



Multi-task Learning for Detection and Classification of Cancer in Screening Mammography

Maria V. Sainz de Cea¹, Karl Diedrich¹, Ran Bakalo², Lior Ness²,
and David Richmond¹(✉)

¹ IBM Watson Health, Cambridge, MA 02142, USA
david.richmond@ibm.com

² IBM Research, Haifa University Campus, Mount Carmel, 3498825 Haifa, Israel

Abstract. Breast screening is an effective method to identify breast cancer in asymptomatic women; however, not all exams are read by radiologists specialized in breast imaging, and missed cancers are a reality. Deep learning provides a valuable tool to support this critical decision point. Algorithmically, accurate assessment of breast mammography requires both detection of abnormal findings (object detection) and a correct decision whether to recall a patient for additional imaging (image classification). In this paper, we present a multi-task learning approach, that we argue is ideally suited to this problem. We train a network for both object detection and image classification, based on state-of-the-art models, and demonstrate significant improvement in the recall vs no recall decision on a multi-site, multi-vendor data set, measured by concordance with biopsy proven malignancy. We also observe improved detection of microcalcifications, and detection of cancer cases that were missed by radiologists, demonstrating that this approach could provide meaningful support for radiologists in breast screening (especially non-specialists). Moreover, we argue that this multi-task framework is broadly applicable to a wide range of medical imaging problems that require a patient-level recommendation, based on specific imaging findings.

Keywords: Decision support · Deep learning · RetinaNet · ResNet

1 Introduction

Breast cancer is the leading cause of cancer death in women world wide [19]. Screening aims to increase early detection by identifying suspicious findings in asymptomatic women, and has been shown to reduce the risk of dying from breast cancer [17]. However, radiologists struggle to keep up with the volume of breast screening, increasing the risk of burn-out and missed cancer [13]. Thus, computer algorithms that can assist radiologists in reading mammography exams have the potential for a significant impact on women's health.

The task of screening mammography is to decide whether or not to recall a patient for additional work-up. This clinical decision is made on the basis of abnormal findings within the breast, but may also be influenced by the patient’s risk profile, which can be inferred from the mammogram’s overall appearance [21]. Despite the dual nature of this problem, requiring both object detection and image classification, and the vast literature on breast screening decision support, we are aware of only one publication utilizing multi-task learning (MTL) [9].

Previous methods based on image classification have produced competitive results [22], somewhat surprisingly, given the fact that suspicious findings may occupy only 1% of a mammogram. Detection-based methods, on the other hand, have the clear advantage that they are trained on highly discriminative findings, and have demonstrated good results on detection of masses and calcifications [1, 2, 7]. However, local annotations are inherently subjective, and are typically not as consistent as outcome-based ground truth used by classification. Furthermore, we observe that detection-only approaches potentially overlook ancillary findings, which can be unlabeled. There are many ancillary findings that have significant implication for determining malignancy, such as skin thickening, nipple retraction, neovascularity, adenopathy and multifocal disease.

A few notable publications have combined image classification and detection. In [8, 12], the authors initialize a classification model from a patch-based model pretrained on locally annotated data. However, weight sharing in this approach is sequential, offering much less flexibility than MTL. In addition, similar to classification methods mentioned above, these models can only provide indirect evidence for their decision via saliency maps. In [4], the authors address detection, segmentation and classification; however, similar to Mask R-CNN [5] and RetinaMask [3], their method applies only to detection ROIs, and does not address classification of the full image, a central focus of our paper. In [20], the authors input heatmaps from a sliding-window classifier as an additional channel to a study-level classification model, whereas in [18], the authors combine a sliding-window classifier and image classifier using a Random Forest, and in [14], the authors employ ensemble averaging over multiple classification and detection models; however, none of these approaches benefit from weight sharing across tasks. The paper that is methodologically most similar to ours is [9]. This work applies MTL of segmentation and image classification; however, they achieve a smaller performance improvement, and only evaluate their model on DDSM.

In this paper, we combine both image classification and object detection in a single multi-class MTL framework to derive the benefits of strong outcome-based ground truth information, global image features, and highly discriminative, albeit possibly noisy, local annotations. Using this flexible approach, we observe a significant performance boost in classification of malignant vs non-malignant images, and an improvement in detection of malignant calcifications.

2 Methods

2.1 Baseline RetinaNet Detector

We use the state-of-the-art RetinaNet model [11] as our baseline detection algorithm. RetinaNet is a single-shot object detector that utilizes a novel focal loss to counteract background-foreground imbalance, and has been used for object detection in several fields, including medical [7, 23]. The overall architecture is composed of a ResNet34-Feature Pyramid Network (FPN) backbone, and two sub-networks performing (i) classification, and (ii) coordinate regression, for each of the candidate detections.

2.2 Proposed Multi-task Algorithm

Building from the baseline RetinaNet model, we added an image classification subnet to ResNet34, to perform full image classification. In this way, the ResNet34 weights were shared between the image classification and detection tasks (Fig. 1). The image classification subnet matches the architecture of the published ResNet34 classification model [6], consisting of global pooling followed by a fully-connected layer. We experimented with additional Conv blocks followed by multiple fully-connected layers; however, it did not further improve performance. The final loss is a combination of categorical cross-entropy for image classification, and the focal loss and regression loss for object detection. We used a relative weighting factor, λ , to balance the two tasks. Best results were obtained for $\lambda = 0.2$, and $\gamma = 2$ and $\alpha = 0.5$ in the focal loss [11].

$$L_{\text{multi-task}} = L_{\text{focal}} + L_{\text{regression}} + \lambda L_{\text{cross entropy}}$$

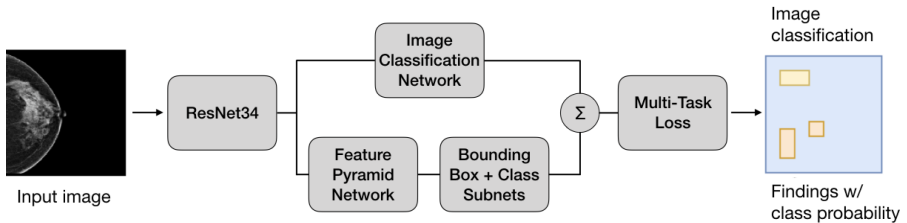


Fig. 1. Architecture of our proposed algorithm. Starting from the state-of-the-art RetinaNet algorithm, we added a classification task with shared weights (ResNet34) and multi-task loss. The trained algorithm outputs both detected findings, each with their own classification (e.g., soft-tissue lesion vs calcification) and probability of malignancy, as well as the probability of malignancy for the full image.

We also made multiple changes to the standard RetinaNet model underlying our algorithm. All changes were applied consistently, for fair comparison of the baseline and proposed algorithms.

First, we addressed the issue that the RetinaNet model is initialized from pre-trained ImageNet weights containing 3 color channels. The typical approach in medical imaging is to replicate the grayscale image 3-fold to match the expected input of ImageNet models. We observe that there is a much simpler and more efficient solution. We take the ImageNet model, and sum each of the first layer kernels over its channel dimension, reducing it to a single channel input. It is trivial to show that this is mathematically equivalent to replicating the input image, due to the linearity of the convolution operation.

Additional changes were made as follows: (i) To address the fact that findings such as microcalcification clusters may have irregular shapes, yielding a low IoU with the generated anchor boxes, we use the best matching policy of [3]. (ii) We also use a wider range of aspect ratios {1:3, 1:2, 1:1, 2:1, 3:1} for the anchor boxes to account for elongated findings. (iii) We modified the FPN architecture by removing the highest resolution level, because the majority of anchors from the highest level were “easy” background anchors that made negligible contribution to the focal loss. In practice, we found that small findings could be adequately explained with anchors from the second highest level, leading to a more efficient network, without loss in performance. (iv) In the focal loss, we normalized positive and negative anchors separately, whereas the original implementation normalized both positive and negative anchors by the number of positive anchors. By normalizing positive and negative anchors separately, we gave a greater relative weight to positive anchors in the loss computation, which we found to improve performance.

2.3 Model Training

Training experiments were conducted on hardware with 80 Intel(R) CPUs and 8 T V100-SXM2 GPUs with 32 GB memory per GPU. The training architecture was developed in Python 3.7.6 using packages: Pytorch 1.3.0, Torchvision 0.4.2, and Apex 0.1 automatic mixed precision (AMP) from Nvidia. Using AMP enabled training with a batch size of 4 full-sized images per GPU. The training environment was deployed in Docker 18.06.2-ce containers based on Ubuntu Linux 18.04.2 LTS. Training time ranged from 3–6 days for up to 60 epochs. At inference time, the model can be run with single-image batches, and fits on a consumer-grade GPU card with 12 GB memory.

2.4 Dataset and Performance Evaluation

Model training and evaluation was done on a multi-site, multi-vendor in-house research data set containing 8613 images (5825 negative, 2788 positive) from 2699 patients (1351 negative, 1348 positive) at 4 geographically distinct sites within the USA. Images were acquired on Hologic (70%), Siemens (25%), and GE (3%) machines (2% Fuji or undefined). Data was split 80/10/10 at the patient level for training, validation and test, respectively, ensuring that images from the same patient weren’t included in multiple splits. Images were scaled to a pixel size of $100 \times 100 \mu\text{m}$, cropped to eliminate background, padded to match the largest

image in the batch (approx. 2400×1200 pixels), and then normalized to $[-1, 1]$. Training time augmentation was used, including zoom.

Patient-level ground truth was defined as follows. Positive cases were screening exams with biopsy proven malignancy within 12 months. Negative exams were either screening negative (BI-RADS 1 or 2 with 24–48 months of normal follow up), diagnostic negative (recalled with negative diagnostic exam), or biopsy negative (screening exams with biopsy proven benign findings within 12 months). For positive cases, only the biopsied breast was used in training and test, the contralateral breast was discarded due to lack of follow up.

Finding-level ground truth was generated by expert annotation of all biopsy positive findings. No additional findings were annotated. During training and test, annotated findings were assigned to two classes: (i) soft-tissue lesions (masses and asymmetries; 1655 in train, 165 in test), and (ii) calcifications (980 in train, 102 in test). Architectural distortions were not included in this study.

To generate test results that are representative of performance on a screening population [10], inverse weighting [16] was applied to simulate the following prevalence: screening negative (88.4%), diagnostic negative (9.9%), biopsy negative (1.2%) and biopsy positive (0.5%).

We also evaluated our algorithm on a separate, small data set containing 24 interval cancer cases, where the screening exam was assessed as negative by a radiologist, and the patient returned with cancer within 12 months. These cases were selected from a larger pool of interval cancers, on the basis that in each case, the malignant finding is visible upon retrospective investigation.

3 Results

Below we demonstrate that our method improves both image classification performance (Sect. 3.1), and detection performance (Sect. 3.2) over strong baseline algorithms. All results are summarized in Table 1.

3.1 Image Classification Analysis

We compared five methods of image classification: (i) ResNet34, a popular classification architecture and the backbone of RetinaNet, (ii) RetinaNet, using the max detection score as the image classification score, as in [15] (iii) an ensemble of ResNet34 and the max detection score from RetinaNet, (iv) our proposed method, using the max detection score from the detection head as the image classification score, and (v) our proposed method, using the classification head.

Using the classification head from our proposed method increases the AUC for image classification by 0.058 (p -value $< 10^{-4}$), from 0.851 (95% CI (0.824–0.876)) for the second-best performing algorithm (ensemble of ResNet34 and RetinaNet) to 0.909 (0.890–0.927). Compared to the naive method of using the maximum detection score from RetinaNet [15], we see an even greater improvement of 0.195 (p -value $< 10^{-4}$) from an AUC of 0.714 (0.679–0.750).

We show all ROC curves with corresponding AUCs in Fig. 2. Our method yields better performance for a continuum of operating points. For example, at an operating point with sensitivity of 0.80, our method increases the specificity from 0.471 using the maximum finding output in RetinaNet, 0.495 using the maximum finding output with our proposed architecture, 0.712 using ResNet34, and 0.723 using an ensemble of ResNet and RetinaNet, to 0.876 with the classification head of our method (p -value $< 10^{-4}$ in all cases).

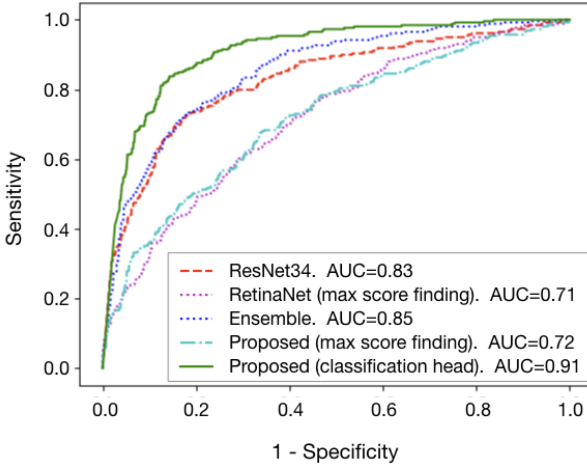


Fig. 2. ROC analysis. ROC curves obtained for (i) ResNet34, (ii) baseline RetinaNet using the max-detection score, (iii) ensemble of ResNet34 and the max-detection score from baseline RetinaNet, (iv) our proposed method using the max-detection score, and (v) our proposed method using the classification head.

3.2 Finding Detection Analysis

As reported in Sect. 2.4, we train the detection algorithm with two different kinds of annotated findings: soft tissue lesions (masses and asymmetries) and calcifications. We report detection performance for each finding type individually, using Free Response ROC Curve (FROC) analysis (Fig. 3).

In Table 1, we show the trade-off between sensitivity and the number of FPPIs at different operating points for soft tissue lesions, calcifications, and both finding types combined (average). For example, at a sensitivity of 0.7, the number of FPPIs is reduced from 9.04 to 2.09 (77% reduction; $p < 10^{-4}$) for calcifications, from 1.38 to 1.39 (0.72% increase; not significant) for masses, and from 3.08 to 1.65 (46% reduction; $p < 10^{-4}$) for both finding types combined. Similarly, at a fixed number of 2 FPPIs, the sensitivity increases from 0.58 to 0.7 for calcifications (21% increase; p 0.04), from 0.74 to 0.77 for masses (4%

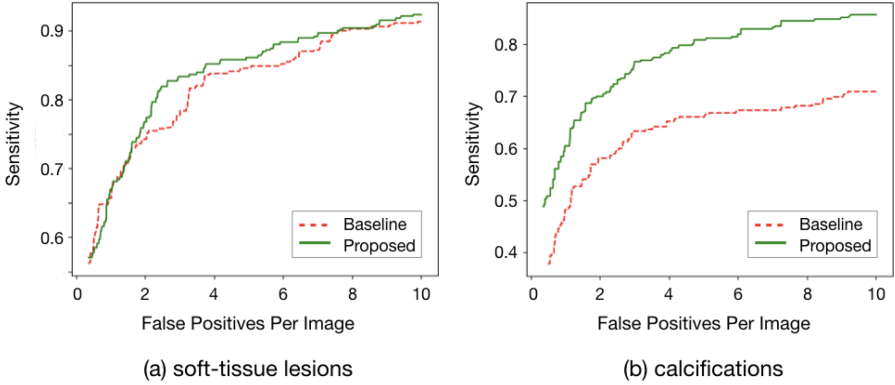


Fig. 3. FROC analysis. FROC curves show sensitivity against the average number of false positives per image (FPPI) for detection of (a) soft-tissue lesions, and (b) calcifications. Baseline: RetinaNet; proposed: detection head of our multi-task network.

Table 1. Detection and classification results for baseline and proposed algorithms. p -values are reported as follows: $^*(p < 0.05)$; $^{**}(p < 0.01)$; $^{***}(p < 10^{-3})$

	Resnet34	RetinaNet	Ensemble	Proposed
Classification				
$AUC_{classification}$	0.83	–	0.85	0.91^{***}
$AUC_{max\ detection}$	–	0.71	–	0.72
Detection average soft-tissue lesions and calcifications				
Se @ FPPI = 2.0	–	0.67	–	0.76^{***}
Se @ FPPI = 3.0	–	0.74	–	0.80^{**}
FPPI @ Se = 0.6	–	1.51	–	0.89^{***}
FPPI @ Se = 0.7	–	3.08	–	1.65^{***}
Detection calcifications				
Se @ FPPI = 2.0	–	0.58	–	0.70[*]
Se @ FPPI = 3.0	–	0.63	–	0.77^{**}
FPPI @ Se = 0.6	–	2.66	–	1.12^{***}
FPPI @ Se = 0.7	–	9.04	–	2.09^{***}
Detection soft-tissue lesions				
Se @ FPPI = 2.0	–	0.74	–	0.77
Se @ FPPI = 3.0	–	0.77	–	0.83[*]
FPPI @ Se = 0.6	–	0.54	–	0.69
FPPI @ Se = 0.7	–	1.38	–	1.39

increase; not significant), and from 0.67 to 0.76 for both findings combined (13% increase; $p < 10^{-3}$).

We also evaluated our algorithm on a separate set of 24 interval cancer cases, where the screening exam was assessed as negative by a radiologist, and the patient returned with cancer within 12 months. At a threshold of 0.85, our algorithm achieved a sensitivity of 0.67 with 1.4 FPPI for soft tissue lesions, and a sensitivity of 0.5 with 3 FPPI for calcifications on the interval cancer cases. In Fig. 4, we show example detections from our algorithm. Figure 4(a) is a case of malignancy that was detected during screening, and Fig. 4(b) is a case of an interval cancer.

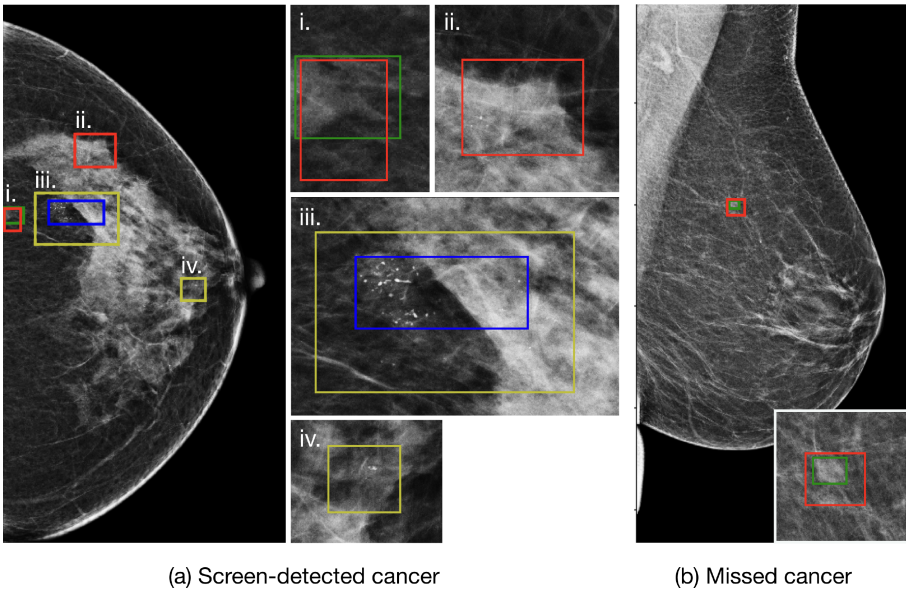


Fig. 4. Example detections from our proposed method. Predicted boxes obtained with a detection threshold of 0.85 are shown for soft-tissue lesions (prediction: red; ground truth: green) and calcifications (prediction: yellow; ground truth: blue). (a) A mammogram containing lesions detected during breast screening, with algorithm output overlaid. Enlarged candidate detections show: (i) a true positive detection of a malignant mass, (ii) a false positive detection of a malignant mass, (iii) a true positive detection of malignant calcifications, and (iv) a false positive detection of malignant calcifications (the calcifications are benign). (b) Detection, by our algorithm, of a malignant mass that was missed by a radiologist. (Color figure online)

4 Conclusion

We have presented a flexible and efficient single-shot, multi-class MTL algorithm that takes as input a screening mammogram and returns a probability of malig-

nancy for the image, as well as detected findings. This method leverages both subjective expert annotations and non-subjective outcome-based ground truth.

We tested our method on a simulated screening population and achieved an AUC of 91% for image classification, an improvement of 6% ($p < 10^{-4}$) over the second-best performing method. We also observed an improvement in the detection of malignant microcalcifications, but not for soft-tissue lesions. This may be due to the fact that the baseline performance was much higher for soft-tissue lesions, leaving less room for improvement. Compared to other published results on public datasets [1, 2, 7], we achieved a lower detection performance; however, this may be due to the challenging multi-site, multi-vendor nature of our data.

In future, we plan to apply methods such as RetinaMask [3], to address the issue of low IOU between findings and RetinaNet's bounding box representation. We anticipate this will further improve the detection performance. We will also focus on detecting biopsy negative findings, which currently have a 10x higher false positive rate compared to screening negative cases when evaluated at an operating point with sensitivity of 80%.

We demonstrated the potential impact of our algorithm by using it to successfully detect cancer that was missed by radiologists during screening but visible retrospectively. Moreover, we feel that this approach will be useful for a wide range of medical imaging problems, where a clinical decision is made at a patient or organ level, but finding-level information confers significant advantage, both during training, and as a form of direct explanatory output at run time.

References

1. Akselrod-Ballin, A., Karlinsky, L., Alpert, S., Hasoul, S., Ben-Ari, R., Barkan, E.: A region based convolutional network for tumor detection and classification in breast mammography. In: Carneiro, G., et al. (eds.) LABELS/DLMIA -2016. LNCS, vol. 10008, pp. 197–205. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46976-8_21
2. Akselrod-Ballin, A., et al.: Deep learning for automatic detection of abnormal findings in breast mammography. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 321–329. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_37
3. Fu, C.Y., Shvets, M., Berg, A.C.: RetinaMask: learning to predict masks improves state-of-the-art single-shot detection for free. arXiv preprint [arXiv:1901.03353](https://arxiv.org/abs/1901.03353) (2019)
4. Gao, F., Yoon, H., Wu, T., Chu, X.: A feature transfer enabled multi-task deep learning model on medical imaging. *Expert Syst. Appl.* **143**, 112957 (2020)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: International Conference on Computer Vision (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
7. Jung, H., et al.: Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. *PLoS ONE* **13**(9), e0203355 (2018)

8. Kim, H.E., et al.: Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digit. Health* **2**(3), e138–e148 (2020)
9. Le, T.L.T., Thome, N., Bernard, S., Bismuth, V., Patoureaux, F.: Multitask classification and segmentation for cancer diagnosis in mammography. In: International Conference on Medical Imaging with Deep Learning - Extended Abstract Track, London, UK (2019)
10. Lehman, C.D., et al.: National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology* **283**(1), 49–58 (2017)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
12. Lotter, W., Sorensen, G., Cox, D.: A multi-scale CNN and curriculum learning strategy for mammogram classification. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 169–177. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_20
13. Mainiero, M.B., Parikh, J.R.: Recognizing and overcoming burnout in breast imaging. *J. Breast Imaging* **1**(1), 60–63 (2019)
14. McKinney, S.M., et al.: International evaluation of an AI system for breast cancer screening. *Nature* **577**(7788), 89–94 (2020)
15. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. *Sci. Rep.* **8**(1), 1–7 (2018)
16. Seaman, S.R., White, I.R.: Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.* **22**(3), 278–295 (2013)
17. Tabár, L., et al.: The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening. *Cancer* **125**(4), 515–523 (2019)
18. Teare, P., Fishman, M., Benzaquen, O., Toledano, E., Elnekave, E.: Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. *J. Digit. Imaging* **30**(4), 499–505 (2017). <https://doi.org/10.1007/s10278-017-9993-2>
19. International Agency for Research on Cancer: World Health Organization: Global cancer observatory database (2018)
20. Wu, N., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans. Med. Imaging* **39**(4), 1184–1194 (2020)
21. Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**(1), 60–66 (2019)
22. Yala, A., Schuster, T., Miles, R., Barzilay, R., Lehman, C.: A deep learning model to triage screening mammograms: a simulation study. *Radiology* **293**(1), 38–46 (2019)
23. Zlocha, M., Dou, Q., Glocker, B.: Improving RetinaNet for CT lesion detection with dense masks from weak RECIST labels. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 402–410. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_45