



UKRI AI CENTRE FOR DOCTORAL TRAINING IN
DIGITAL HEALTHCARE

IMPERIAL COLLEGE LONDON

Paper Summary and Report

International evaluation of an AI system for breast cancer
screening

McKinney *et al.* 2020

Author: Alfred BALSTON

April 10, 2025

1 Paper Details

Title: International evaluation of an AI system for breast cancer screening

Authors: McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F.J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C.J., King, D., Ledsam, J.R., Melnick, D., Mostofi, H., Peng, L., Reicher, J.J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K.C., De Fauw, J., & Shetty, S.

Publication details: Published in Nature, Issue 577, January 2020

2 Paper Summary

This paper was a collaboration between DeepMind, Google Health, Imperial College University, Northwestern Medicine, and several other institutions including National Health System (NHS) trusts. In their paper, McKinney et al. describe the development of an AI model that predicts the probability of breast cancer from screening mammography with a degree of accuracy that is equal, and potentially superior, to human experts[1].

2.1 Background

Breast cancer is the most common type of cancer in women worldwide[2]. When diagnosed at an early stage survival rates are high, with more than 90% of individuals living more than five years[2, 3]. Mammography based screening programmes enable earlier detection of cancer and have a beneficial effect on breast cancer survival[2]. For this reason, many countries have developed screening programmes to detect breast cancer at an earlier stage. Mammography is recommended three-yearly for women aged 50-71 in the UK and two-yearly for women aged 40-74 in the United States of America (USA)[4, 5]. In the UK, mammograms are interpreted by two readers and recalled for further investigations if an abnormality is identified. In the USA, mammograms are interpreted by one reader.

However, the accuracy of mammography could be improved to reduce the adverse consequences of diagnostic errors. Specifically, false negatives (i.e. a missed cancer) cause delayed diagnosis and false positives (i.e. flagging a benign lesion) cause unnecessary investigations and anxiety[6, 7]. Furthermore, a worsening shortage of radiologists needed to interpret mammograms may threaten the efficiency of screening programmes[8]. Previous computational methods developed to support radiologists to read mammograms such as computer-aided detection (CAD) software failed to improve the accuracy of mammogram interpretation in real-world studies [6]. The combination of advances in convolutional neural networks (CNNs) for computer vision and the availability of data from large-scale breast cancer screening programmes meant several groups began developing artificial intelligence (AI) systems for classifying mammograms.

2.2 Methods

2.2.1 Data

The authors collected two large datasets from the UK and USA with 121,455 and 22,225 individuals respectively. The data consisted of four mammogram images (mediolateral oblique and craniocaudal views of each breast) and age. For each case, the ground-truth was established by determining if breast cancer was subsequently diagnosed within the screening interval plus

three months (figure 1). More concretely, individuals with biopsy confirmed breast cancer in the 39 (UK) or 27 (US) months following their mammogram were considered positive, and individuals that had two negative interval mammograms were considered negative. In this manner, cases were labelled with a diagnosis in a more robust fashion than simply taking the outcome of the mammogram. By including cancers missed by the screening programme, the AI model has the potential to surpass human performance. A subset of abnormal images were also annotated by 6 US radiologists to identify particular regions of interest.

The UK dataset comprised individuals who had breast screening at three NHS sites between 2010 and 2018 that had sufficient metadata and follow-up scans available. The test set was a subset of the data from two sites between 2012 and 2015. The resulting train/validation/test splits were 11%/62%/27%. This unusual choice of split ratio was not explained by the authors. The US dataset was collected from a single site by retrieving the mammograms of all women who had a breast biopsy between 2001 and 2018, supplemented by a random sample of 5% of women who had a screening mammogram but no biopsy in the same time period (n=7522). The dataset was split 55%/15%/30% train/validation/test.

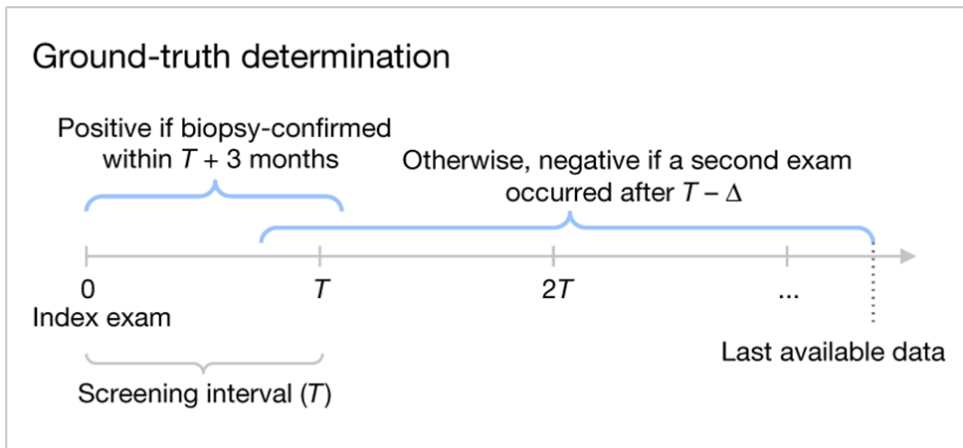


Figure 1: Ground-truth determination strategy. $T = 36$ in the UK and 24 in the USA datasets. Adapted from figure 1 of [1]

2.2.2 Model

The AI model presented is an ensemble of three CNNs aimed at detecting cancer in a complementary way. Each model is a binary classifier that outputs a predicted probability of breast cancer, and the final score is the mean of the three predictions[1].

The first model, referred to as the lesion model, detects suspicious regions within each breast. The second model, the breast model, predicts the per-breast likelihood of cancer using the two mammogram views of each breast. The third model, the full case model, assess all four mammographic views together to create a per-case cancer classification.

The lesion model uses a RetinaNet[9] to detect and extract the ten most suspicious regions from the four mammogram images and classifies each region with MobileNetV2[10]. These region-wise cancer predictions are then aggregated using a Noisy-OR function[11] to give a case level prediction. The RetinaNet was trained using region of interest annotations provided by the US radiologists. Both the breast model and full case model use ResNet architectures. The breast model uses ResNet-V2-50[12] to provide a per-breast cancer prediction, after which the highest prediction is outputted. The full case model is a ResNet-V1-50[13].

Neural network parameters were initially set from ImageNet pre-trained where possible, and data augmentation techniques were applied to each image or region in the lesion model. During training, cancer cases were oversampled to address class imbalance.

2.3 Results

The primary results are shown in figure 2. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for the AI model was 0.89 in the UK test set and 0.81 in the US test set. The authors reported that by varying the operating point of the model (i-iii in figure 2), it performed superiorly to the mean first reader, and non-inferiorly to the mean second reader and consensus opinion (the final conclusion in the case of disagreement between the two readers). The AUC in the US test set improved by 0.05 when trained on both US and UK data compared to training on the US data alone.

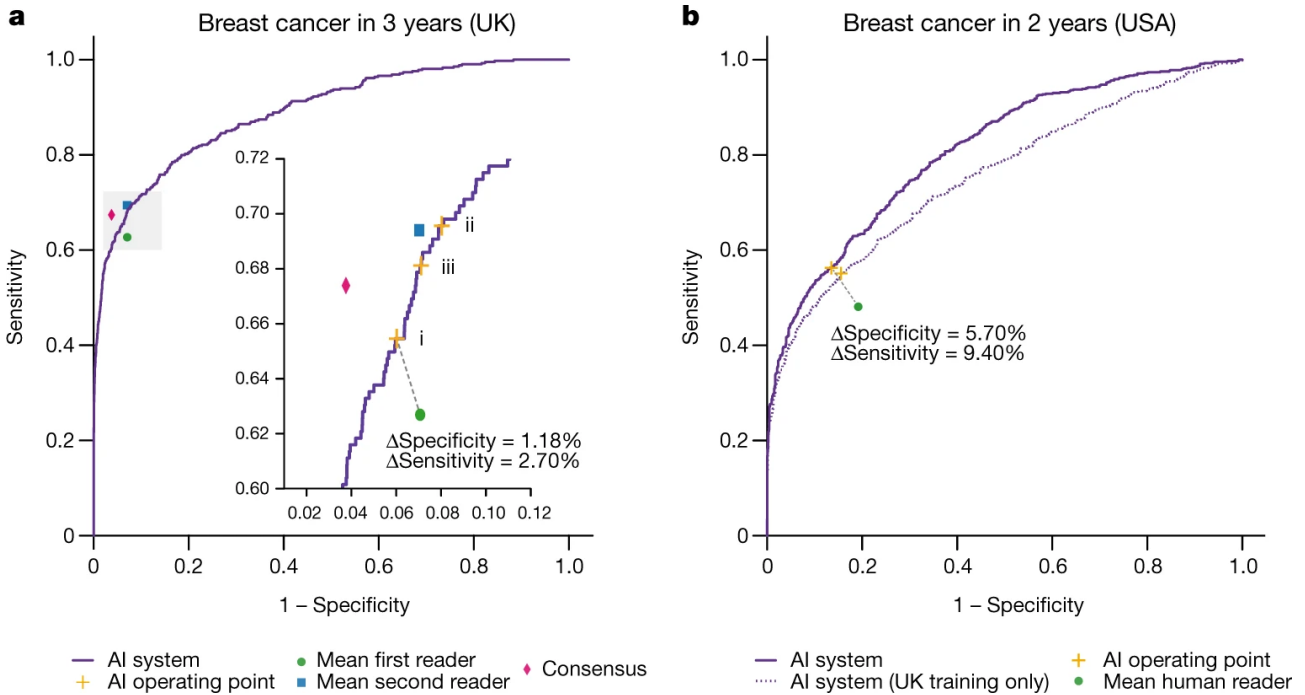


Figure 2: ROC curves for the AI model performance with human reader performance plotted for the UK (a) and USA (b) test data. a, UK AUC is 0.889 (95% confidence interval 0.871-0.907; $n=25,856$ patients). Inset shows the magnification of the grey area. b, USA AUC is 0.8107 (95% confidence interval 0.791-0.831; $n=3,097$ patients) when trained on UK and US data (solid line) and 0.757 (95% confidence interval 0.732-0.780) when trained on only US data (dotted line) [1]

The report also detailed a comparison of the AI model to six individual US readers on a small subset of the US test set ($n=465$). They compared the model AUC to a derived AUC for each individual reader based on the Breast Imaging Reporting and Data System (BI-RADS) scores assigned to the cases by each reader. The authors reported the AI model had an AUC of 0.74; an increase of 0.115 compared to the average performance of the six radiologists.

3 Critique and Discussion

3.1 Strengths

This paper had many key strengths. In particular, the timely application of advances in machine vision with an important healthcare problem meant this paper attracted considerable attention, being accessed over 110,000 times and having 1693 citations as of March 2025. Applying state of the art CNN architectures to healthcare had resulted in human level diagnostic

performance in dermatology and ophthalmology[14, 15]. At the time, mammography remained a difficult problem for AI due to the small sample sizes of publicly available datasets, large file sizes meaning difficulty in implementing efficient training, and the mixture of film and digital images[16]. McKinney et al. were the first researchers to use the large OPTIMAM database for AI, and were the first to report super-human level performance.

Beyond the author’s ability to access a large, well curated dataset, they also achieved a better ground truth compared to previous work[17]. The most comparable published work used histology reports from biopsies performed within 120 days of mammography to assign positive cases[17]. All other cases were labelled as negative[17]. McKinney *et al.* used a longer follow-up period, ensuring that all cancers diagnosed between mammography could be captured. In this way, the authors trained their AI model under real-world screening conditions. As women in the UK have mammography three yearly, it is crucial that any AI model can accurately predict the change of malignancy within this time period, not just after 120 days. McKinney *et al.* also labelled negative cases in the presence of two negative mammograms, not simply the absence of a positive biopsy as Wu *et al.* did[17].

3.2 Limitations

Whilst this work achieved state of the art performance in breast cancer prediction, it came under heavy scrutiny for a lack of transparency and reproducibility in its methodological reporting. In an article published in October 2020 alongside the author’s response, Haibe-Kains *et al.* criticised the limited description of the models, the lack of code availability, and that neither the trained model or the US dataset would be made available to the public or other researchers[18]. Specifically, Haibe-Kains *et al.* pointed out certain hyperparameters (e.g. learning rate, optimizer, number of training epochs) were not reported for all models, and that the image augmentation and data preprocessing strategy were not explained[18].

McKinney *et al.* argued that the majority of the code was not of scientific value and that releasing the full trained model could pose a risk to patient safety if used in an improper context[19]. In contrast, similar work released in pre-print prior to the submission of McKinney *et al.*’s work made their full model available with all of the code used for training[17]. This provided an opportunity for McKinney *et al.* to compare the performance of their model to another AI model, but this was not done.

The study could also have been strengthened in other areas. The authors did not explain the reasoning behind their model architecture and did not report the performance of each part of the ensemble with an ablation study. In common with other work, no detailed demographic information (e.g. ethnicity, employment, education level) were reported and no experiments assessing bias or fairness were performed.

The AI model was compared to human experts by plotting the mean human performance on the AUC curve. Whilst this is common practice in many medical prediction tasks, some have argued that AUC is not a good comparison to human performance, as AUC summarises model performance over a range of operating thresholds which mean human performance does not[20]. Other work attempted to address this by asking radiologists to assign a confidence score to their mammogram interpretation to represent the operating threshold for each human reader[17]. They also surveyed a larger number (14 vs. 6) of readers, and required readers to review more mammograms (720 vs. 465)[1, 17]. The authors conclude that their model exceeds the performance of expert radiologists based on the difference in specificity and sensitivity of the model compared to the mean first reader, however it is not possible to confidently conclude this given the shortcomings in comparing an AUC to the average human performance.

3.3 Subsequent work

Following the publication of this work a significant volume of research has focussed on applying AI to breast cancer screening[21–23]. There are now multiple commercial enterprises developing and testing AI-assisted mammography software, with many testing their models in prospective observational studies[21, 22, 24] some even progressing to a randomised-control trial[23]. In February 2025, the UK government announced a national randomised-control trial in which five AI systems will be evaluated within the UK breast cancer screening programme with recruitment starting in April 2025[25]. The rapid pace of development from proof of concept to randomised-control trials in five years underscores the significant contributions that his paper made to the field and the importance of this research application.

4 Conclusion

In summary, McKinney *et al.* developed and evaluated a model that can accurately predict breast cancer from screening mammography in a representative population. The paper demonstrated that AI could be used to screen mammograms with a degree of accuracy that approached, or even exceeded human experts, potentially improving diagnostic performance as well as saving valuable and costly clinician time. Whilst criticism was levelled that the paper would be hard to reproduce due to a lack of detailed methodological description, many researchers have subsequently been successful in developing AI systems to improve breast cancer screening. Many of these AI systems are being actively evaluated in clinical trials and are getting closer to being rolled out more widely in the real-world.

5 Acknowledgments

This work was supported by UK Research and Innovation [UKRI AI Centre for Doctoral Training in Digital Healthcare grant number EP/Y030974/1].

References

1. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94. <https://www.nature.com/articles/s41586-019-1799-6> (2020).
2. Barclay, N. L. *et al.* Trends in incidence, prevalence, and survival of breast cancer in the United Kingdom from 2000 to 2021. *Sci. Rep.* **14**, 19069 (Aug. 2024).
3. *Breast cancer statistics* Cancer Research UK. <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer> (2025).
4. *Breast cancer screening* Cancer Research UK. <https://www.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/screening-breast> (2025).
5. US Preventive Services Task Force. Screening for Breast Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **331**, 1918–1930 (2024).
6. Lehman, C. D. *et al.* Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern Med* **175**, 1828–1837 (2015).
7. Lehman, C. D. *et al.* National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* **283**, 49–58. <https://pubs.rsna.org/doi/abs/10.1148/radiol.2016161174> (2017).
8. The Royal College of Radiologists. *Clinical radiology workforce census 2023* <https://www.rcr.ac.uk/news-policy/policy-reports-initiatives/clinical-radiology-census-reports/>.
9. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. *Focal Loss for Dense Object Detection* 2018. arXiv: 1708.02002 [cs.CV]. <https://arxiv.org/abs/1708.02002>.
10. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. *MobileNetV2: Inverted Residuals and Linear Bottlenecks* 2019. arXiv: 1801.04381 [cs.CV]. <https://arxiv.org/abs/1801.04381>.
11. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* ISBN: 1558604790 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988).
12. He, K., Zhang, X., Ren, S. & Sun, J. *Identity Mappings in Deep Residual Networks* 2016. arXiv: 1603.05027 [cs.CV]. <https://arxiv.org/abs/1603.05027>.
13. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* 2015. arXiv: 1512.03385 [cs.CV]. <https://arxiv.org/abs/1512.03385>.
14. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118. <https://www.nature.com/articles/nature21056> (2017).
15. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410. <https://doi.org/10.1001/jama.2016.17216> (2016).
16. Logan, J., Kennedy, P. J. & Catchpole, D. A review of the machine learning datasets in mammography, their adherence to the FAIR principles and the outlook for the future. *Sci Data* **10**, 595. <https://www.nature.com/articles/s41597-023-02430-6> (2023).
17. Wu, N. *et al.* Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging* **39**, 1184–1194. <https://ieeexplore.ieee.org/document/8861376/?arnumber=8861376> (2020).
18. Haibe-Kains, B. *et al.* Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16. <https://www.nature.com/articles/s41586-020-2766-y> (2020).

19. McKinney, S. M. *et al.* Reply to: Transparency and reproducibility in artificial intelligence. *Nature* **586**, E17–E18. <https://www.nature.com/articles/s41586-020-2767-x> (2020).
20. Oakden-Rayner, L. & Palmer, L. *Docs are ROCs: A simple off-the-shelf approach for estimating average human performance in diagnostic studies* 2020. arXiv: 2009.11060[stat]. <http://arxiv.org/abs/2009.11060>.
21. Ng, A. Y. *et al.* Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nat Med* **29**, 3044–3049. <https://www.nature.com/articles/s41591-023-02625-9> (2023).
22. Eisemann, N. *et al.* Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nat Med*, 1–8. <https://www.nature.com/articles/s41591-024-03408-6> (2025).
23. Hernström, V. *et al.* Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. *The Lancet Digital Health* **7**, e175–e183. [https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(24\)00267-X/fulltext](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(24)00267-X/fulltext) (2025).
24. Chang, Y.-W. *et al.* Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study. *Nat Commun* **16**, 2248. <https://www.nature.com/articles/s41467-025-57469-3> (2025).
25. Venkatesan, P. Largest trial of AI in breast cancer screening launched. *The Lancet Oncology* **26**, 285. [https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045\(25\)00080-4/fulltext](https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(25)00080-4/fulltext) (2025).