# SCIENTIFIC REP𝗢RTS

# Detecting and classifying lesions in mammograms with Deep Learning

Dezső Ribli[1], Anna Horváth[2], Zsuzsa Unger[3], Péter Pollner[4] & István Csabai [1]

**In the last two decades, Computer Aided Detection (CAD) systems were developed to help radiologists analyse screening mammograms, however benefits of current CAD technologies appear to be contradictory, therefore 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 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.**

**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 world[1]. Screening mammography has been shown to reduce breast cancer mortality by 38–48% among participants[2]. 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 stage[3]. 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 diagnosed cancers could be found retrospectively on the previous negative screening exam by blinded reviewers[4–9]. 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 has been implemented in many countries[13]. Multiple readings can further improve diagnostic performance up to more than 10 readers, proving that there is room for improvement in mammogram evaluation beyond double reading[14].

**Computer-aided detection in mammographic screening.**     Computer-aided detection (CAD) solutions were developed to help radiologists in reading mammograms. These programs usually analyse a mammogram and mark the suspicious regions, which should be reviewed by the radiologist[15]. The technology was approved by the FDA and had spread quickly. By 2008, in the US, 74% of all screening mammograms in the Medicare population were interpreted with CAD, however 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

[1]Department of Physics of Complex Systems, Eötvös Loránd University, Budapest, Hungary. [2]3rd Department of Internal Medicine, Semmelweis University, Budapest, Hungary. [3]Department of Radiology, Semmelweis University, Budapest, Hungary. [4]MTA-ELTE Statistical and Biological Physics Research Group, Hungarian Academy of Sciences, Budapest, Hungary. Correspondence and requests for materials should be addressed to D.R. (email: dkrib@caesar. elte.hu)

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 an 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 methods[28]. 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 analyse 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 whether breast cancer was diagnosed within the next 12 months after the exam. Each side of the breasts were treated as separate cases that we will call breast-level prediction in this paper. The participants had to upload their programs to a secure cloud platform, and were not able to download or view the images, nor 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.

## Material and Methods

**Data.** Mammograms with pixel level annotations were needed to train a lesion detector and test the classification and localisation performance. We 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 INbreast[36] 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 expected that training with both kinds of lesions will help our model to find more lesions of interest, and differentiate between malignant and benign examples.

The DDSM dataset contains 2620 digitised 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 digitised 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 calibration functions from the DDSM website, and rescaled the pixel values to the 0–255 range.

The dataset from the Department of Radiology at 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, therefore it was only used for improvement in the second stage of the DM challenge, after pixel level annotation by the authors.

The INbreast dataset contains 115 FFDM cases with pixel-level ground truth annotations, and histological proof for cancers[36]. We adapted the INbreast pixel level annotations to suit our testing scenario. We ignored all benign annotations, and converted the malignant lesion annotations to bounding boxes. We excluded 8 exams which had unspecified other findings, artefacts, previous surgeries, or ambiguous pathological outcome. The images had low contrast, therefore we 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.

The dataset acquired from Semmelweis University (http://semmelweis.hu/radiologia/) was used with a special license therefore is not publicly available, however the authors can supply data upon reasonable request and permission from the University.

**Methods.** The heart of our model is a state of the art object detection framework, Faster R-CNN[37]. Faster R-CNN is based on a convolutional neural network with additional components for detecting, localising and classifying objects in an image. Faster R-CNN has a branch of convolutional layers, called Region Proposal Network (RPN), on top of the last convolutional layer of the original network, which is trained to detect and localise objects on the image, regardless of the class of the object. It uses 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 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, which is 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 article[37]. An outline of the model can be seen in Fig. 1.

The base CNN used in our model was a VGG16 network, which is a 16 layer deep CNN[38]. The final layer can detect 2 kinds of objects in the images, benign or malignant lesions. The model's output is a bounding box for each
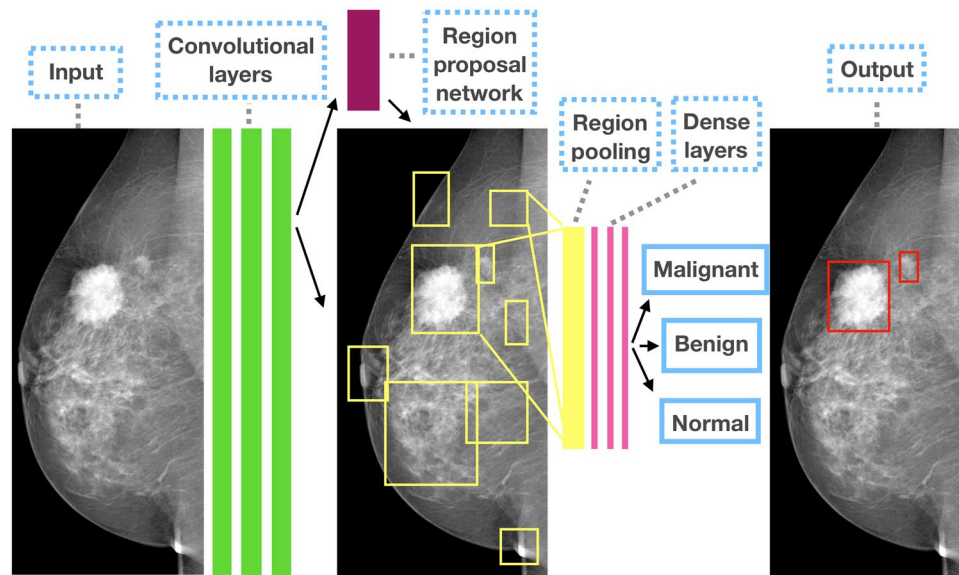
**Figure 1.** The outline of the Faster R-CNN model for CAD in mammography.

detected lesion, and a score, which reflects the confidence in the class of the lesion. To describe an image with one score, we calculate 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 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 was motivated by a previous study on independent human readers, and it has proven reasonably effective, while being both simple and flexible[14].

We used the framework developed by the authors of Faster R-CNN[37], which was built in the Caffe framework for deep learning[39]. During training, we optimized both the object detection and classifier part of the model at the same time, which is called joint optimization[37]. We used backpropagation and stochastic gradient descent with weight decay. The initial model used for training was pretrained on 1.2 million images from the ImageNet dataset[38].

The mammograms were downscaled isotropically so their longer side were smaller than 2100 pixels and their shorter side were smaller than 1700 pixels. This resolution is close to the maximum size that fits in the memory of the graphics card used. The aspect ratio was selected to fit the regular aspect ratio of Hologic images. We have found that higher resolution yields better results as a previous model with images downscaled to $1400 \times 1700$ pixel resolution had 0.08 lower AUC in the DM challenge. We applied vertical and horizontal flipping to augment the training dataset. Mammograms contain fewer objects than ordinary images and during the initial inspection of the training behaviour we have observed pathologically few positive regions in minibatches. To solve the class balance problem we decreased the Intersection over Union (IoU) threshold for foreground objects in the region proposal network from 0.7 to 0.5. This choice allowed more positive examples in a minibatch, and effectively stabilized training. 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 compressed, and relatively thin 3D space compared to ordinary images, therefore overlapping detections are expected to occur 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. The results presented in this article using the INbreast dataset were obtained with a single model. Our final entry in the DM challenge was an ensemble of 2 models, which was the maximum number of models we could run, given the runtime restrictions in the challenge.

## 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 contains many exams with only one laterality, therefore we evaluated predictions for each breast. The system achieved AUC = 0.95, (95 percentile interval: 0.91 to 0.98, estimated from 10000 bootstrap samples). This is the highest AUC score reported on the INbreast dataset with a fully automated system based on a single model, to our best knowledge.

**FROC analysis.** In order to test the model's ability to detect and accurately localise malignant lesions, we evaluated the predictions on the INbreast dataset using the Free-response ROC (FROC) curve[40]. The FROC curve shows sensitivity (fraction of correctly localised lesions) as a function of the number of false positive marks put on an image Fig. 3.
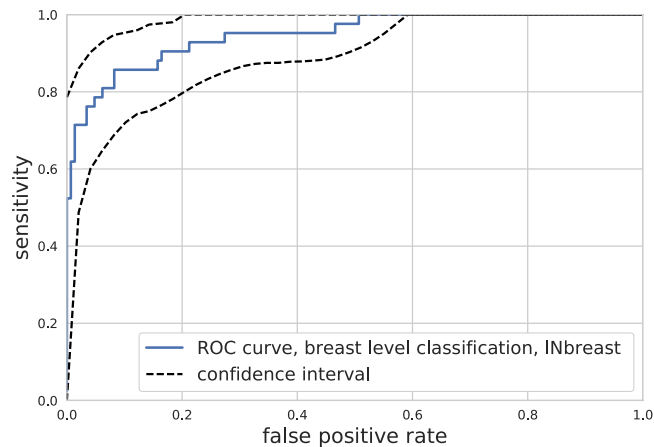
**Figure 2.** Classification performance. The solid blue line shows the ROC curve on the INbreast dataset on breast level, AUC = 0.95, the dashed lines show the 95 percentile interval of the curve based on 10000 bootstrap samples.
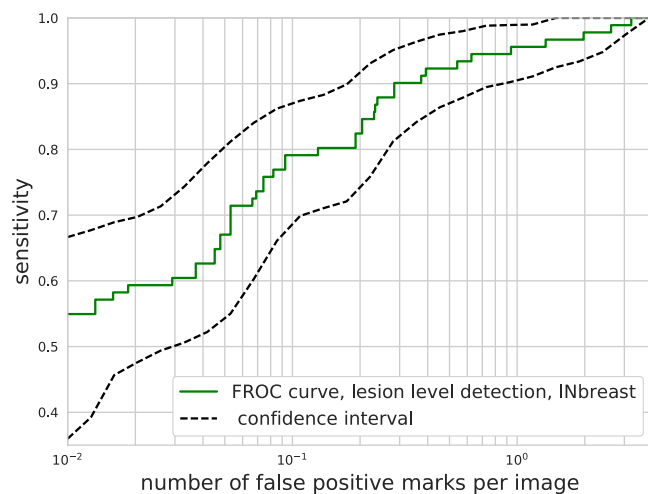


**Figure 3.** FROC curve on the INbreast dataset. Sensitivity is calculated on a per lesion basis. The solid curve with squares shows the results using all images, while the dashed lines show the 95 percentile interval from 10000 bootstrap samples.

A detection was considered correct if the center of the proposed lesion fell within 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 cannot use it for FROC analysis.

As seen in Fig. 3, our model was able to detect malignant lesions with a sensitivity of 0.9 and 0.3 false positive marks per image. The reported number of false positive marks per image for commercially available CAD systems ranges from 0.3 to 1.25[11,16–20,41–43]. The lesion based sensitivities of commercially available CAD systems are generally reported to be around 0.75–0.77 for scanned film mammograms[17,19,20], and 0.85 for FFDM[42,43]. Our model achieves slightly better detection performance on the INbreast dataset than the reported characteristics of commercial CAD systems, although it is important to note that results obtained on different datasets are not directly comparable.

**Examples.** To demonstrate the characteristics and errors of the detector, we 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 a sensitivity of 0.9 and 0.3 false positive marks per image.

After inspecting the false positive detections, we 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 different samples of images, indicated by the area in Fig. 3.
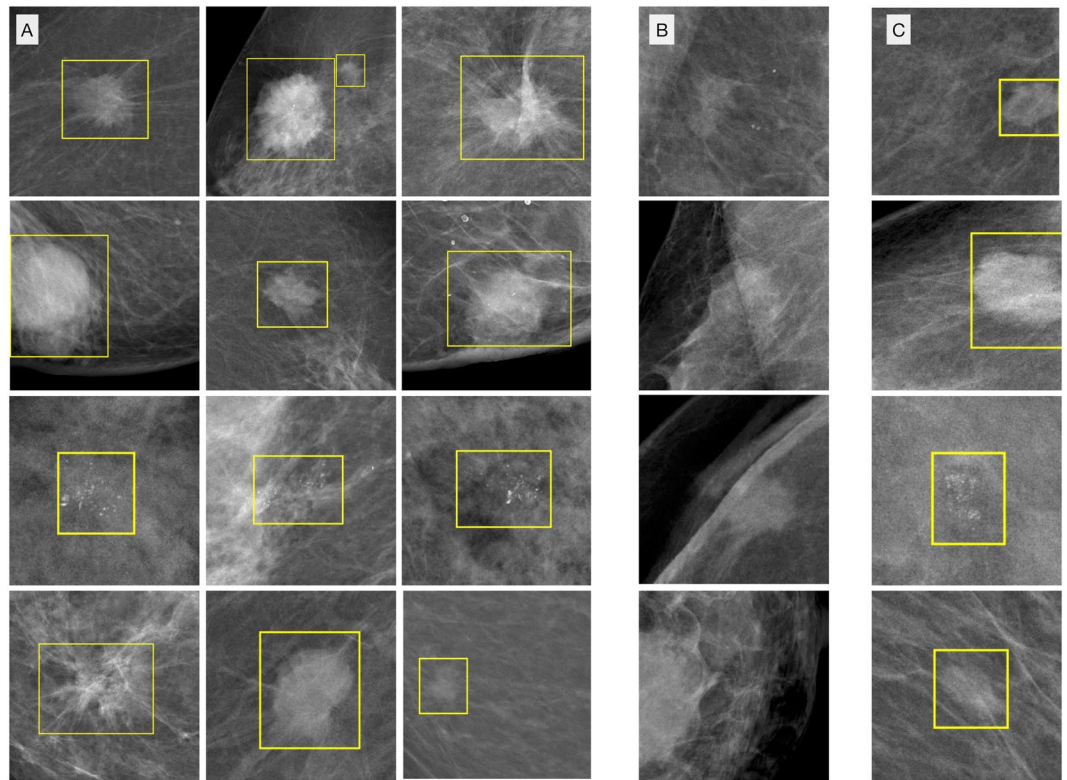
**Figure 4.** Detection examples: The yellow boxes show the lesion proposed by the model. The threshold for these detections was selected to be at lesion detection sensitivity = 0.9. (**A**) Correctly detected malignant lesions. (**B**) Missed malignant lesions. (**C**) False positive detections, Courtesy of the Breast Research Group, INESC Porto, Portugal[36].

## Discussion

We have proposed a Faster R-CNN based CAD approach, which achieved 2nd place in the Digital Mammography DREAM Challenge with an AUC = 0.85 score on the final validation dataset. The competition results proved that the method described in this article is one of the best approaches for cancer classification in mammograms. Out of the top contenders in the DM challenge, our method was the only one that was based on the detection of malignant lesions. For our approach, whole image classification is just a trivial step from the detection task. We believe that a lesion detector is clinically much more useful than a simple image classifier. A classifier only gives a single score per case or breast, and 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 a regular object detection task. Therefore the expensive, traditional CAD solutions, which have controversial efficiency, could be replaced with the recently developed, deep learning based, open source object detection methods in the near future. Provided with more training data, these models have the potential to become significantly more accurate. The FROC analysis 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. *Journal of 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. *Radiology* **270**, 369–377 (2014).
5. Bird, R. E., Wallace, T. W. & Yankaskas, B. C. Analysis of cancers missed at screening mammography. *Radiology* **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. *Radiology* **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. American journal of 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. *Radiology* **264**, 378–386 (2012).
9. Martin, J. E., Moskowitz, M. & Milbrath, J. R. Breast cancer missed by mammography. *American Journal of Roentgenology* **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. *Journal of the National Cancer Institute* **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. *Journal of Medical 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. *Computerized medical imaging and graphics* **26**, 309–319 (2002).
16. Brem, R. F. *et al.* Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. *American Journal of Roentgenology* **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. *European journal of 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. *Radiology* **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. *Radiology* **239**, 375–383 (2006).
20. Warren Burhenne, L. J. *et al.* Potential contribution of computer-aided detection to the sensitivity of screening mammography 1. *Radiology* **215**, 554–562 (2000).
21. Gilbert, F. J. *et al.* Single reading with computer-aided detection for screening mammography. *New England Journal of Medicine* **359**, 1675–1684 (2008).
22. Fenton, J. J. *et al.* Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine* **356**, 1399–1409 (2007).
23. Fenton, J. J. *et al.* Effectiveness of computer-aided detection in community mammography practice. *Journal of the National 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 Medical Imaging* **28**, 2033–2041 (2009).
26. Hupse, R. *et al.* Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. *European radiology* **23**, 93–100 (2013).
27. Kooi, T. *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Medical 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. *Investigative Radiology* (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. *Academic 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. *Radiology* **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. *European journal of radiology* **77**, 457–461 (2011).
43. Kim, S. J. *et al.* Computer-aided detection in full-field digital mammography: sensitivity and reproducibility in serial examinations. *Radiology* **246**, 71–80 (2008).

## Acknowledgements

## Author Contributions

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 Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.