Radiologist-level breast cancer detection for clinical practice

Florian Hillen and Nikhil Mallareddy
MIT, Institute for Data, Systems, and Society

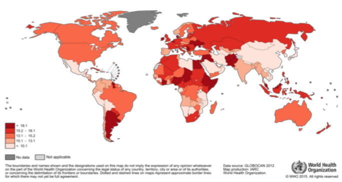
CS109b
Spring 2018

Executive summary: We develop a VGG16-based deep learning algorithm for classifying pathological versus non-pathological mammograms. We set the algorithmic threshold of our final model as 0.17 prioritizing sensitivity over specificity to reduce false negative cases in medical practice and increase women's life expectancy through early detection.

INTRODUCTION

1 in **8** **WOMEN**

Will be diagnosed with breast cancer in their lifetime[1]



40,000 Women die of breast cancer in the US every year[2]

- Early detection can strongly increase the survival rate[3]
- Mammography is the most common screening tool for breast cancer, identifying masses and calcifications
- Radiologists are prone to make diagnostic errors causing significant risks for the women[4]

DATA EXPLORATION

- Digital Database for Screening Mammography(DDSM)[4]
- 2,620 cases split into 10,713 patches
- Information about masses, calcifications, and non-pathological images[5]
- Labels: "benign", "malignant", "benign without callback", "unproven" and "normal"

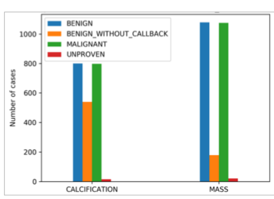


Figure 1. Classes and labels of the DDSM dataset

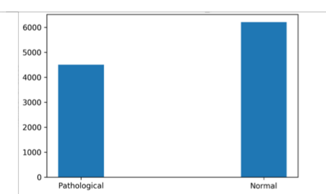


Figure 2. Split of the DDSM dataset

Objective

→ Model a deep learning algorithm as a support tool for radiologists to decrease false negative and false positive diagnoses.

METHODOLOGY

- Binary classification for normal vs. pathological (combining masses and calcifications)

1. Comparing different architectures
2. Fine tuning best architecture by pre-processing to meet benchmark of an AUC of 85% of Shen, L. (2017)[5]
3. Analyze clinical relevance

RESULTS – MODELS

Model[6]	Batch size	Epochs	Special pre-processing	Accuracy on test set
Simple model	32	15	no	0.759
MobileNet	32	15	no	0.777
ResNet50	32	15	no	0.751
VGG16	32	15	no	0.819
VGG16 + Augmentation	32	30	Flips, shifts, rotations	0.808
Final model				
VGG16 + ImageNet	32	15	Pretrained on ImageNet	0.869

Table 1. Comparison of different architectures and parameters

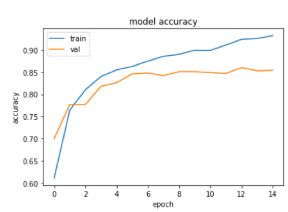


Figure 3. Loss function of final model

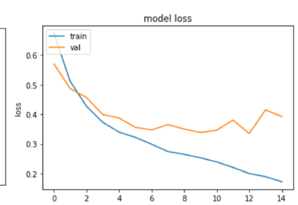


Figure 4. Accuracy of final model

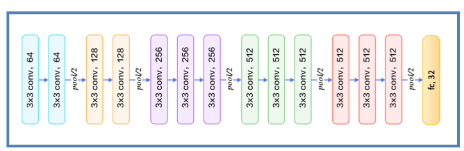


Figure 5. Final customized VGG16-Model

RESULTS – CLINICAL RELEVANCE

Risks and costs of an error	
False positive	• Additional test: Costs and minimal-invasive biopsy • Short-term distress/long-term risk of anxiety[7]
False negative	• 5-year survival rate is strongly impacted by later detection: Decreases from 93% to 72% from stage III to stage II[8]

→ From a clinical point of view, having high sensitivity is more important[9]

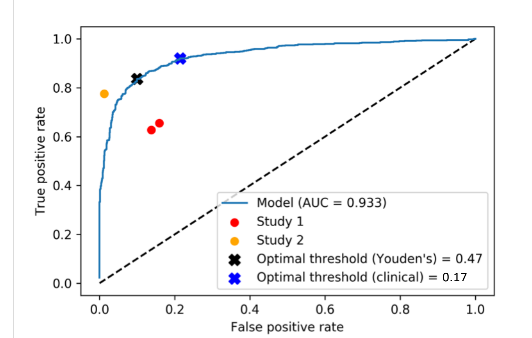


Figure 6. Comparison of final model's AUC with radiologists from study one[10] and two[11] and the mathematically optimal threshold vs. our clinically relevant threshold obtained through our risk evaluation

CONCLUSIONS

- Our best performing model was a customized VGG16 architecture pre-trained on ImageNet with an accuracy of ~87% and an AUC of 0.933
- Clinically, higher sensitivity outweighs the costs of lower specificity, resulting in a lower threshold value (0.17) compared to the mathematically optimal (0.47)

CITATIONS AND LINKS

1. McGuire, A., Brown, J. A. L., Malone, C., McLaughlin, R., & Kerin, M. J. (2015). Effects of Age on the Detection and Management of Breast Cancer. *Cancers, 7*(2), 908–929. http://doi.org/10.3390/cancers7020815
2. https://seer.cancer.gov/statfacts/html/breast.html
3 Sam Shapiro, Wanda Venet, Philip Strax, Louis Venet, Ruth Roeser; Ten- to Fourteen-Year Effect of Screening on Breast Cancer Mortality, *JNCI: Journal of the National Cancer Institute*, Volume 69, Issue 2, 1 August 1982, Pages 349–355, https://doi.org/10.1093/jnci/69.2.349
4 R. Heath, M., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, P. (2000). The digital database for screening mammography. *Digital mammography*, 431-434.
5 Shen, L. (2017). End-to-end Training for Whole Image Breast Cancer Diagnosis using An All Convolutional Design. *arXiv preprint arXiv:1708.09427*.
6 The model architectures have been customized from us for this task. Data-split was 75%/10%/15% for training, validation and test set.
7 Aro, A. R., Absetz, S. P., van Elderen, T. M., van der Ploeg, E., & van der Kamp, L. T. (2000). False-positive findings in mammography screening induces short-term distress—breast cancer-specific concern prevails longer. *European Journal of Cancer, 36*(9), 1089-1097.
8 https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html
9 Interview with a Gynecologist, 2nd year resident (April 28th)
10 Rafferty, E. A., Park, J. M., Philpotts, L. E., Poplack, S. P., Sumkin, J. H., Halpern, E. F., & Niklason, L. T. (2013). Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: results of a multicenter, multireader trial. *Radiology, 266*(1), 104-113.
11 Kolb, T. M., Lichy, J., & Newhouse, J. H. (2002). Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology, 225*(1), 165-175.