LOW RESOURCE BREAST CANCER DETECTION WITH MAMMOGRAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning has evolved in healthcare and gone further to have a full clinical deployment. We use recent advancements in breast cancer detection to answer the following questions:

- Is it better to use pretrained models on natural images with medical applications or pretrained models on medical images?
- Which models will transfer the required knowledge to the new medical target tasks?
- Or using both pretrained models as feature extractor will help in the detection?

We experiment with the INBreast dataset of 410 mammograms. We test two neural networks pretrained on (ImageNet and NYU v1.0 dataset of mammograms). The results were not sufficient to fully answer the first two questions; the INBreast dataset has few images, and this leads to rapid overfitting. However, when we use both models as feature extractors, apply an oversampling technique for malignant cases and then classify with a linear SVM, the performance metrics outperform the deep neural network fine tuning results. And the NYU model as a feature extractor outperforms Resnet50 (pretrained on ImageNet) extracted features.

1 INTRODUCTION

Breast cancer is the most common type of cancer among women in the USA (Atlanta, 2019) and worldwide. In Egypt, Ibrahim et al. (2014) found that the breast cancer rate is 32.0% while liver cancer comes the second with 13.5% among women when analyzing cancer types across different parts of Egypt. Other countries in Africa suffer from the same problem: a high number of breast cancer cases. There are approximately 94,378 diagnosed breast cancer cases in sub-Saharan Africa annually(Ltd, 2016; Adeloye et al., 2018). All of these statistics show that breast cancer is a worldwide problem, and it motivated us to practically contribute to solutions with the recent advancements in deep learning.

Deep learning has been growing in the field of medical imaging and diseases prediction systems. It has proven that it can push the automatic diagnosis systems forward and provide more reliable and helpful assistance tools for doctors and radiologists. Three success stories where deep learning excelled in radiology applications are: identifying skin cancer from dermatologist level photographs(Esteva et al., 2017), automatic detection of diabetic retinopathy on a publicly available dataset of retinal fundus images(Abràmoff et al., 2016) and classification and detection of thorax diseases from chest x-ray images(Wang et al., 2017).

Section two provides an overview of the previous work that we are using in our study. Section three explains the two methodologies we use to answer the questions of interest. Section four defines the dataset we are using. Finally we put the experiments results and discuss them in the last two sections.

2 PREVIOUS WORK

Recently, researchers at New York University (NYU) published a new methodology using deep learning to detect breast cancer from mammograms(Wu et al., 2019b). The study is based on a recently curated dataset (NYU Breast Cancer Screening Dataset v1.0) that was reported in (Wu et al., 2019a). The dataset was not publicly released, but the report explained in detail every step to curate the dataset and make it ready for deep learning training process. It consists of 229,426 digital screening mammography exams (1,001,093 images) from 141,473 patients screened between 2010 and 2017 at **NYU Langone Health**. Each exam has four images of the breasts (each breast with two views, CC and MLO, as shown in Figure 1. The dataset has three labels: a breast level label, a pixel level label (from biopsied findings) and an exam level BIRADS label[1], a number from one to six indicating the initial diagnosis of the radiologist after screening a mammography (1 indicates low risk while 6 indicates highest risk). The technical report is rich in information to any research center or hospital which needs to start using AI in their screening and wants to prepare data for deep learning. They start by excluding images that do not look like a normal mammogram, and the images are segmented by radiologists indicating benign or malignant findings according to biopsies. Then they cropped the images to exclude the background of the breast image.

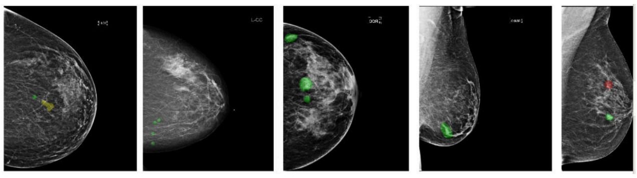

Figure 1: Segmented regions by radiologists were used in the NYU model training [red: malignant finding, green: benign finding, yellow: high risk benign finding]. The first three images from the left are in the CC view, while the last two images are in the MLO view. Source:(Wu et al., 2019a)

Wu et al. (2019b) use deep learning with the aforementioned NYU v1.0 dataset. They train the dataset in a two-stage training procedure: breast-level and pixel-level training tasks. In the breast-level training (they call it the image-only model), they use a down-scaled version of the mammogram and run it through four Convolutional Neural Networks (CNNs) (one CNN for each breast view) because the computation resources limited their ability to train with high resolution images. The second stage is a patch-level (i.e. pixel-level) training task. This stage uses a sliding window/patch of size 256×256 to classify (presence/absence of a malignant or benign finding) according to the provided pixel-level annotation of the findings by the radiologists. The second stage produces heatmaps of the original mammograms (i.e. green patches for benign findings and red patches for malignant findings), they tested with CNNs which take these heatmaps (for both findings) as additional two channels to the low resolution mammogram images in the training phase and called this model (image + heatmap). They concluded that using a hybrid model (the average of the radiologists and deep learning model predictions) is more accurate than using any one of them alone (one of the author wrote a blog post (Phang) about it for more details).

This work was the motivation of this report, their network helps in the cancer detection and achieves an AUC of 0.895 in predicting whether there is a cancer in the breast. The question was: can we extend the publicly released model by (Wu et al., 2019b) to other smaller datasets? or will it be useful only for the patients in **NYU Langone Health**?

3 Methodology

3.1 Transfer Learning

Since most of the successful new deep learning models in breast cancer detection that were published after (Wu et al., 2019b) were based on the NYU v1.0 non-public dataset, we seek to answer the previously mentioned question "can we extend the publicly released model, by (Wu et al., 2019b), trained on mammograms only, to other smaller mammogram datasets? would a pretrained network on natural images (e.g. ImageNet dataset) perform better than the former model? . We fine-tune pretrained models on ImageNet and NYU v1.0 datasets (for classifying single image) to see which model will generalize to unseen smaller mammogram datasets. We also used both pretrained networks as feature extractor then used Support Vector Machines (SVM) classifier with different kernels. We used the common performance metrics for image classification task in our experiments.

[1] https://breast-cancer.ca/bi-rads/

3.2 Feature Extractors

We extend our way of addressing the problem to apply traditional machine learning algorithms. We extract the features of the INBreast dataset of 410 image from both models (NYU and ImageNet) and use these features to train Logistic Regression (LR) and Support Vector Machine (SVM) classifiers with different kernel functions. We evaluate on the INBreast binary classification task (i.e. BIRADS 1,2 and 3 are benign and BIRADS 4,5 and 6 are malignant). We use SVM and LR with reduced number of features by using Lasso feature selection method. When we use the NYU model we evaluate against NYU L-CC feature extractor only and an average of all four extractors used to train the NYU model. We evaluate SVM with two different kernel functions (Linear and RBF) . We apply a 10-fold cross validation of the balanced accuracy score. We over-sample the malignant cases using the Synthetic Minority Over-sampling Technique (SMOTE)(Chawla et al., 2002).

4 Dataset

We aim to answer the question of interest with the available public datasets of mammograms. It is surprising that there are few of them. There was another factor for choosing the dataset: the similarity to the NYU dataset in terms of the resolution and intensity of the breast tissues in the images.

INBreast dataset (Moreira et al., 2012) was the first option in terms of similarity to the NYU dataset. The dataset includes full-field digital mammograms and consists of 115 cases (410 mammogram images) from which 90 cases have both breasts affected (four images per case) and 25 cases have (two images per case). However, it is still a small number of images to train deep learning models from, which require a larger number of data points to produce reliable results. Figure 2 shows an example of a mammogram with BIRADS 6, the cropped version as in the pre-processing of the NYU's deep learning model and the generated heatmaps (benign and malignant) when using the (image+heatmap) NYU's model.

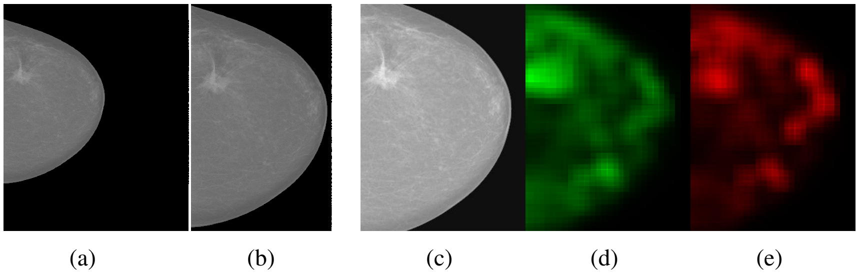

Figure 2: INBreast Mammogram Example of BIRADS 6. (a) CC View. (b) Cropped CC-view. (c) NYU Input. (d) Benign Heatmap. (e) Malignant Heatmap

5 Experiments and Results

First, we used the NYU model for inference on the INBreast dataset, without any fine-tuning. More specifically, we used the single image prediction which is one of the Resnet-22 feature extractors to predict four labels (benign, not benign, malignant and not malignant). The reason why the NYU model uses four labels instead of binary labels (benign/malignant) is that a breast can have more than finding with different labels for each and having four labels is a kind of alleviating the misinterpretation for multiple findings. Figure 3 shows the distribution of how many data points in INBreast are predicted as benign or malignant (we compare the two label values of 'malignant' and 'benign' instead of choosing a threshold to determine whether the prediction is malignant or benign).

(Raghu et al., 2019) studied the effect of transfer learning in the medical domain. They concluded that using large pretrained networks on natural images (e.g. ImageNet) works as well as a small network trained from scratch on medical images. They experimented with a few medical tasks where the dataset has few hundred thousand images. Here, we are trying to extend their conclusion; is it also useless to fine-tune large networks trained on ImageNet as compared to using small networks trained-from-scratch on medical images?.

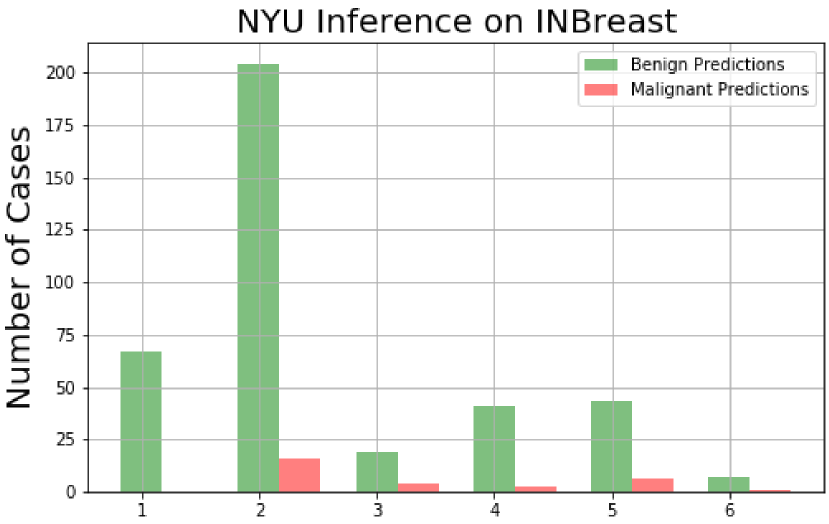

Figure 3: Inference results of using NYU single-image model on the INBreast dataset. Each number indicates a BIRADS number. Each BIRADS number has two bars indicating the number of predicted cases as benign (green bar) and the number of detected malignant cases (red bar). The model predicts most of the images as benign (this is more beneficial for BIRADS 1,2 and 3 than BIRADS 4, 5 and 6 which indicate high risk of having a malignant tumor.

INBreast dataset has accurate BIRADS labels but the biopsy results are not known; that is why we experimented with both tasks: BIRADS classification and binary classification (assuming BIRADS 1,2 and 3 are b benign cases and BIRADS 4, 5 and 6 as malignant cases). We used few pretrained models on ImageNet (resnet50, resnet18 and squeeznet) besides the NYU model. Table 1 shows different performance metrics across the different pretrained models and settings (the stage indicates which phase in the training process, e.g. using as feature extractor or fine tuning, the size of the batch and the progressive resizing setting).

The results of feature extractor experiments are illustrated in Table 2. We evaluate on the INBreast binary classification task (i.e. BIRADS 1,2 and 3 are benign and BIRADS 4,5 and 6 are malignant). We apply a 10-fold cross validation of the balanced accuracy metric and report their mean and standard deviation in the last two columns. Finally, We see a boost in performance metrics when we over-sample the malignant cases using the SMOTE.

Model	Pretrained on	Task	Accuracy	AUROC	Precision	Recall
resnet50	ImageNet	BIRADS	0.549	0.475	0.513	0.481
resnet18	ImageNet	BIRADS	0.573	0.390	0.686	0.524
squeeznet	ImageNet	BIRADS	0.427	0.459	0.270	0.210
resnet50	**ImageNet**	**Binary**	**0.866**	**0.9**	**0.737**	**0.7**
resnet18	ImageNet	Binray	0.854	0.872	0.75	0.6
squeeznet	ImageNet	Binary	0.732	0.695	0.437	0.35
NYU (L-MLO)	NYU v1.0	Binary	0.780	0.544	0.625	0.25
NYU (R-CC)	**NYU v1.0**	**Binary**	**0.793**	**0.581**	**0.667**	**0.3**

Table 1: Fine tuning both Resnet50 pretrained on ImageNet and the NYU network trained on the NYU v1.0 dataset. Common performance metrics are in the last four columns.

[HTML]C0C0C0 Model	balanced	Feature Extractor	# Features	Accuracy	AUC	mean	std
SVM+Linear	No	ImageNet Resnet50	2048	0.6544	0.562	0.58	0.09
SVM+Linear	No	ImageNet Resnet50	217	0.8676	0.8362	0.74	0.11
SVM+RBF	No	ImageNet Resnet50	217	0.75	0.5143	0.51	0.03
LR	No	ImageNet Resnet50	217	0.8676	0.8362	0.76	0.12
SVM+Linear	**SMOTE**	**ImageNet Resnet50**	217	**0.9317**	**0.9345**	**0.94**	**0.03**
LR	SMOTE	ImageNet Resnet50	217	0.9122	0.9157	0.95	0.04
SVM+Linear	No	NYU L-CC	4096	0.6912	0.5680	0.56	0.07
SVM+RBF	No	NYU L-CC	4096	0.7426	0.5	0.50	0.03
SVM+Linear	No	NYU L-CC	350	0.8750	0.8412	0.74	0.10
SVM+RBF	No	NYU L-CC	350	0.7426	0.5	0.50	0.03
LR	No	NYU L-CC	350	0.8456	0.8027	0.72	0.11
SVM+Linear	SMOTE	NYU L-CC	350	0.9610	0.9631	0.96	0.02
LR	SMOTE	NYU L-CC	350	0.9317	0.9361	0.94	0.04
SVM+Linear	No	NYU Avg	325	0.8162	0.7269	0.73	0.07
LR	No	NYU Avg	325	0.8162	0.7082	0.72	0.08
SVM+Linear	**SMOTE**	**NYU Avg**	325	**0.9756**	**0.9758**	**0.97**	**0.04**
SVM+RBF	SMOTE	NYU Avg	325	0.6683	0.6733	0.72	0.07
LR	SMOTE	NYU Avg	325	0.9610	0.9623	0.97	0.03

Table 2: Results of the binary task classification of the INBreast dataset (i.e. BIRADS1,2 and 3 are benign and BIRADS4,5 and 6 are malignant). The used models are SVMs with two different kernel functions (Linear and RBF) and Logistic Regression (LR). Lasso feature selection is used and the number of used features is stated in the 4th column. A 10-fold cross validation of the balanced accuracy is depicted in the last two columns (the mean and standard deviation of the 10 folds). SMOTE was used as an oversampling technique for the malignant class. We use both one feature extractor from NYU model and an average over the four extractors.

6 Discussion

Figure 3 shows that we can not use the NYU model for inference even if we think the target dataset was close in distribution to the NYU v1.0 dataset; because it miss-classifies most of the high risk cases as benign ones. Table 1 shows that using pre-trained models on ImageNet is still effective in this case. However, we can not conclude that the network is robust and can be used for inference with other datasets. Resnet50 with ImageNet performs better that Resnet18 and Squeeznet which is something we need to investigate more by visualizing the learnt features and measuring the similarity of the networks. We need to apply the same methodology to other datasets and see if they have similar results, only then we can make conclusions.

On the other hand, the extracted features from both models seem to be linearly separable in a high dimension but probably lower than the total number of extracted features by a factor of 0.1. Linear SVM with SMOTE oversampling outperforms the finetuning experiments results as well as an RBF SVM classifier. Comparing the two models (trained on ImageNet and NYU v1.0 datasets) it seems the latter has the capacity to extract more meaningful features than the ImageNet model which is reflected in the difference in their performance with the linear SVM.

7 Conclusion and Future Work

It is not clear whether using a smaller networks trained on medical images from scratch would be useless because we have tested with a small dataset. We may need traditional machine learning algorithms if we think of small scale healthcare institutions. We might need to see how radiologists are trained to interpret such images so that we can mimic the process: we need few-hundred learning techniques to study the effect of transfer learning with a dataset of the INBreast size. Also, studying when transfer learning is useful, redundant or harmful is highly required.

In the future, we may head to other techniques that try narrowing down the gap between the source task and target task such as domain adaptation techniques to extend the usability to other different datasets of mammograms. There is a final note about the behaviour of the transfer learning in the performed experiments: if we are interested in detecting cancer cells in the human body through any medical image type, maybe we need to collect an ImageNet-like medical dataset of various types of medical images and different body parts which share similar characteristics of cancer cells.

REFERENCES

Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science*, 57(13): 5200–5206, 2016.

Davies Adeloye, Olaperi Y Sowunmi, Wura Jacobs, Rotimi A David, Adeyemi A Adeosun, Ann O Amuta, Sanjay Misra, Muktar Gadanya, Asa Auta, Michael O Harhay, et al. Estimating the incidence of breast cancer in africa: a systematic review and meta-analysis. *Journal of global health*, 8(1), 2018.

GA Atlanta. American cancer society: Cancer facts and figures. 2019. URL https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Amal S Ibrahim, Hussein M Khaled, Nabiel NH Mikhail, Hoda Baraka, and Hossam Kamel. Cancer incidence in egypt: results of the national population-based cancer registry program. *Journal of cancer epidemiology*, 2014, 2014.

Elsevier Ltd. How advanced is breast cancer in africa, December 2016. URL https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(16)30283-2/fulltext.

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*, 19(2):236–248, 2012.

Jason Phang. Deep neural networks improve radiologists' performance in breast cancer screening.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, pp. 3342–3352, 2019.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Huang, et al. The nyu breast cancer screening dataset v1.0, 2019a. URL https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf.

Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzębski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists' performance in breast cancer screening. 2019b.