

[2] Breast Cancer Care. Facts and Statistics 2018. Accessed: 14 June 2018. Url: https://www.breastcancercare.org.uk/about-us/media/press-pack-breast-cancer-awareness-month/facts-statistics.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, vol. 42, pp. 60–80, 2017.

[4] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor and R. Zwiggelaar. Deep Learning in Mammography and Breast Histology, an Overview and Future Trends. *Medical Image Analysis*, vol. 47, pp. 45–67, 2018.

[5] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin. Curated breast imaging subset of DDSM. *The cancer imaging archive*, 2016.

[6] A. Rampun and P. J. Morrow and B. W. Scotney and J. Winder. Fully automated breast boundary and pectoral muscle segmentation in mammograms. *Artificial Intelligence in Medicine*, vol. 79, pp. 28–41, 2017.

[7] A. Rampun and B. W. Scotney and P. J. Morrow and H. Wang and J. Winder. Breast Density Classification Using Local Quinary Patterns with Various Neighbourhood Topologies. *J. Imaging*, vol. 4, 2018.

[8] A. Rampun, P. J. Morrow, B. W. Scotney and R. J. Winder. Breast density classification in mammograms using local ternary patterns. *Proc. of the International Conference Image Analysis and Recognition ICIAR 2017*, pp. 463–470, 2017.

[9] A. Rampun, H. Wang, B. Scotney, P. Morrow and R. Zwiggelaar. Confidence analysis for breast mass image classification. International Conference on Image Processing (ICIP), 2018 (in press).

[10] A. Rampun, H. Wang, B. Scotney, P. Morrow and R. Zwiggelaar. Classification of mammographic microcalcification clusters with machine learning confidence levels. 14th International Workshop on Breast Imaging (IWBI 2018) 10718, 107181B, 2018.

[11] C. Muramatsu, T. Hara, T. Endo and H. Fujita. Breast mass classification on mammograms using radial local ternary patterns. *Computers in Biology and Medicine*, vol. 72 (1), pp. 43–53, 2016.

[12] F. Narváez and E. Romero. Breast Mass Classification Using Orthogonal Moments. Proc. of International workshop of digital mammography. IWDM 2012: Breast Imaging pp. 64–71, 2012.

[13] Y. A. Reyad, M. A. Berbar and M. Hussain. Comparison of statistical, LBP, and multi-resolution analysis features for breast mass classification. *J Med Syst.* 38(9):100, 2014. doi: 10.1007/s10916-014-0100-7.

[14] J. Y. Choi and Y. M. Ro. Multiresolution local binary pattern texture analysis combined with variable selection for application to false-positive reduction in computer-aided detection of breast masses on mammograms. *Physics in Medicine & Biology*, Vol. 57 (21), 2012.

[15] D. Lévy and A. Jain. Breast mass classification from mammograms using deep convolutional neural networks, 2016. https://arxiv.org/abs/1612.00542

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich. Going deeper with convolutions. *In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19, 2015.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *In Proc. of the 25th International Conference on Neural Information Processing Systems (NIPS)*, pp. 1097-1105, 2012.

[18] I. Bakkouri and K. Afdel. Breast tumor classification based on deep convolutional neural networks. *In Proc. 3rd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2017.

[19] Z. Jiao, X. Gao, Y. Wang, J. Li. A deep feature based framework for breast masses classification. *Journal of Neurocomputing*, vol. 197, pp. 221-231, 2016.

[20] J. Arevalo, F.A. Gonzlez, R. Ramos-Pollán, J. Oliveira, M.A. Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks, *Comput. Methods Program Biomed*, vol. 127, pp. 248-257, 2016.

[21] N. Dhungel, G. Carneiro, A.P. Bradley. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical Image Analysis*, vol.37, pp. 114–128, 2017.

[22] L. Shen. End-to-end Training for Whole Image Breast Cancer Diagnosis using An All Convolutional Design, 2017. arXiv:1708.09427 [cs, stat].

[23] D. P. Kingma and J. L. Ba. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG], December 2014.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In ICCV, 2015.

[25] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In ICML, 2010.

[26] A. Rampun, B. Tiddeman, R. Zwiggelaar and P. Malcolm. Computer aided diagnosis of prostate cancer: A texton based approach. *Medical physics*, vol. 43 (10), pp.5412–5425, 2016

[27] A Rampun and L. Zheng, P. Malcolm and B. Tiddeman and R. Zwiggelaar. Computer-aided detection of prostate cancer in T2-weighted MRI within the peripheral zone. *Physics in Medicine & Biology*, vol. 61 (13), pp.4796, 2016

[28] A Hamidinekoo, Z Suhail, T Qaiser and R Zwiggelaar. Investigating the Effect of Various Augmentations on the Input Data Fed to a Convolutional Neural Network for the Task of Mammographic Mass Classification. In Proc. Annual Conference on Medical Image Understanding and Analysis, pp. 398–409, 2017.

[29] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv technical report, 2014.

[30] K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition, 2015. arXiv:1512.03385 [cs.CV]