**Paper review 8DM50 Deep learning in medical imaging and biology, Sjoerd Vossen**  
*N. Wu et al., "Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening," in IEEE Transactions on Medical Imaging, vol. 39, no. 4, pp. 1184-1194, April 2020, doi: 10.1109/TMI.2019.2945514.*

Summary of the application domain

The paper focuses on the application of deep convolution neural networks (CNNs) in the domain of breast cancer screening and classification. Breast cancer is the second leading cancer-related cause of death among women in the US. Mammography is the only imaging test that has reduced breast cancer mortality. However, there has been discussion regarding the false positive recalls and associated false positive biopsies. 10-15% of the women is asked to return as a consequence of a inconclusive screening mammogram. Only 10-20% are recommended to undergo a needle biopsy for further work up. Among these, only 20-40% yield a diagnosis of cancer. Hence, there is an unmet need to shift the balance of routine breast cancer screening towards more benefit and less harm.

They contribute to the development of neural networks to support radiologists in interpreting breast cancer screening exams by

1. Introducing a novel two-stage neural network for incorporating global and local information with an appropriate training procedure
2. Demonstrate the feasibility of training and evaluating the network with over 1,000,000 high-resolution mammographic images
3. Using a building block their network, they propose a novel variant of a ResNet specifically designed for medical imaging, which has a balance of depth and width that allows the model to process a very large image while maintaining reasonable memory consumption
4. Evaluating the utility of pretraining the network using a related task with a more noisy outcome (screening BI-RADS classification) and find it to be a very important part of the pipeline that markedly improves the performance of their models

Summary of machine learning methodology and evaluation metrics

Left craniocaudal (L-CC) and right craniocaudal (R-CC) representations and left mediolateral oblique (L-MLO) and right mediolateral oblique (R-MLO) representations are concatenated. This makes separate predictions for CC and MLO views with are averaged during inference. This model is based on the ResNet architecture that output a fixed-dimension hidden representation for each mammography view, and two fully connected layers to map the computed hidden representations to the output predictions. They train the model by optimizing the loss function that includes the binary cross-entropy. After training they use the AUC (are under the ROC curve) for malignant/not malignant and benign/not benign classification tasks on the breast level. The ROC curve summarizes the trade-off between the true positive rate and false positive rate for a model using different probability thresholds. The precision-recall curve summarizes the trade off between the true positive rate (recall) and the positive predictive value (precision) for a model using different probability thresholds.

Strengths and weaknesses of the methodology and evaluation metrics

Strengths:

* Use of concatenated representations: consider information from multiple views
* ResNet architecture: provides the model with the ability to capture complex hierarchical features
* Binary cross-entropy loss function: appropriate for binary classification tasks
* Using AUC for evaluation: is an appropriate evaluation metric for this type of model

Weaknesses:

* Limited interpretability: the ResNet architecture makes it often challenging to understand the specific features that contribute to the model’s decision-making
* Data imbalance issues: binary cross-entropy assumes a balanced dataset in the number of malignant and benign cases
* Sensitivity to hyperparameters
* Limited exploration of other metrics: AUC-ROC does not take into account other evaluation metrics such as sensitivity, specificity, precision and recall.

Alternative methodology, evaluation metrics, and ideas for improvement

This model could be improved with the use of Generative Adversarial Networks (GANs). GANs can generate realistic mammography images, serving as an effective data augmentation technique. By enlarging the dataset, the model’s ability to generalise to various imaging conditions can be improved. This can also be done by varying pathological conditions, so the model gets provided with a more diverse set of training examples.

When using GANs, other evaluation metrics become important, such as an inception score (IS). This IS measures the quality and diversity of the generated images. Another evaluation metric is the Frechet Inception Distance (FID) which quantifies the similarity between generated and real images.