

Detecting and classifying lesions in mammograms with Deep Learning

Dezső Ribli^{1*}, Anna Horváth², Zsuzsa Unger³, Péter Pollner⁴, and István Csabai¹

¹Department of Physics of Complex Systems, Eötvös Loránd University, Budapest

²3rd Department of Internal Medicine, Semmelweis University, Budapest

³Department of Radiology, Semmelweis University, Budapest

⁴MTA-ELTE Statistical and Biological Physics Research Group, Hungarian Academy of Sciences, Budapest

*dkrib@caesar.elte.hu

ABSTRACT

In the last two decades Computer Aided Diagnostics (CAD) systems were developed to help radiologists analyze screening mammograms. The benefits of current CAD technologies appear to be contradictory and they should be improved to be ultimately considered useful. Since 2012 deep convolutional neural networks (CNN) have been a tremendous success in image recognition, reaching human performance. These methods have greatly surpassed the traditional approaches, which are similar to currently used CAD solutions. Deep CNN-s have the potential to revolutionize medical image analysis. We propose a CAD system based on one of the most successful object detection frameworks, Faster R-CNN. The system detects and classifies malignant or benign lesions on a mammogram without any human intervention. The proposed method sets the state of the art classification performance on the public INbreast database, AUC = 0.95. The approach described here has achieved the 2nd place in the Digital Mammography DREAM Challenge with AUC = 0.85. When used as a detector, the system reaches high sensitivity with very few false positive marks per image on the INbreast dataset. Source code, the trained model and an OsiriX plugin are published online at https://github.com/riblidezso/frcnn_cad.

Introduction

Screening mammography

Breast cancer is the most common cancer in women and it is the main cause of death from cancer among women in the world¹. Screening mammography was shown to reduce breast cancer mortality by 38–48% among participants². In the EU 25 of the 28 member states are planning, piloting or implementing screening programs to diagnose and treat breast cancer in an early stage³. During a standard mammographic screening examination, X-ray images are captured from 2 angles of each breast. These images are inspected for malignant lesions by one or two experienced radiologists. Suspicious cases are called back for further diagnostic evaluation.

Screening mammograms are evaluated by human readers. The reading process is monotonous, tiring, lengthy, costly and most importantly prone to errors. Multiple studies have shown that 20-30% of the diagnosed cancers could be found retrospectively on the previous negative screening exam by blinded reviewers⁴⁻⁹. The problem of missed cancers still persists despite modern full field digital mammography (FFDM)^{4,8}. The sensitivity and specificity of screening mammography is reported to be between 77-87% and 89-97% respectively. These metrics describe the average performance of readers, and there is substantial variance in the performance of individual physicians, with reported false positive rates between 1-29%, and sensitivities between 29-97%¹⁰⁻¹². Double reading was found to improve the performance of mammographic evaluation and it had been implemented in many countries¹³. Multiple reading can further improve diagnostic performance up to more than 10 readers, proving that there is room for improvement in mammogram evaluation beyond double reading¹⁴.

Computer-aided detection in mammographic screening

Computer-aided detection (CAD) solutions were developed to help radiologists in reading mammograms. These programs usually analyze a mammogram and mark the suspicious regions, which should be reviewed by the radiologist¹⁵. The technology was approved by FDA and had spread quickly. By 2008, in the US, 74% of all screening mammograms in the Medicare population were interpreted with CAD, the cost of CAD usage is over \$400 million a year.¹¹

The benefits of using CAD are controversial. Initially several studies have shown promising results with CAD^{6,16-20}. A large clinical trial in the United Kingdom has shown that single reading with CAD assistance has similar performance to double reading.²¹ However, in the last decade multiple studies concluded that currently used CAD technologies do not improve the

performance of radiologists in everyday practice in the United States.^{11,22,23} These controversial results indicate that CAD systems need to be improved before radiologists can ultimately benefit from using the technology in everyday practice.

Currently used CAD approaches are based on describing the X-ray image with meticulously designed hand crafted features, and machine learning for classification on top of these features^{15,24–27}. In the field of computer vision, since 2012, deep convolutional neural networks (CNN) have significantly outperformed these traditional methods²⁸. Deep CNN-s have reached or even surpassed human performance in image classification and object detection.²⁹ These models have tremendous potential in medical image analysis. Several studies have attempted to apply Deep Learning to analyze mammograms^{27,30–32}, but the problem is still far from being solved.

The Digital Mammography DREAM Challenge

The Digital Mammography DREAM Challenge (DM challenge)^{33,34} asked participants to write algorithms which can predict whether a breast in a screening mammography exam will be diagnosed with cancer. The dataset consisted of 86000 exams, with no pixel level annotation, only a binary label indicating if breast cancer was diagnosed in the next 12 months after the exam. Each side of the breasts were treated as a separate case that we will call breast-level prediction in this paper. The participants had to upload their programs to a secure cloud platform, and they were not able to download or view the images, neither interact with their program during training or testing. The DM challenge provided an excellent opportunity to compare the performance of competing methods in a controlled and fair way instead of self-reported evaluations on different or proprietary datasets.

1 Material and methods

Data

Mammograms with pixel level annotations were needed to train a lesion detector and test the classification and localization performance. We have trained the model on the public Digital Database for Screening Mammography (DDSM)³⁵ and a dataset from the Semmelweis University in Budapest, and tested it on the public INbreast³⁶ dataset. The images used for training contain either histologically proven cancers or benign lesions which were recalled for further examinations, but later turned out to be nonmalignant. We expect that training with both kinds of lesions helps our model to find more lesions of interest, and differentiate between malignant and benign examples.

The DDSM dataset contains 2620 digitized film-screen screening mammography exams, with pixel-level ground truth annotation of lesions. Cancerous lesions have histological proof. We have only used the DDSM database for training our model and not evaluating it. The quality of digitized film-screen mammograms is not as good as full field digital mammograms therefore, evaluation on these cases is not relevant. We have converted the lossless jpeg images to png format, mapped the pixel values to optical density using the calibration functions from the DDSM website, and rescaled the pixel values to the 0-255 range.

The dataset from the Department of Radiology at the Semmelweis University in Budapest, Hungary contains 847 FFDM images of 214 exams from 174 patients, recorded with a Hologic LORAD Selenia device. Institutional board approval was obtained for the dataset. This dataset was not available for the full period of the DM challenge, it is used only for improvement in the second stage of the DM challenge, after the pixel level annotation by the authors.

The INbreast dataset contains 115 FFDM cases with pixel-level ground truth annotations, and histological proof for cancers³⁶. We have adapted the INbreast pixel level annotations to suit our testing scenario. We have ignored all benign annotations, and converted the malignant lesion annotations to bounding boxes. We have excluded 8 exams which had other findings, artifacts, previous surgeries or ambiguous pathological outcome. The images have low contrast therefore, we have adjusted the window of the pixel levels. The pixel values were clipped to be minimum 500 pixel lower and maximum 800 pixels higher than the mode of the pixel value distribution (excluding the background) and were rescaled to the 0-255 range.

Data Availability

The DDSM dataset is available online at <http://marathon.csee.usf.edu/Mammography/Database.html>. The INBreast dataset can be requested online at http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database.

For the dataset from the Semmelweis university (<http://semmelweis.hu/radiologia/>) restrictions apply to the availability of these data, which were used under special licence from Semmelweis University, and so are not publicly available. Data are however available from the authors upon reasonable request and permission of the Semmelweis University.

Methods

The heart of our model is a state of the art object detection framework, Faster R-CNN³⁷. Faster R-CNN is based on a convolutional neural network with additional components for detecting, localizing and classifying objects in an image. Faster R-CNN has a branch of convolutional layers, called Region Proposal Network (RPN), after the last convolutional layer of

the original network, which is trained to detect and localize objects on the image, regardless of the class of the object. There are default detection boxes with different sizes and aspect ratios in order to find objects with varying sizes and shapes. The highest scoring default boxes are called the region proposals for the other branch of the network. The other branch of the neural network evaluates the signal coming from each proposed region of the last convolutional layer, resampled to a fix size. Both branches try to solve a classification task to detect the presence of objects and a bounding-box regression task in order to refine the boundaries of the object present in the region. From the detected overlapping objects, the best predictions are selected using non-maximum suppression. Further details about Faster R-CNN can be found in the original article³⁷. An outline of the model can be seen in Fig. 1.

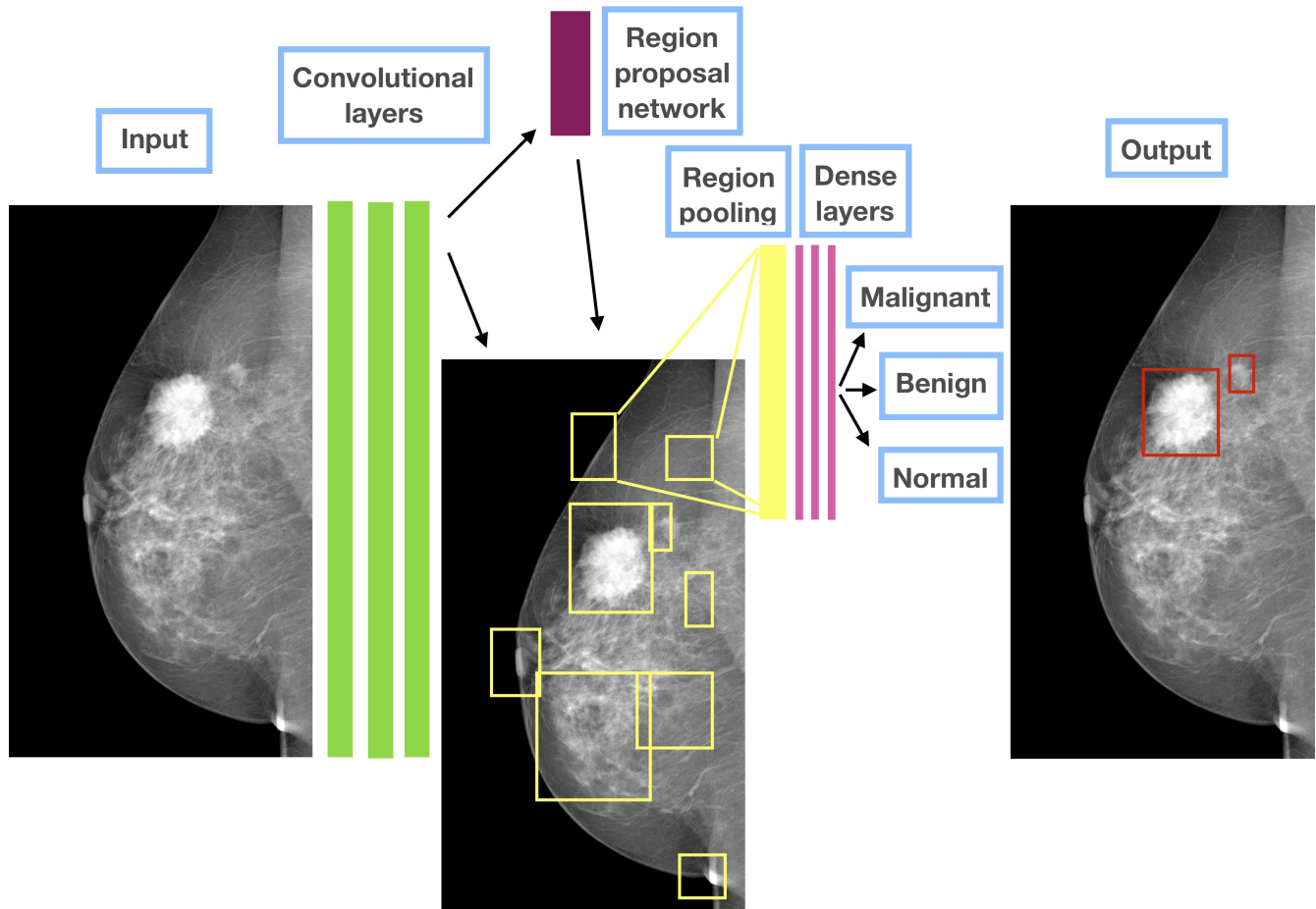


Figure 1. The outline of the Faster R-CNN model for CAD in mammography.

The base CNN used in our model was a VGG16 network, which is 16 layer deep CNN³⁸. The final layer can detect 2 kinds of objects in the images, benign lesions or malignant lesions. The model outputs a bounding box for each detected lesion, and a score, which reflects the confidence in the class of the lesion. To describe an image with one score, we take the maximum of the scores of all malignant lesions detected in the image. For multiple images of the same breast, we take the average of the scores of individual images. For the DM challenge we have trained 2 models using shuffled training datasets. When ensembling these models, the score of an image was the average score of the individual models. This approach has been motivated by a previous study on independent human readers, and it has proven reasonably effective, while simple and flexible¹⁴.

We have used the framework developed by the authors of Faster R-CNN³⁷, which was built in the Caffe framework for deep learning³⁹. During training we optimized the object detection part and the classifier part of the model in the same time, this is called the joint optimization³⁷. We used backpropagation and stochastic gradient descent with weight decay. The initial model used for training has been pretrained on 1.2 million images from the ImageNet dataset³⁸.

We have found that higher resolution yields better results, therefore the mammograms were rescaled isotropically to longer side smaller than 2100 pixels or shorter side smaller than 1700 pixels. This resolution is close to the maximum size which fits in the memory of the graphics card used. The aspect ratio was selected to fit the regular aspect ratio of Hologic images.

We applied vertical and horizontal flipping to augment the training dataset. Mammograms contain fewer objects than ordinary images, therefore negative proposals dominate minibatches. The Intersection over Union (IoU) threshold for foreground objects in the region proposal network was relaxed from 0.7 to 0.5 to allow more positive samples in each minibatch. Relaxation of positive examples is also supported by the fact that lesions on a mammogram have much less well defined boundaries than a car or a dog on a traditional image. The IoU threshold of the final non maximum suppression (nms) was set to 0.1, because mammograms represent a smaller and compressed 3D space compared to ordinary images, therefore overlapping detections are expected to happen less often than in usual object detection. The model was trained for 40k iterations, this number was previously found to be close to optimal by testing multiple models on the DM challenge training data. The model was trained and evaluated on an Nvidia GTX 1080Ti graphics card. Our final entry in the DM challenge was an ensemble of 2 models.

Results

Cancer classification

We also evaluated the model's performance on the public INbreast dataset with the receiver operating characteristics (ROC) metric, Fig. 2. The INbreast dataset has many exams with only one laterality, therefore we have evaluated predictions for each breast. The system achieved $AUC = 0.95$, (95 percentile interval: 0.91 to 0.98, estimated from 10000 bootstrap samples). To our best knowledge this is the highest AUC score reported on the INbreast dataset with a fully automated system based on a single model.

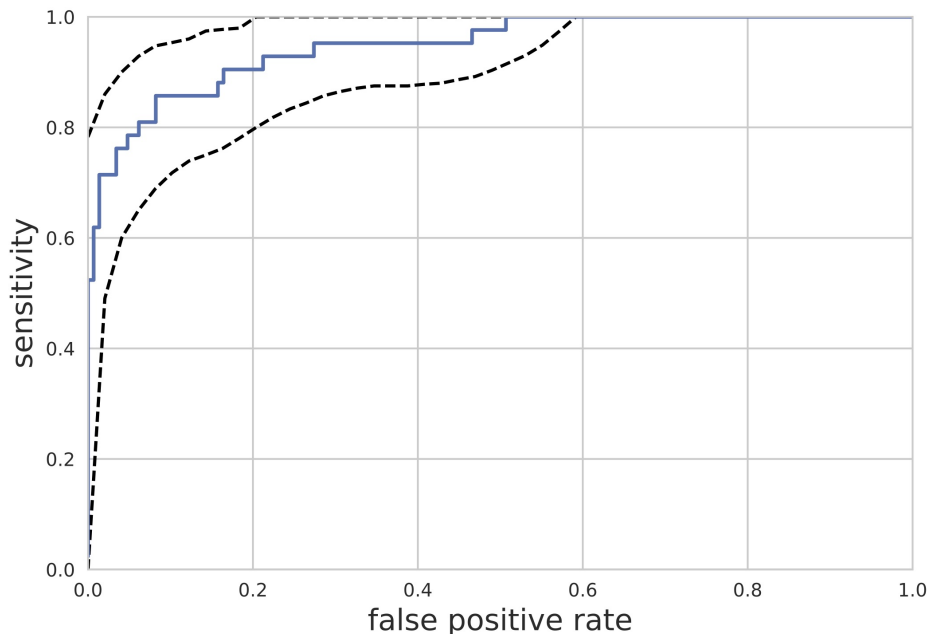


Figure 2. Classification performance. The solid blue line shows the ROC curve on the INbreast dataset on breast level, $AUC = 0.95$, the dashed lines show the 95 percentile interval of the curve based on 10000 bootstrap samples.

FROC analysis

In order to test the model's ability to detect and accurately localize malignant lesions, we evaluated the predictions on the INbreast dataset using the Free-response ROC (FROC) curve⁴⁰. The FROC curve shows the sensitivity (fraction of correctly localized lesions) as a function of the number of false positive marks put on an image Fig. 3.

A detection was considered correct if the center of the proposed lesion fell inside a ground truth box. The same criteria is generally used when measuring the performance of currently used CAD products^{24,41,42}. The DM challenge dataset has no lesion annotation, therefore we can not use it for an FROC analysis.

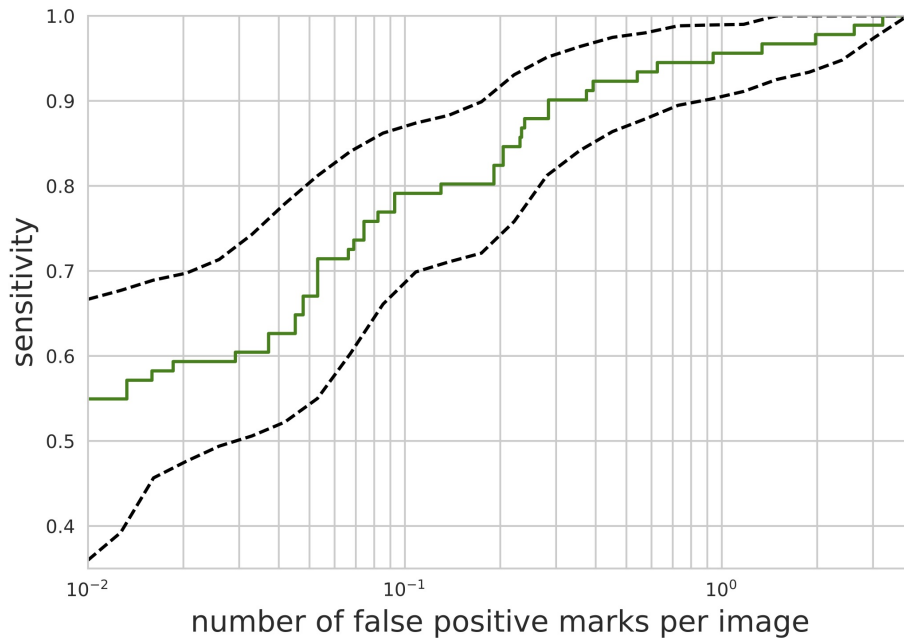


Figure 3. FROC curve on the INbreast dataset. Sensitivity is calculated on a per lesion basis. The solid curve with squares shows the results using all images, while the dashed lines show the 95 percentile interval from 10000 bootstrap samples.

Examples

To demonstrate the characteristics and errors of the detector, we have created a collection of correctly classified, false positive and missed malignant lesions of the INbreast dataset, see in Fig. 4. The score threshold for the examples was defined at sensitivity = 0.9 and 0.3 false positive marks per image.

After inspecting the false positive detections, we have found that most were benign masses or calcifications. Some of these benign lesions were biopsy tested according to the case descriptions of the INbreast dataset. While 10% of the ground truth malignant lesions were missed at this detection threshold, these were not completely overlooked by the model. With a score threshold which corresponds to 3 false positive marks per image, all the lesions were correctly detected (see Fig. 3). Note that the exact number of false positive and true positive detections slightly varies with with different samples of images, indicated by the area in Fig. 3.

Discussion

We have proposed a Faster R-CNN based CAD approach, which achieved the 2nd position in the Digital Mammography DREAM Challenge with an AUC = 0.85 score on the final validation dataset. The competition results have proved that the method described in this article is one of the best approach for cancer classification in mammograms. Our method was the only one of the top contenders in the DM challenge which is based on the detection of malignant lesions, and whole image classification is just a trivial step from the detection task. We think that a lesion detector is clinically much more useful than a simple classifier. A classifier only gives a single score for a case or breast, but it is not able to locate the cancer which is essential for further diagnostic tests or treatment.

We have evaluated the model on the publicly available INbreast dataset. The system is able to detect 90% of the malignant lesions in the INbreast dataset with only 0.3 false positive marks per image. It also sets the state of the art performance in cancer classification on the publicly available INbreast dataset. The system uses the mammograms as the only input without any annotation or user interaction.

An object detection framework developed to detect objects in ordinary images shows excellent performance. This result indicates that lesion detection on mammograms is not very different from the regular object detection task. Therefore the expensive, traditional CAD solutions, which have controversial efficiency, could be replaced with the recently developed, potentially more accurate, deep learning based, open source object detection methods in the near future. The FROC analysis

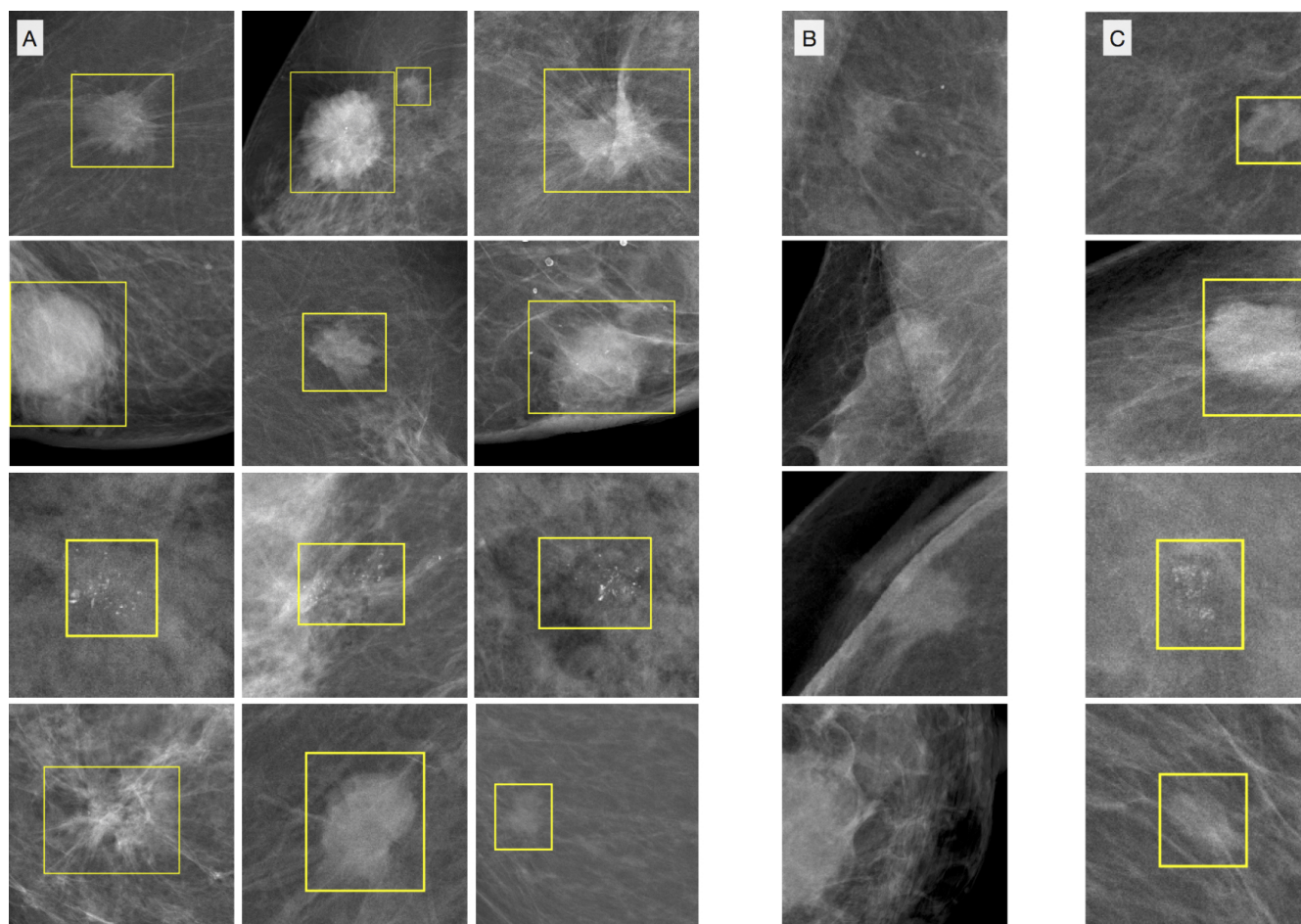


Figure 4. Detection examples: The yellow boxes show the lesion proposed by the model. The threshold for these detections was selected to be at lesion detection sensitivity = 0.9. A: Correctly detected malignant lesions B: Missed malignant lesions C: False positive detections, Courtesy of the Breast Research Group, INESC Porto, Portugal³⁶

results suggest that the proposed model could be applied as a perception enhancer tool, which could help radiologists to detect more cancers.

A limitation of our study comes from the small size of the publicly available pixel-level annotated dataset. While the classification performance of the model has been evaluated on a large screening dataset, the detection performance could only be evaluated on the small INbreast dataset.

References

1. Ferlay, J., Héry, C., Autier, P. & Sankaranarayanan, R. Global burden of breast cancer. In *Breast cancer epidemiology*, 1–19 (Springer, 2010).
2. Broeders, M. *et al.* The impact of mammographic screening on breast cancer mortality in europe: a review of observational studies. *J. medical screening* **19**, 14–25 (2012).
3. Ponti, A. *et al.* Cancer screening in the european union. final report on the implementation of the council recommendation on cancer screening (2017).
4. Bae, M. S. *et al.* Breast cancer detected with screening us: reasons for nondetection at mammography. *Radiol.* **270**, 369–377 (2014).
5. Bird, R. E., Wallace, T. W. & Yankaskas, B. C. Analysis of cancers missed at screening mammography. *Radiol.* **184**, 613–617 (1992).

6. Birdwell, R. L., Ikeda, D. M., O'Shaughnessy, K. F. & Sickles, E. A. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection 1. *Radiol.* **219**, 192–202 (2001).
7. Harvey, J. A., Fajardo, L. L. & Innis, C. A. Previous mammograms in patients with impalpable breast carcinoma: retrospective vs blinded interpretation. 1993 arrs president's award. *AJR. Am. journal roentgenology* **161**, 1167–1172 (1993).
8. Hoff, S. R. *et al.* Breast cancer: missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography—results from a retrospective review. *Radiol.* **264**, 378–386 (2012).
9. Martin, J. E., Moskowitz, M. & Milbrath, J. R. Breast cancer missed by mammography. *Am. J. Roentgenol.* **132**, 737–739 (1979).
10. Banks, E. *et al.* Influence of personal characteristics of individual women on sensitivity and specificity of mammography in the million women study: cohort study. *Bmj* **329**, 477 (2004).
11. Lehman, C. D. *et al.* Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine* **175**, 1828–1837 (2015).
12. Smith-Bindman, R. *et al.* Physician predictors of mammographic accuracy. *J. Natl. Cancer Inst.* **97**, 358–367 (2005).
13. Blanks, R., Wallis, M. & Moss, S. A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the uk national health service breast screening programme. *J. Med. screening* **5**, 195–201 (1998).
14. Karssemeijer, N., Otten, J. D., Roelofs, A. A., van Woudenberg, S. & Hendriks, J. H. Effect of independent multiple reading of mammograms on detection performance. In *Medical Imaging 2004*, 82–89 (International Society for Optics and Photonics, 2004).
15. Christoyianni, I., Koutras, A., Dermatas, E. & Kokkinakis, G. Computer aided diagnosis of breast cancer in digitized mammograms. *Comput. medical imaging graphics* **26**, 309–319 (2002).
16. Brem, R. F. *et al.* Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. *Am. J. Roentgenol.* **181**, 687–693 (2003).
17. Ciatto, S. *et al.* Comparison of standard reading and computer aided detection (cad) on a national proficiency test of screening mammography. *Eur. journal radiology* **45**, 135–138 (2003).
18. Freer, T. W. & Ulissey, M. J. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiol.* **220**, 781–786 (2001).
19. Morton, M. J., Whaley, D. H., Brandt, K. R. & Amrami, K. K. Screening mammograms: interpretation with computer-aided detection—prospective evaluation. *Radiol.* **239**, 375–383 (2006).
20. Warren Burhenne, L. J. *et al.* Potential contribution of computer-aided detection to the sensitivity of screening mammography 1. *Radiol.* **215**, 554–562 (2000).
21. Gilbert, F. J. *et al.* Single reading with computer-aided detection for screening mammography. *New Engl. J. Medicine* **359**, 1675–1684 (2008).
22. Fenton, J. J. *et al.* Influence of computer-aided detection on performance of screening mammography. *New Engl. J. Medicine* **356**, 1399–1409 (2007).
23. Fenton, J. J. *et al.* Effectiveness of computer-aided detection in community mammography practice. *J. Natl. Cancer institute* **103**, 1152–1161 (2011).
24. Hologic. *Understanding ImageChecker® CAD 10.0 User Guide – MAN-03682 Rev 002* (2017).
25. Hupse, R. & Karssemeijer, N. Use of normal tissue context in computer-aided detection of masses in mammograms. *IEEE Transactions on Med. Imaging* **28**, 2033–2041 (2009).
26. Hupse, R. *et al.* Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. *Eur. radiology* **23**, 93–100 (2013).
27. Kooi, T. *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Med. image analysis* **35**, 303–312 (2017).
28. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105 (2012).

29. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034 (2015).
30. Becker, A. S. *et al.* Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investig. Radiol.* (2017).
31. Dhungel, N., Carneiro, G. & Bradley, A. P. Fully automated classification of mammograms using deep residual neural networks. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, 310–314 (IEEE, 2017).
32. Lotter, W., Sorensen, G. & Cox, D. A multi-scale cnn and curriculum learning strategy for mammogram classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 169–177 (Springer, 2017).
33. DREAM. The digital mammography dream challenge. https://www.synapse.org/Digital_Mammography_DREAM_challenge (2017).
34. Trister, A. D., Buist, D. S. & Lee, C. I. Will machine learning tip the balance in breast cancer screening? *JAMA oncology* (2017).
35. Heath, M., Bowyer, K., Kopans, D., Moore, R. & Kegelmeyer, W. P. The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, 212–218 (Medical Physics Publishing, 2000).
36. Moreira, I. C. *et al.* Inbreast: toward a full-field digital mammographic database. *Acad. radiology* **19**, 236–248 (2012).
37. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99 (2015).
38. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
39. Jia, Y. *et al.* Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* (2014).
40. Bunch, P. C., Hamilton, J. F., Sanderson, G. K. & Simmons, A. H. A free response approach to the measurement and characterization of radiographic observer performance. In *Application of Optical Instrumentation in Medicine VI*, 124–135 (International Society for Optics and Photonics, 1977).
41. Ellis, R. L., Meade, A. A., Mathiason, M. A., Willison, K. M. & Logan-Young, W. Evaluation of computer-aided detection systems in the detection of small invasive breast carcinoma. *Radiol.* **245**, 88–94 (2007).
42. Sadaf, A., Crystal, P., Scaranelo, A. & Helbich, T. Performance of computer-aided detection applied to full-field digital mammography in detection of breast cancers. *Eur. journal radiology* **77**, 457–461 (2011).

Author contributions statement

D.R., I.C. and P.P. contributed to the conception and design of the study. A.H. and Z.U. contributed to the acquisition, analysis and interpretation of data. All authors reviewed the manuscript.

Additional information

Competing financial interests

The authors declare no competing financial interests.

Acknowledgements

This work was supported by the Novo Nordisk Foundation Interdisciplinary Synergy Programme [Grant NNF15OC0016584]; and National Research, Development and Innovation Fund of Hungary, [Project no. FIEK.16-1-2016-0005]. The funding sources cover computational and publishing costs.