eScholarship@UMassChan

Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach

Item Type	Journal Article
Authors	Lotter, William;Diab, Abdul Rahman;Haslam, Bryan;Kim, Jiye G.;Grisot, Giorgia;Wu, Eric;Wu, Kevin;Onieva, Jorge Onieva;Boyer, Yun;Boxerman, Jerrold L.;Wang, Meiyun;Bandler, Mack;Vijayaraghavan, Gopal;Gregory Sorensen, A.
Citation	Lotter W, Diab AR, Haslam B, Kim JG, Grisot G, Wu E, Wu K, Onieva JO, Boyer Y, Boxerman JL, Wang M, Bandler M, Vijayaraghavan GR, Gregory Sorensen A. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat Med. 2021 Feb;27(2):244-249. doi: 10.1038/s41591-020-01174-9. Epub 2021 Jan 11. PMID: 33432172. Link to article on publisher's site
DOI	10.1038/s41591-020-01174-9
Download date	2025-02-22 10:03:52
Link to Item	https://hdl.handle.net/20.500.14038/48490

nature
medicine

Check for updates

Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach

William Lotter[1 ✉], Abdul Rahman Diab[1,9], Bryan Haslam[1,9], Jiye G. Kim[1,9], Giorgia Grisot[1], Eric Wu[1,7], Kevin Wu[1,8], Jorge Onieva Onieva[1], Yun Boyer[1], Jerrold L. Boxerman[2,3], Meiyun Wang[4], Mack Bandler[5], Gopal R. Vijayaraghavan[6] and A. Gregory Sorensen[1 ✉]

Breast cancer remains a global challenge, causing over 600,000 deaths in 2018 (ref. [1]). To achieve earlier cancer detection, health organizations worldwide recommend screening mammography, which is estimated to decrease breast cancer mortality by 20–40% (refs. [2,3]). Despite the clear value of screening mammography, significant false positive and false negative rates along with non-uniformities in expert reader availability leave opportunities for improving quality and access[4,5]. To address these limitations, there has been much recent interest in applying deep learning to mammography[6–18], and these efforts have highlighted two key difficulties: obtaining large amounts of annotated training data and ensuring generalization across populations, acquisition equipment and modalities. Here we present an annotation-efficient deep learning approach that (1) achieves state-of-the-art performance in mammogram classification, (2) successfully extends to digital breast tomosynthesis (DBT; '3D mammography'), (3) detects cancers in clinically negative prior mammograms of patients with cancer, (4) generalizes well to a population with low screening rates and (5) outperforms five out of five full-time breast-imaging specialists with an average increase in sensitivity of 14%. By creating new 'maximum suspicion projection' (MSP) images from DBT data, our progressively trained, multiple-instance learning approach effectively trains on DBT exams using only breast-level labels while maintaining localization-based interpretability. Altogether, our results demonstrate promise towards software that can improve the accuracy of and access to screening mammography worldwide.

Despite technological improvements such as the advent of DBT[19], studies that have reviewed mammograms where cancer was detected estimate that indications of cancer presence are visible 20–60% of the time in earlier exams that were interpreted as normal[20–22]. This missed cancer rate is driven in part by the difficulty of reading mammograms—abnormalities are often indicated by small, subtle features, and malignancies are present in approximately only 0.5% of screened women. These challenges are exacerbated by the high volume of mammograms (over 40 million per year in the

United States alone) and the additional time required to interpret DBT, both of which pressure radiologists to read faster.

Given these challenges, there have been many efforts in developing computer-aided diagnosis (CAD) software to assist radiologists in interpreting mammograms. The rationale behind initial versions of CAD was that even if its standalone performance was inferior to expert humans, it could still boost sensitivity when used as a 'second look' tool. In practice however, the effectiveness of traditional CAD has been questioned[23–26]. A potential reason for the limited accuracy of traditional CAD is that it relied on hand-engineered features. Deep learning relies instead on learning the features and classification decisions end-to-end. Applications of deep learning to mammography have shown great promise, including recent important work in which McKinney et al.[17] presented evidence of a system that exceeded radiologist performance in the interpretation of screening mammograms. Their system was trained and tested on two-dimensional (2D) mammograms from the United Kingdom and United States, and demonstrated generalization from training on UK data to testing on data collected from a US clinical site.

Despite such strong prior work, there remains meaningful room for improvement, especially in developing methods for DBT and demonstrating more widespread generalization. One of the key motivations for developing artificial intelligence (AI) applications for mammography lies in increasing access to screening. Demonstrating that the performance of an AI system generalizes to populations with currently low screening rates would be an important initial step towards a tool that could help alleviate the scarcity of expert clinicians[27,28]. Furthermore, AI would ideally demonstrate benefit in extending the window in which many cancers can be detected. Finally, given the rapid rise in the use of DBT for screening and its additional interpretation challenges, developing AI systems for DBT would be particularly impactful. In sum, prior efforts have illustrated the potential of applying AI to mammography, but further robustness and generalization are necessary to move towards true clinical utility.

Both data-related and algorithmic challenges contribute to the difficulty of developing AI solutions that achieve the aforementioned goals. Deep learning models generally perform best when trained on large amounts of highly-annotated data, which are

[1]DeepHealth Inc., RadNet AI Solutions, Cambridge, MA, USA. [2]Department of Diagnostic Imaging, Rhode Island Hospital, Providence, RI, USA. [3]Department of Diagnostic Imaging, Alpert Medical School of Brown University, Providence, RI, USA. [4]Department of Medical Imaging, Henan Provincial People's Hospital, Zhengzhou, Henan, China. [5]Medford Radiology Group, Medford, OR, USA. [6]Department of Radiology, University of Massachusetts Medical School, Worcester, MA, USA. [7]Present address: Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [8]Present address: Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [9]These authors contributed equally: Abdul Rahman Diab, Bryan Haslam, Jiye G. Kim. ✉e-mail: wlotter@deep.health; asorensen@deep.health
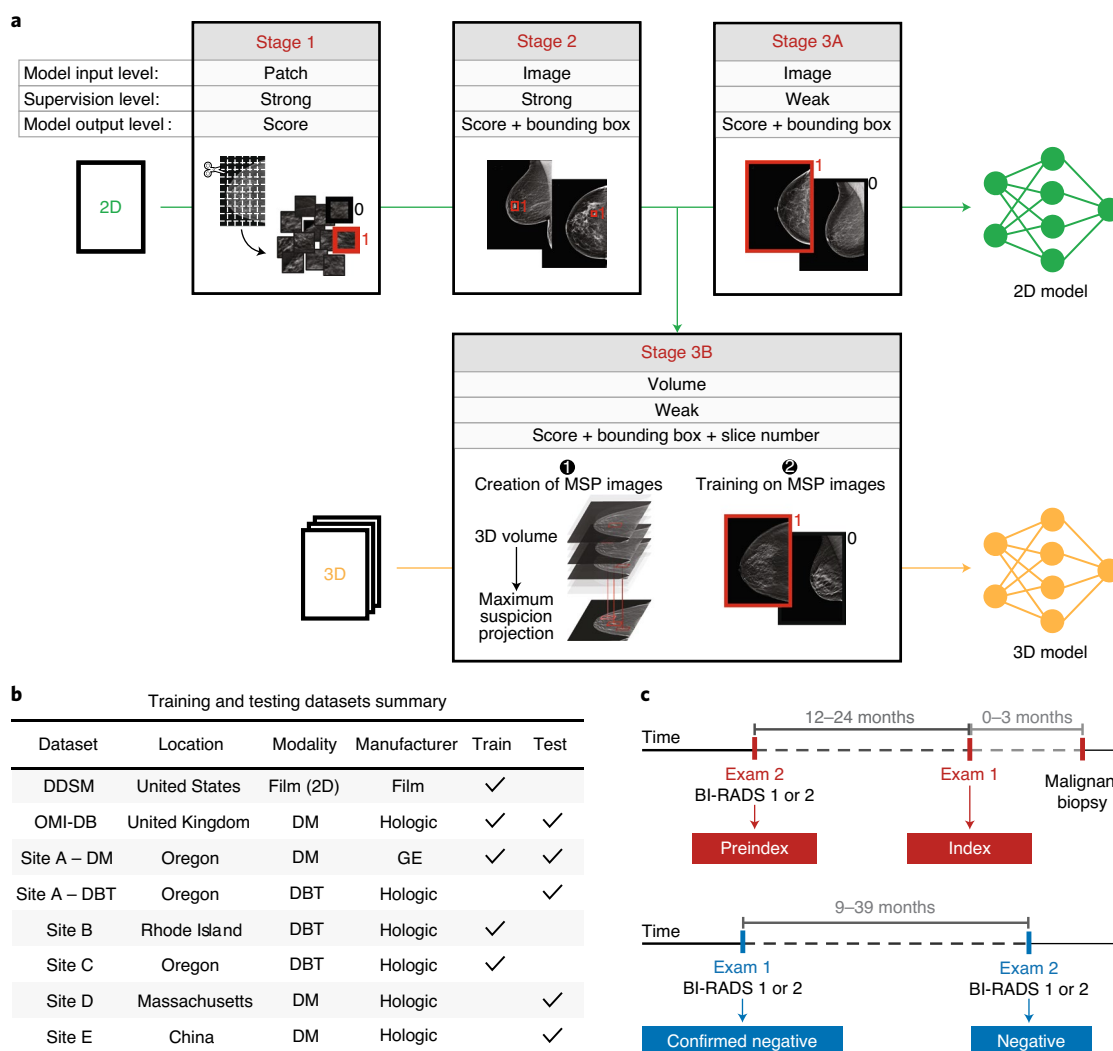
Fig. 1 | Model training approach and data summary. a, To effectively leverage both strongly and weakly labeled data while mitigating overfitting, we progressively trained our deep learning models in a series of stages. Stage 1 consists of patch-level classification using cropped image patches from 2D mammograms[15]. In Stage 2, the model trained in Stage 1 is used to initialize the feature backbone of a detection-based model. The detection model, which outputs bounding boxes with corresponding classification scores, is then trained end-to-end in a strongly supervised manner on full images. Stage 3 consists of weakly supervised training, for both 2D and 3D mammography. For 2D mammography (Stage 3A), the detection network is trained for binary classification in an end-to-end, multiple-instance learning fashion where an image-level score is computed as a maximum over bounding box scores. For 3D mammography (Stage 3B), the model from Stage 2 is used to condense each DBT stack into an optimized 2D projection by evaluating the DBT slices and extracting the most suspicious region of interest at each x–y spatial location. The model is then trained on these MSP images using the approach in Stage 3A. **b**, Summary of training and testing datasets. **c**, Illustration of exam definitions used here.

difficult to obtain for mammography. The two most prominent public datasets for mammography are the Digital Database of Screening Mammography (DDSM)[29] and the Optimam Mammography Imaging Database (OMI-DB)[30], both of which consist of 2D data. While relatively small compared with some natural image dataset standards[31], DDSM and OMI-DB both contain strong annotations (for example, expert-drawn bounding boxes around lesions). As opposed to weak annotations, such as knowing only the breast laterality of a cancer, strong annotations are particularly valuable for mammography given its 'needle in a haystack' nature. This is especially true for DBT, which can contain over 100 times as many images (or 'slices') as digital mammography (DM), and on which malignant features are often visible in only a few of the slices. This combination of large data size with small, subtle findings can cause deep learning models to memorize spurious correlations in the training set that do not generalize to independent test sets.

Strong annotations can mitigate such overfitting, but are costly and often impractical to collect. Nonetheless, many mammography AI efforts rely predominantly on strongly labeled data, and those that use weakly labeled data often lack a consistent framework to simultaneously train on both data types while maintaining intuitive localization-based interpretability[7,11,17,32].

Here we have taken steps to address both the data and algorithmic challenges of deep learning in mammography. We have assembled three additional training datasets, focusing especially on DBT, and have developed an algorithmic approach that effectively makes use of both strongly and weakly labeled data by progressively training a core model in a series of increasingly difficult tasks. We evaluate the resulting system extensively, assessing both standalone performance and performance in direct comparison with expert breast-imaging specialists in a reader study. In contrast to other recent reader studies[11,17], our study involves data from a site that was

a

i) ROC of index cases

ii)

	Sensitivity (%)		
Reader number	Reader	Proposed model	Difference
1	49.6	70.2	20.6 ± 10.1
2	85.5	96.2	10.7 ± 3.7
3	86.3	95.4	9.2 ± 3.3
4	94.7	97.7	3.1 ± 2.2
5	93.9	97.7	3.8 ± 2.5

iii)

	Specificity (%)		
Reader number	Reader	Proposed model	Difference
1	96.1	99.4	3.2 ± 1.8
2	68.2	89.6	21.4 ± 5.7
3	69.5	88.3	18.8 ± 5.4
4	51.9	73.4	21.4 ± 10.3
5	48.7	76.6	27.9 ± 10.1

b

i) ROC of pre-index cases

ii)

	Sensitivity (%)		
Reader number	Reader	Proposed model	Difference
1	22.5	29.2	6.7 ± 6.9
2	58.3	73.3	15.0 ± 6.7
3	52.5	69.2	16.7 ± 6.3
4	65.8	82.5	16.7 ± 6.9
5	80.2	86.7	6.7 ± 6.2

iii)

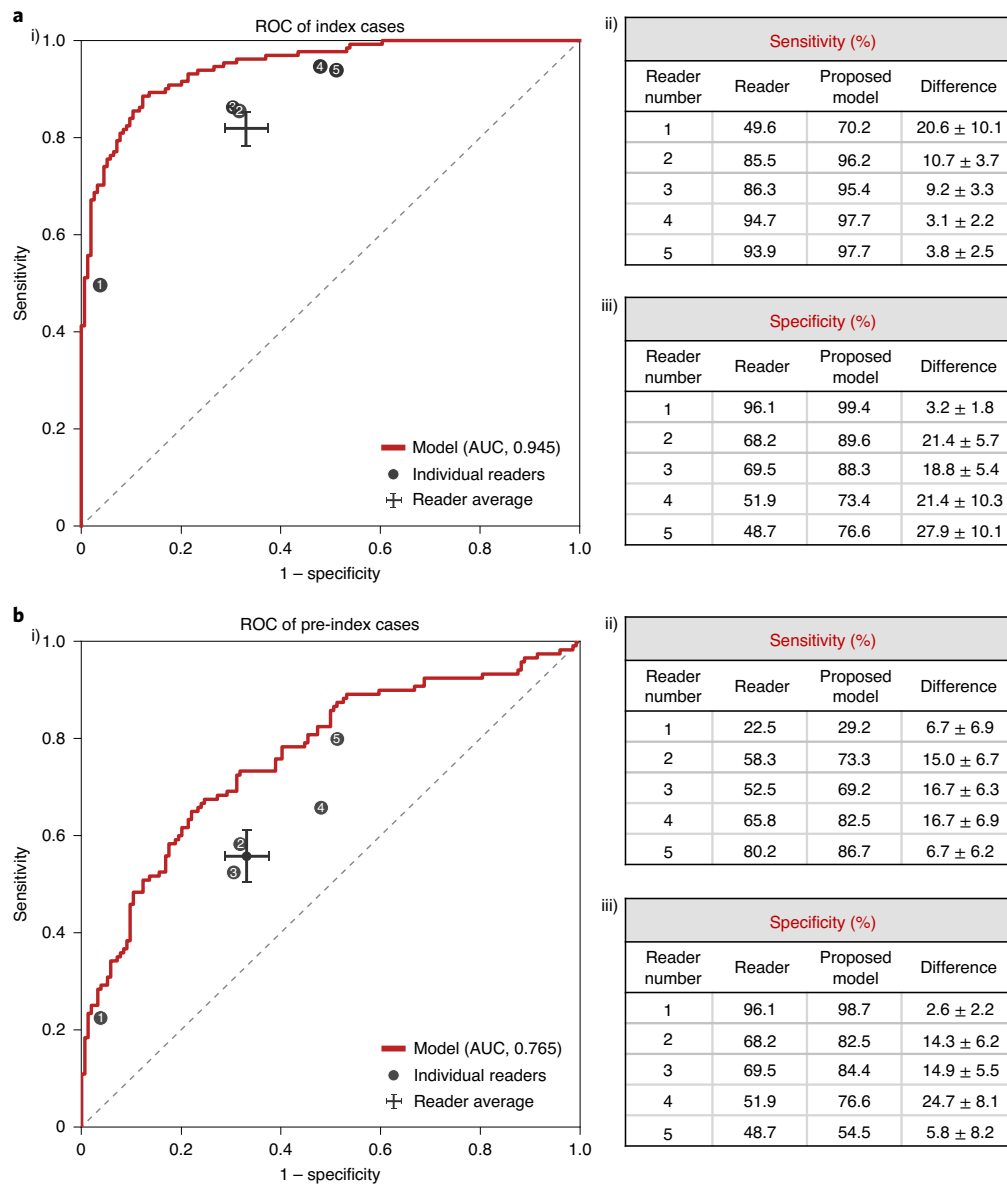|	Specificity (%)		
Reader number	Reader	Proposed model	Difference
1	96.1	98.7	2.6 ± 2.2
2	68.2	82.5	14.3 ± 6.2
3	69.5	84.4	14.9 ± 5.5
4	51.9	76.6	24.7 ± 8.1
5	48.7	54.5	5.8 ± 8.2

Fig. 2 | Reader study results. a, Index cancer exams and confirmed negatives. i, The proposed deep learning model outperformed all 5 radiologists on the set of 131 index cancer exams and 154 confirmed negatives. Each data point represents a single reader, and the ROC curve represents the performance of the deep learning model. The cross corresponds to the mean radiologist performance with the lengths of the cross indicating 95% confidence intervals. ii, Sensitivity of each reader and the corresponding sensitivity of the proposed model at a specificity chosen to match each reader. iii, Specificity of each reader and the corresponding specificity of the proposed model at a sensitivity chosen to match each reader. **b**, Pre-index cancer exams and confirmed negatives. i, The proposed deep learning model also outperformed all five radiologists on the early-detection task. The dataset consisted of 120 pre-index cancer exams—which are defined as mammograms interpreted as negative 12–24 months prior to the index exam in which cancer was found—and 154 confirmed negatives. The cross corresponds to the mean radiologist performance, with the lengths of the cross indicating 95% confidence intervals. ii, Sensitivity of each reader and the corresponding sensitivity of the proposed model at a specificity chosen to match each reader. iii, Specificity of each reader and the corresponding specificity of the proposed model at a sensitivity chosen to match each reader. For the sensitivity and specificity tables, the s.d. of the model minus reader difference was calculated via bootstrapping.

never used for model training to enable a more fair and generalizable comparison between AI and readers. Altogether, we use five test sets spanning different modalities (DM and DBT), different acquisition equipment manufacturers (General Electric (GE) and Hologic) and different populations (United States, United Kingdom and China). Additionally, our reader study investigates different timeframes of 'ground truth': (1) 'index' cancer exams, which are typically used in reader studies and are the screening mammograms acquired most recently prior to biopsy-proven malignancy; and (2) what we term 'pre-index' cancer exams—screening mammograms

acquired a year or more prior to the index cancer exams that were interpreted as normal by the clinician at the time of acquisition.

Figure 1 details our model training pipeline, our training and testing data and our definition of different exam types. In the first step of our approach, we train a convolutional neural network (CNN) to classify whether lesions are present in cropped image patches[15]. Next, we use this CNN to initialize the backbone of a detection-based model that takes an entire image as input and outputs bounding boxes with corresponding scores that indicate the likelihood that the corresponding enclosed region represents a

Fig. 3 | Examples of index and pre-index cancer exam pairs. Images from three patients with biopsy-proven malignancies are displayed. For each patient, an image from the index exam from which the cancer was discovered is shown on the right, and an image from the prior screening exam acquired 12–24 months earlier and interpreted as negative is shown on the left. From top to bottom, the number of days between the index and pre-index exams is 378, 629, and 414. The dots below each image indicate reader and model performance. Specifically, the number of infilled black dots represent how many of the five readers correctly classified the corresponding case, and the number of infilled red dots represent how many times the model would correctly classify the case if the model score threshold was individually set to match the specificity of each reader. The model is thus evaluated at five binary decision thresholds for comparison purposes, and we note that a different binary score threshold may be used in practice. Red boxes on the images indicate the model's bounding box output. White arrows indicate the location of the malignant lesion. **a**, A cancer that was correctly classified by all readers and the deep learning model at all thresholds in the index case, but detected by only the model in the pre-index case. **b**, A cancer that was detected by the model in both the pre-index and index cases, but detected by only one reader in the index case and zero readers in the pre-index case. **c**, A cancer that was detected by the readers and the model in the index case, but detected by only one reader in the pre-index case. The absence of a red bounding box indicates that the model did not detect the cancer.

malignancy. These first two stages both use strongly labeled data. In Stage 3A, we train the detection model on weakly labeled 2D data using a multiple-instance learning formulation where a maximum is computed over all of the bounding box scores. This results in a single image score that intuitively corresponds to the most suspicious region-of-interest (ROI) in the image. Importantly, even though the model at this stage is trained only with image-level labels, it retains its localization-based explainability, mitigating the 'black box' nature of standard classification models. Stage 3B consists of our weakly supervised training approach for DBT. Given a DBT volume, the strongly supervised 2D model is evaluated on each slice to produce a set of bounding boxes that are then filtered to retain the highest scoring box at each spatial location. The image patches defined by the boxes are then collapsed into a single 2D image array, which we term a MSP image. After creating the MSP images, the strongly supervised 2D detection model from Stage 2 is then trained on these images using the same multiple-instance learning formulation described above.

Figure 1b summarizes the data sources used to train and test our models. In addition to the OMI-DB and DDSM, we use datasets collected from three US clinical sites for training, denoted as Sites A, B and C. The data used for testing include test partitions of the OMI-DB and 'Site A – DM' datasets in addition to three datasets that were never used for model training or selection. These testing-only datasets include a screening DM dataset from a Massachusetts health system used for our reader study (Site D), a diagnostic DM dataset from an urban hospital in China (Site E) and a screening DBT dataset from a community hospital in Oregon (Site A – DBT). We note that we use screening mammograms for testing whenever possible, but the low screening rates in China necessitate using diagnostic exams (those in which the woman presents with symptoms) in the Site E dataset.

As briefly outlined above, we conducted a reader study using both 'index' and 'pre-index' cancer exams to directly compare model performance with that of expert radiologists in both regimes. Specifically, we define the index exams as mammograms acquired up to three months prior to biopsy-proven malignancy (Fig. 1c). We define pre-index exams as those that were acquired 12–24 months prior to the index exams and were interpreted as negative in clinical practice. Following the Breast Imaging Reporting and Data System
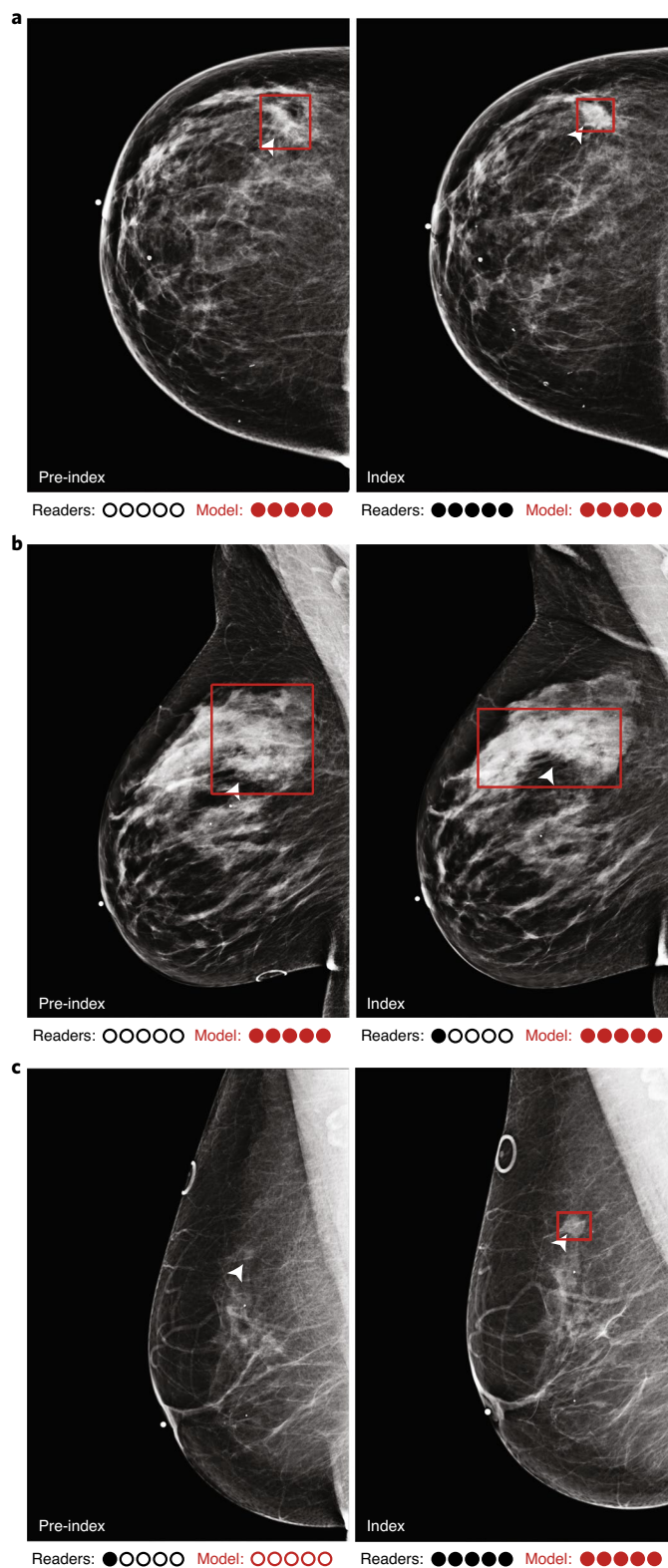
(BI-RADS) standard[33], we consider a BI-RADS score of 1 or 2 as a negative interpretation and further define a 'confirmed negative' as a negative exam followed by an additional BI-RADS 1–2 screen. All of the negatives used in our reader study are confirmed negatives.

Figure 2a summarizes the results of the 'index' component of the reader study. The study involved five radiologists, each fellowship-trained in breast imaging and practicing full-time in the field. The data consisted of screening DM cases retrospectively

Table 1 | Summary of additional DM and DBT evaluation

Dataset	Location	Manufacturer	Model	Input type	AUC
OMI-DB	UK	Hologic	2D	DM	0.963 ± 0.003
Site A – DM	Oregon	GE	2D	DM	0.927 ± 0.008
Site E	China	Hologic	2D	DM	0.971 ± 0.005
Site E (resampled)	China	Hologic	2D	DM	0.956 ± 0.020
Site A – DBT	Oregon	Hologic	2D*	DBT manufacturer synthetics	0.922 ± 0.016
Site A – DBT	Oregon	Hologic	3D	DBT slices	0.947 ± 0.012
Site A – DBT	Oregon	Hologic	2D + 3D	DBT manufacturer synthetics + slices	0.957 ± 0.010

All results correspond to using the index exam for cancer cases and confirmed negatives for the non-cancer cases, except for Site E where the negatives are unconfirmed. Pre-index results, where possible, and additional analysis are included in Extended Data Fig. 8. Rows 1 and 2, performance of the 2D deep learning model on held-out test sets of the OMI-DB (1,205 cancers, 1,538 negatives) and Site A (254 cancers, 7,697 negatives) datasets. Rows 3 and 4, performance on a dataset collected at a Chinese hospital (Site E; 533 cancers, 1,000 negatives). The dataset consists entirely of diagnostic exams given the low prevalence of screening mammography in China. Nevertheless, even when adjusting for tumor size using bootstrap resampling to approximate the distribution of tumor sizes expected in an American screening population (see Methods), the model still achieves high performance (Row 4). Rows 5–7, performance on DBT data (Site A – DBT; 78 cancers, 518 negatives). Row 5 contains results of the 2D model fine-tuned on the manufacturer-generated synthetic 2D images, which are created to augment/substitute DM images in a DBT study (the '*' symbol indicates this fine-tuned model). Row 6 contains the results of the weakly supervised 3D model, illustrating strong performance when evaluated on the MSP images computed from the DBT slices. We note that when scoring the DBT volume as the maximum bounding box score over all of the slices, the strongly supervised 2D model used to create the MSP images exhibits an AUC of 0.865 ± 0.020. Thus, fine-tuning this model on the MSP images significantly improves its performance. Row 7, results when combining predictions across the final 3D model and 2D models. The s.d. for each AUC value was calculated via bootstrapping.

collected from a regional health system located in a different US state than any of the sources of training data. Figure 2a contains a receiver operating characteristic (ROC) plot based on case-level performance comparing the readers with the proposed deep learning model on the set of 131 index cancer exams and 154 confirmed negatives. The points representing each reader all fall below the model's ROC curve, indicating that the model outperformed all five radiologists. At the average reader specificity, the model achieved an absolute increase in sensitivity of 14.2% (95% confidence interval (CI): 9.2–18.5%; $P < 0.0001$). At the average reader sensitivity, the model achieved an absolute increase in specificity of 24.0% (95% CI: 17.4–30.4%; $P < 0.0001$). Reader ROC curves based on a continuous 'probability of malignancy' score are also contained in Extended Data Fig. 1 and illustrate similar higher performance by the model. Additionally, the model outperformed every simulated combination of the readers (Extended Data Fig. 2) and also compares favorably to other recently published models on this dataset[7,11,32] (Extended Data Fig. 3).

Figure 2b summarizes the second component of the reader study involving pre-index exams from the same patients. Pre-index exams can largely be thought of as challenging false negatives, as studies estimate that breast cancers typically exist 3+ years prior to detection by mammography[34,35]. The deep learning model outperformed all five readers in the early detection, pre-index paradigm as well. The absolute performances of the readers and the model were lower on the pre-index cancer exams than on the index cancer exams, which is expected given the difficulty of these cases. Nonetheless, the model still demonstrated an absolute increase in sensitivity of 17.5% (95% CI: 6.0–26.2%; $P = 0.0009$) at the average reader specificity, and an absolute increase in specificity of 16.2% (95% CI: 7.3–24.6%; $P = 0.0008$) at the average reader sensitivity. At a specificity of 90% (ref. [36]), the model would have flagged 45.8% (95% CI: 28.8–57.1%) of the pre-index ('missed') cancer cases for additional workup. The model additionally exhibited higher performance than recently published models on the pre-index dataset as well[7,11,32] (Extended Data Fig. 4).

Given the interpretable localization outputs of the model, it is possible to evaluate sensitivity while requiring correct localization. For both laterality-level and quadrant-level localization, we find that the model again demonstrated improvements in sensitivity for both the index and pre-index cases, as detailed in Extended Data Fig. 5. Examples of pre-index cancer cases detected by the model are shown in Fig. 3. The trend of higher model performance also holds when considering factors such as lesion type, cancer type, cancer size and breast density (Extended Data Fig. 6). Nevertheless, there are examples in which the model missed cancers that were detected by the readers, and vice versa (Extended Data Fig. 7).

Building upon the reader study performance, we evaluated standalone performance of our approach on larger, diverse datasets spanning different populations, equipment manufacturers and modalities. These results are summarized in Table 1, which are calculated using index cancer exams. Additional results using pre-index exams and other case definitions are contained in Extended Data Fig. 8, and a summary of performance across all datasets is contained in Extended Data Fig. 9. Beginning with a test partition of the OMI-DB including 1,205 cancers and 1,538 confirmed negatives, our approach exhibits strong performance on DM exams from a UK screening population with an area under the curve (AUC) of 0.963 ± 0.003 (0.961 ± 0.003 using all 1,967 negatives, confirmed and unconfirmed; s.d. for AUC was calculated via bootstrapping). On a test partition of the Site A – DM dataset with 254 cancers and 7,697 confirmed negatives, the model achieved an AUC of 0.927 ± 0.008 (0.931 ± 0.008 using all 16,369 negatives), which is not statistically different from the results on the other tested US screening DM dataset (Site D; $P = 0.22$). The Site A – DM dataset consists of mammograms acquired using GE equipment, as opposed to the Hologic equipment used for the majority of the other datasets.

To further test the generalizability of our model, we assessed performance on a DM dataset collected at an urban Chinese hospital (Site E). Testing generalization to this dataset is particularly meaningful given the low screening rates in China[28] and the known (and potentially unknown) biological differences found in mammograms between Western and Asian populations, including a greater proportion of women with dense breasts in Asian populations[37]. The deep learning model, which was evaluated locally at the Chinese hospital, generalized well to this population, achieving an AUC of 0.971 ± 0.005 (using all negatives – 'confirmation' is not possible given the lack of follow-up screening). Even when adjusting for tumor size to approximately match the statistics expected in an American screening population, the model achieved an AUC of 0.956 ± 0.020 (see Table 1 and Methods).

Finally, our DBT approach performs well when evaluated at a site not used for DBT model training. Our method, which generates an optimized MSP image from the DBT slices and then classifies this image, achieved an AUC of 0.947 ± 0.012 (with 78 cancers and 519 confirmed negatives; 0.950 ± 0.010 using all 11,609 negative exams). If we instead simply fine-tune our strongly supervised 2D model on the manufacturer-generated synthetic 2D images that are generated by default with each DBT study, the resulting model achieved 0.922 ± 0.016 AUC on the test set (0.923 ± 0.015 AUC using all negatives). Averaging predictions across the manufacturer-generated synthetic images and our MSP images results in an overall performance

of 0.957 ± 0.010 (0.959 ± 0.008 using all negatives). Examples of the MSP images can be found in Extended Data Fig. 10, which illustrate how the approach can be useful in mitigating tissue superposition compared with the manufacturer-generated synthetic 2D images.

In summary, we have developed a deep learning approach that effectively leverages both strongly and weakly labeled data by progressively training in stages while maintaining localization-based interpretability. Our approach also extends to DBT, which is especially important given its rising use as state-of-the-art mammography screening and the additional time required for its interpretation. In a reader study, our system outperformed five out of five full-time breast-imaging specialists. This performance differential occurred on both the exams in which cancers were found in practice and the prior exams of these cancer cases. Nevertheless, prospective clinical cohort studies will ultimately provide the best comparison to the current standard of care. Furthermore, while we have aimed to test performance across various definitions of 'positive' and 'negative', assigning ground truth is non-trivial for screening mammography and further real-world, regulated validation is needed before clinical use[38]. An encouraging aspect regarding generalization, nonetheless, is that our reader study involved data from a site never used for model development. We additionally observe similar levels of performance in four other larger datasets, including independent data from a Chinese hospital. One particular reason why our system may generalize well is that it has also been trained on a wide array of sources, including five datasets in total. Altogether, our results show great promise towards earlier cancer detection and improved access to screening mammography using deep learning.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-01174-9.

References

1. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
2. Berry, D. A. et al. Effect of screening and adjuvant therapy on mortality from breast cancer. *N. Engl. J. Med.* **353**, 1784–1792 (2005).
3. Seely, J. M. & Alhassan, T. Screening for breast cancer in 2018—what should we be doing today? *Curr. Oncol.* **25**, S115–S124 (2018).
4. Majid, A. S., Shaw De Paredes, E., Doherty, R. D., Sharma, N. R. & Salvador, X. Missed breast carcinoma: pitfalls and pearls. *Radiographics* **23**, 881–895 (2003).
5. Rosenberg, R. D. et al. Performance benchmarks for screening mammography. *Radiology* **241**, 55–66 (2006).
6. Yala, A., Lehman, C., Schuster, T., Portnoi, T. & Barzilay, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**, 60–66 (2019).
7. Yala, A., Schuster, T., Miles, R., Barzilay, R. & Lehman, C. A deep learning model to triage screening mammograms: a simulation study. *Radiology* **293**, 38–46 (2019).
8. Conant, E. F. et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol. Artif. Intell.* **1**, e180096 (2019).
9. Rodriguez-Ruiz, A. et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J. Natl Cancer Inst.* **111**, 916–922 (2019).
10. Rodríguez-Ruiz, A. et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* **290**, 305–314 (2019).
11. Wu, N. et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans. Med. Imaging* **39** 1184–1194 (2019).
12. Ribli, D., Horváth, A., Unger, Z., Pollner, P. & Csabai, I. Detecting and classifying lesions in mammograms with deep learning. *Sci. Rep.* **8**, 4165 3 (2018).
13. Kooi, T. et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017).
14. Geras, K. J. et al. High-resolution breast cancer screening with multi-view deep convolutional neural networks. Preprint at https://arxiv.org/abs/1703.07047 (2017).
15. Lotter, W., Sorensen, G., and 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* (eds. Cardoso, M. J. et al.) (2017).
16. Schaffter, T. et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw. Open* **3**, e200265 (2020).
17. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
18. Kim, H.-E. et al Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit. Health* **2**, e138–e148 (2020).
19. Kopans, D. B. Digital breast tomosynthesis from concept to clinical care. *Am. J. Roentgenol.* **202**, 299–308 (2014).
20. Saarenmaa, I. et al. The visibility of cancer on earlier mammograms in a population-based screening programme. *Eur. J. Cancer* **35**, 1118–1122 7 (1999).
21. Ikeda, D. M., Birdwell, R. L., O'Shaughnessy, K. F., Brenner, R. J. & Sickles, E. A. Analysis of 172 subtle findings on prior normal mammograms in women with breast cancer detected at follow-up screening. *Radiology* **226**, 494–503 (2003).
22. Hoff, S. R. et al. Missed and true interval and screen-detected breast cancers in a population based screening program. *Acad. Radiol.* **18**, 454–460 (2011).
23. Fenton, J. J. et al. Influence of computer-aided detection on performance of screening mammography. *N. Engl. J. Med.* **356**, 1399–1409 4 (2007).
24. Lehman, C. D. et al. Diagnostic accuracy of digital screening mammography with and without computer aided detection. *JAMA Intern. Med.* **33**, 839–841 (2016).
25. Henriksen, E. L., Carlsen, J. F., Vejborg, I. M., Nielsen, M. B. & Lauridsen, C. A. The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. *Acta Radiol.* **60**, 13–18 1 (2019).
26. Tchou, P. M. et al. Interpretation time of computer-aided detection at screening mammography. *Radiology* **257**, 40–46 (2010).
27. Bowser, D., Marqusee, H., Koussa, M. E. & Atun, R. Health system barriers and enablers to early access to breast cancer screening, detection, and diagnosis: a global analysis applied to the MENA region. *Public Health* **152**, 58–74 (2017).
28. Fan, L. et al. Breast cancer in China. *Lancet Oncol.* **15**, e279–e289 (2014).
29. 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* (ed Yaffe, M. J.) 212–218 (2001).
30. Halling-Brown, M. D. et al. OPTIMAM mammography image database: a large scale resource of mammography images and clinical data. Preprint at https://arxiv.org/abs/2004.04742 (2020).
31. Deng, J. et al. ImageNet: a large-scale hierarchical image database. in *Conference on Computer Vision and Pattern Recognition* (2009).
32. Wu, K. et al. Validation of a deep learning mammography model in a population with low screening rates. in *Fair ML for Health Workshop. Neural Information Processing Systems* (2019).
33. Sickles, E., D'Orsi, C. & Bassett, L. *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System* 5th edn (American College of Radiology, 2013).
34. Hart, D., Shochat, E. & Agur, Z. The growth law of primary breast cancer as inferred from mammography screening trials data. *Br. J. Cancer* **78**, 382–387 (1998).
35. Weedon-Fekjær, H., Lindqvist, B. H., Vatten, L. J., Aalen, O. O. & Tretli, S. Breast cancer tumor growth estimated through mammography screening data. *Breast Cancer Res.* **10**, R41 (2008).
36. Lehman, C. D. et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology* **283**, 49–58 (2017).
37. Bae, J.-M. & Kim, E. H. Breast density and risk of breast cancer in Asian women: a meta-analysis of observational studies. *J. Preventive Med. Public Health* **49**, 367 (2016).
38. Park, S. H. Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance. *Radiology* **290**, 272–273 (2018).

Methods

Ethical approval. All non-public datasets (data from Sites A, B, C, D and E) were collected under Institutional Review Board (IRB) approval. The following review boards were used for each dataset: Site A, Southern Oregon IRB; Site B, Rhode Island Hospital IRB; Site C, Providence IRB; Site D, Advarra IRB; and Site E, Henan Provincial People's Hospital IRB. All of the datasets used in the study were de-identified prior to model training and testing.

Dataset descriptions. Details of all utilized datasets are provided below. Each dataset was partitioned into one or more of the following splits: training, model selection and/or testing. The model-selection split was specifically used to choose final models and to determine when to stop model training. Data splits were created at the patient level, meaning that exams from a given patient were all in the same split. Rules for label assignment and case selection for training data varied slightly across datasets given variability in collection time periods and available metadata (as described below). However, the definitions of testing sets and label criteria were standardized across datasets unless otherwise stated. In the main text, the following definitions were used in assigning labels (as summarized in Fig. 1c): index cancer, a mammogram obtained within the 3 months preceding a cancer diagnosis; pre-index cancer, a mammogram interpreted as BI-RADS category 1 or 2 and obtained 12–24 months prior to an index exam; negative, a mammogram interpreted as BI-RADS 1 or 2 from a patient with no known prior or future history of breast cancer; confirmed negative, a negative exam followed by an additional BI-RADS 1 or 2 interpretation at the next screening exam 9–39 months later (which represents 1–3 yr of follow-up depending on the screening paradigm with a 3-month buffer). We extend the time window beyond 3 yr to include triennial screening (for example, as performed in the United Kingdom). In Extended Data Fig. 8, we include additional results using a 12-month time window for defining an index cancer exam, as well as including pathology-proven benign cases. Throughout, we treat pre-index cases as positives because, while it is not guaranteed that a pathology-proven cancer could have been determined with appropriate follow-up, it is likely that cancer existed at the time of acquisition for the vast majority of these exams[34,35].

All datasets shared the same algorithm for creating test sets, except Site D (which is described in detail in the corresponding section below). Studies were labeled as 'index', 'pre-index', 'confirmed negative', 'unconfirmed negative' or 'none' on the basis of the aforementioned criteria. For each patient in the test set, one study was chosen in the following order of descending priority: 'index', 'pre-index', 'confirmed negative' or 'unconfirmed negative'. If a patient had multiple exams with the chosen label, one exam was randomly sampled. If a patient had an index exam, a single pre-index exam was also included when available. For all training and testing, only craniocaudal (CC) and mediolateral oblique (MLO) mammographic views were used. All test sets included only screening exams except for Site E, for which all tested exams were diagnostic given the low screening rates in China. A summary of all the testing datasets and corresponding results is contained in Extended Data Fig. 9. We note that the proportion of confirmed versus unconfirmed negatives varies by site largely because of differing time periods of exam acquisition (for example, not enough time may have passed for confirmation for some exams), screening paradigms and/or BI-RADS information collection ranges. We report performance using both confirmed and unconfirmed negatives when possible to consider results on a stringent definition of negative while also evaluating on larger amounts of data.

For training, labeling amounts to assigning each training instance (either an image or bounding box) a label of '1' for cancer and '0' for non-cancer. The chief decision for assigning images a label of '1' (cancer) is in the time window allowed between cancer confirmation (biopsy) and image acquisition. For US datasets, we set this window to 15 months. This time window was chosen to balance the risk of overfitting with still including some pre-index cancer exams for training. Localization annotations were not available for the US datasets (except DDSM, which has only index cancers), so extending the time window further could lead to overfitting on more subtle cancers. Nonetheless, the mix of yearly and biyearly screening in the United States enables the inclusion of some pre-index cancers using a 15-month time window. For the OMI-DB from the United Kingdom, we extend this window by a year since this dataset includes a high proportion of strongly labeled data and because the standard screening interval is longer in the United Kingdom. For non-cancers, unless otherwise noted, we use screening negative exams (BI-RADS 1 or 2) from patients with no history of cancer and, when available, pathology-confirmed benign cases from patients with no history of cancer. For the pathology-confirmed benign cases, we trained on both screening and diagnostic exams. For cancers, we additionally included both screening and diagnostic exams for training. We have found that training on diagnostic exams can improve performance even when evaluating on only screening exams (and vice versa). The only dataset where we train exclusively on screening exams is the Site A – DM dataset, where the lack of benign biopsy information would entail that all of the diagnostic exams to be included in training would be cancers, so we exclude diagnostics altogether to avoid such bias. As opposed to model testing where only one exam per patient is included, we use all qualified exams for a given patient for training. Below, we provide additional details of the datasets.

Digital Database of Screening Mammography. DDSM is a public database of scanned film mammography studies from the United States containing cases categorized as normal, benign, and malignant with verified pathology information[29]. The dataset includes radiologist-drawn segmentation maps for every detected lesion. We split the data into 90%/10% training/model selection splits, resulting in 732 cancer, 743 benign and 807 normal studies for training. We did not use any data from DDSM for testing given that it is a scanned film dataset.

OPTIMAM Mammography Imaging Database. The OMI-DB is a publicly available dataset from the United Kingdom, containing screening and diagnostic digital mammograms primarily obtained using Hologic equipment[30]. We split the unique list of patients into 60%/20%/20% training/model selection/testing splits. This results in a training set of 5,233 cancer studies (2,332 with bounding boxes), 1,276 benign studies (296 with bounding boxes) and 16,887 negative studies. We note that although the proportion of positives to negatives in OMI-DB is much higher than the ratio expected in a screening population, the positives and negatives themselves are randomly sampled from their respective populations. Thus, given the invariance of ROC curves to incidence rates, we would not expect bias in the test set AUC in this population compared to the full population with a natural incidence rate.

Site A. Site A is a community hospital in Oregon. The dataset from Site A primarily consists of screening mammograms, with DM data from 2010 to 2015 collected almost entirely from GE equipment, and DBT data from 2016 to 2017 collected almost entirely from Hologic equipment. For the DM data, 40% of the patients were used for training, 20% were used for model selection and 40% were used for testing. We use the DBT data solely for testing, given its high proportion of screening exams compared with the other utilized DBT datasets. Ground-truth cancer status for both modalities was obtained using a local hospital cancer registry. A radiology report also accompanied each study and contained BI-RADS information. For the DBT data, a list of benigns was additionally provided by the hospital, but such information was not available for the DM data. Given the extent of longitudinal data present in the DM dataset and the lack of confirmed benign pathology information for this data, we are slightly more strict when choosing the non-cancers for training, specifically requiring the negatives to have no record of non-screening procedures or non-normal interpretations for the patient for 18 months prior to and following the exam. This results in 466 cancer studies and 48,248 negative studies for training in the Site A – DM dataset.

Site B. Site B consists of an inpatient medical center and affiliated imaging centers in Rhode Island. The data from this site contain DBT mammograms from Hologic equipment, with a mix of screening and diagnostic exams collected retrospectively between 2016 and 2017. Cancer status, benign results and BI-RADS were determined using a local database. We split the list of unique patients into 80%/20% training/model selection splits. Given the relatively smaller amount of DBT available for training and the desire to test on datasets not used for training, Site B was solely used for model development. The training split consists of 13,767 negative cases, 379 benign cases and 263 cancer cases. We note that the manufacturer-generated synthetic 2D images were also included in the weakly supervised training for the final 2D model.

Site C. Site C is a health system in Oregon separate from the one in Site A. From Site C, DBT cases were retrospectively collected between 2017 and 2018. The data consist of a mix of screening and diagnostic cases acquired almost entirely using Hologic equipment. We split the unique list of patients into 70%/30% training/model selection splits. A regional cancer registry was used to determine cancer status. Like Site B, we use Site C solely for model development. Historical BI-RADS information was not readily available for all of the cases in Site C, so we use cases from patients with no entry in the regional cancer registry as non-cancers for training. Given the geographic proximity of Site C and Site A, we exclude a small number of patients that overlap in both sets when performing testing on Site A. We note that the manufacturer-generated synthetic 2D images were also included in the weakly supervised training for the final 2D model.

Site D. Data from Site D were used for the reader study and consisted of 405 screening DM exams that were collected retrospectively from a single health system in Massachusetts with 4 different imaging collection centers. No data from this site were ever used for model training or selection. The exams included in the study were acquired between July 2011 and June 2014. Out of the 405 studies, 154 were negative, 131 were index cancer exams and 120 were pre-index cancer exams. All of the negatives were confirmed negatives. The index cancer exams were screening mammograms interpreted as suspicious and confirmed to be malignant by pathology within three months of acquisition. The pre-index exams came from the same set of women as the index exams and consisted of screening exams that were interpreted as BI-RADS 1 or 2 and acquired 12–24 months prior to the index exams. All studies were acquired using Hologic equipment. Case selection was conducted over several steps. First, the patients included in the study were selected by taking all patients with qualifying index and pre-index exams over the specified time period using a local cancer registry. Due to PACS limitations, it was not

possible to obtain some pre-index cases. Next, the non-cancer cases were chosen to have a similar distribution in patient age and breast density compared with the cancer cases using bucketing. In total, 154 non-cancer, 131 index cancer and 120 pre-index cancer mammograms were collected from 285 women. Additional details on case composition are contained in Extended Data Fig. 6, including breast density for all patients and cancer type, cancer size and lesion type for the index cancer exam. Breast density and lesion type were obtained from the initial radiology reports. Cancer type and size were obtained from pathology reports.

Site E. Site E consists of a dataset from an urban hospital in China collected retrospectively from a contiguous period between 2012 and 2017. Over this time period, all pathology-proven cancers were collected along with a uniformly random sample of non-cancers, resulting in 533 cancers, 1,000 negatives (BI-RADS 1 or 2 interpretation) and 100 pathology-proven benigns. Due to the low screening rates in China, the data came from diagnostic exams, so the distribution of tumor sizes from the cancer cases contained more large tumors (for example, 64% larger than 2 cm) than would be expected in a typical US screening population. For better comparison with a US screening population, results on Site E were also calculated using a bootstrap resampling method to approximately match the distribution of tumor sizes from a US population according to the American College of Radiology National Radiology Data Registry (https://nrdr.acr.org/Portal/NMD/Main/page.aspx). Using this approach, a mean AUC was computed over 5,000 bootstrapped populations. Site E was solely used for testing and never for model development. Furthermore, the deep learning system was evaluated locally at the hospital and data never left the site.

Model development and training. The first stage of model training consisted of patch-level classification[15] (Stage 1 in Fig. 1). Patches of size 275 × 275 pixels were created from the DDSM and OMI-DB datasets after the original images were resized to a height of 1,750 pixels. Data augmentation was also used when creating the patches, including random rotations of up to 360°, image resizing by up to 20% and vertical mirroring. Preprocessing consisted of normalizing pixel values to a range of (−127.5, 127.5). When creating patches containing lesions, a random location within the lesion boundary was selected as the center of the patch. If the resulting patch had fewer than 6 pixels containing the lesion mask, the patch was discarded and a new patch was sampled. For all patches, if the patch contained <10% of the breast foreground, as determined by Otsu's method[39] for DDSM and by thresholding using the minimal pixel value in the image for OMI-DB, then the patch was discarded. In total, two million patches were created with an equal number of patches with and without lesions. For the patch classification model, we use a popular convolutional neural network, ResNet-50 (ref. [40]). The patch-based training stage itself consisted of two training sequences. First, starting from ImageNet[31] pretrained weights, the ResNet-50 model was trained for five-way classification of lesion type: mass, calcifications, focal asymmetry, architectural distortion or no lesion. Patches from DDSM and OMI-DB were sampled in proportion to the number of cancer cases in each dataset. The model was trained for 62,500 batches with a batch size of 16, sampling equally from all lesion types. The Adam optimizer[41] was used with a learning rate of 1×10^{-5}. Next, the patch-level model was trained for three-way classification using labels of normal, benign or malignant, again sampling equally from all categories. The same training parameters were also used for this stage of patch-level training.

After patch-level training, the ResNet-50 weights were used to initialize the backbone of a popular detection model, RetinaNet[42], for the second stage of training: strongly supervised, image-level training (Stage 2 in Fig. 1). Image preprocessing consisted of resizing to a height of 1,750 pixels (maintaining the original aspect ratio), cropping out the background using the thresholding methods described above and normalizing pixel values to a range of (−127.5, 127.5). Data augmentation during training included random resizing of up to 15% and random vertical mirroring. Given the high class imbalance of mammography (far fewer positives than negatives), we implemented class balancing during training by sampling malignant and non-malignant examples with equal probability[7,15,16]. This class balancing was additionally implemented within datasets to prevent the model from learning biases in the different proportions of cancers across datasets. For this strongly supervised, image-level training stage, we use the bounding boxes in the OMI-DB and DDSM datasets. Three-way bounding box classification was performed using labels of normal, benign or malignant. The RetinaNet model was trained for 100,000 iterations with a batch size of 1. The Adam optimizer[41] was used, with a learning rate of 1×10^{-5} and gradient norm clipping with a value of 0.001. Default hyperparameters were used in the RetinaNet loss, except for a weight of 0.5 that was given to the regression loss and a weight of 1.0 that was given to the classification loss.

For the weakly supervised training stage (Stage 3 in Fig. 1), binary cancer/no-cancer classification was performed with a binary cross entropy loss. The same image input processing steps were used as in the strongly supervised training stage. The RetinaNet architecture was converted to a classification model by taking a maximum over all of the bounding box classification scores, resulting in a model that remains fully differentiable while allowing end-to-end training with binary labels. For 2D, training consisted of 300,000 iterations using the Adam optimizer[41], starting with a learning rate of 2.5×10^{-6}, which was decreased by a factor of 4

every 100,000 iterations. Final model weights were chosen by monitoring AUC performance on the validation set every 4,000 iterations.

For DBT, our MSP approach was motivated by the value of DBT in providing an optimal view into a lesion that could otherwise be obscured by overlapping tissue, and by the similarity between DBT and DM images which suggests the applicability of transfer learning. Furthermore, we especially consider that the aggregate nature of 2D mammography can help reduce overfitting compared with training end-to-end on a large DBT volume. To this end, in Stage 3B the MSP images were created using the model resulting from 2D strongly supervised training as described above, after an additional 50,000 training iterations with a learning rate of 2.5×10^{-6}. To create the MSP images, the 2D model was evaluated on every slice in a DBT stack except for the first and last 10% of slices (which are frequently noisy). A minimal bounding box score threshold was set at a level that achieved 99% sensitivity on the OMI-DB validation set. The bounding boxes over all evaluated slices were filtered using non-maximum suppression using an intersection-over-union threshold of 0.2. The image patches defined by the filtered bounding boxes were then collapsed into a single 2D image array representing an image optimized for further model training. Any 'empty' pixels in the projection were infilled with the corresponding pixels from the center slice of the DBT stack, resulting in the final MSP image. Overall, the MSP process is akin to a maximum-intensity projection except that the maximum is computed over ROI malignancy suspicion predicted by an AI model instead of over pixel-level intensity. Training on the resulting MSP images was conducted similar to the 2D weakly supervised approach, except that the model was trained for 100,000 iterations. The input processing parameters used for 2D images were reused for DBT slices and MSP images.

After weakly supervised training for both the 2D and 3D models, we fine-tune the regression (that is, localization) head of the RetinaNet architecture on the strongly labeled data used in Stage 2. Specifically, the backbone and classification head of the network are frozen, and only the regression head is updated during this fine-tuning. This allows the regression head to adapt to any change in the weights in the backbone of the network during the weakly supervised training stage, where the regression head is not updated. For this regression fine-tuning stage, the network is trained for 50,000 iterations with a learning rate of 2.5×10^{-6} using the same preprocessing and data augmentation procedures as the previous stages.

Final model selection was based on performance on the held-out model selection data partition. The final model was an aggregation of three equivalently-trained models starting from different random seeds. A prediction score for a given image was calculated by averaging across the three models' predictions for both horizontal orientations of the image (resulting in an average over six scores). Regression coordinates of the bounding box anchors were additionally averaged across the three models. Each breast was assigned a malignancy score by taking the average score over all of its views. Each study was assigned a score by taking the greater of its two breast-level scores. Finally, while we note that random data augmentation was used during model training as described above, data processing during testing is deterministic, as are the models themselves.

The deep learning models were developed and evaluated using the Keras (https://keras.io/) and keras-retinanet (https://github.com/fizyr/keras-retinanet) libraries with a Tensorflow backend (https://www.tensorflow.org). Data analysis was performed using the Python language with the numpy, pandas, scipy and sklearn packages. DCMTK (https://dicom.offis.de/dcmtk.php.en) and Pydicom (https://pydicom.github.io/) were used for processing DICOM files.

Reader study. The reader study was performed to directly assess the performance of the proposed deep learning system in comparison with expert radiologists. While a reader study is certainly an artificial setting, such studies avoid the 'gatekeeper bias' inherent in retrospective performance comparison[17], since the ground truth of each case is established a priori in reader studies. Recent evidence also suggests that the rate of positive enrichment itself in reader studies may have little effect on reader aggregate ROC performance[43,44].

Reader selection. Five board-certified and Mammography Quality Standards Act (MQSA)-qualified radiologists were recruited as readers for the reader study. All readers were fellowship-trained in breast imaging and had practiced for an average of 5.6 yr post-fellowship (range 2–12 yr). The readers read an average of 6,969 mammograms over the year preceding the reader study (range 2,921–9,260), 60% of which were DM and 40% of which were DBT.

Study design. The data for the reader study came from Site D as described above. The study was conducted in two sessions. During the first session, radiologists read the 131 index cancer exams and 76 of the negative exams. During the second session, radiologists read the 120 pre-index exams and the remaining 78 negative exams. There was a washout period of at least 4 weeks in between the 2 sessions for each reader. The readers were instructed to give a forced BI-RADS score for each case (1–5). BI-RADS 1 and 2 were considered no recall, and BI-RADS 3, 4 and 5 were considered recall[36]. Radiologists did not have any information about the patients (such as previous medical history, radiology reports or other patient records), and were informed that the study dataset is enriched with

cancer mammograms relative to the standard prevalence observed in screening; however, they were not informed about the proportion of case types. All readers viewed and interpreted the studies on dedicated mammography workstations in an environment similar to their clinical practice. Readers recorded their interpretations in electronic case report forms using SurveyMonkey. In addition to a forced BI-RADS, readers provided breast-density classification and, for each detected lesion, the lesion type, laterality, quadrant and a 0–100 probability of malignancy score (for up to 4 lesions). In the main text, reader binary recall decisions using BI-RADS were used for analysis because this more closely reflects clinical practice. In Extended Data Fig. 1, reader ROC curves using the probability of malignancy score were also computed, which show similar results.

Localization-based analysis. While the reader study results in the main text correspond to case-level classification performance, localization-based analysis was also performed. In the study, readers reported the breast laterality and quadrant for each lesion that was determined to warrant recall. Ground-truth laterality and quadrant for malignant lesions were provided by the clinical lead of the reader study by inspecting the mammogram images along with pathology and radiology reports. For the pre-index cases, the ground-truth location was set to the ground-truth location of the corresponding index case, even if the lesion was deemed not visible in the pre-index exam. The proposed deep learning model provides localization in the form of bounding boxes. To compare to the readers and to also act as an exercise in model output interpretability, a MQSA-qualified radiologist from a different practice than the reader study site mapped the outputted boxes of the model to breast laterality and quadrant. This radiologist was blinded to the ground truth of the cases and was instructed to estimate the location of the centroid for each given bounding box, restricting the estimation to one quadrant.

Localization-based results are contained in Extended Data Fig. 5. Both laterality-based and quadrant-based localization sensitivities are considered, requiring correct localization at the corresponding level in addition to recalling the case. Since the readers reported at most 1 lesion for the vast majority of cases (90%) and to avoid scenarios where predictions involving many locations are rewarded, our primary analysis restricts the predicted locations to the location corresponding to the highest scoring lesion. For the model, this corresponds to taking the highest scoring bounding box in the highest scoring laterality. For the readers, the probability of malignancy score provided for each lesion was used to select the highest scoring location. In cases with more than one malignant lesion, a true positive was assigned if the reader or model location matched the location of any of the malignant lesions. As in Fig. 2a(ii) and Fig. 2b(ii), we compared the sensitivity of each reader to the model by choosing a model score threshold that matches the specificity of the given reader. The model was also compared with the reader average in a similar fashion. We additionally report results where we considered a prediction a true positive if any reported lesion by the reader matches the ground truth location (that is, instead of just the top scoring lesion), while still restricting the model to one box per study.

Statistical analysis. ROC curves were used throughout the text as a main evaluation method. We note that ROC analysis is the standard method for assessing diagnostic performance because it is invariant to the ratio of positive to negative cases, and because it allows the comparison of performance across the entire range of possible recall rates (in other words, operating points). Confidence intervals and s.d. values for AUCs and average readers' sensitivity and specificity were computed via bootstrapping with 10,000 random resamples. The *P* values for comparing the model's sensitivity and specificity with the average reader sensitivity and specificity were computed by taking the proportion of times the difference between the model and readers was less than 0 across the bootstrap resamples. The *P* value for comparing AUCs between two models on the same dataset was computed using the DeLong method[45]. Bootstrapping with 10,000 random resamples was used to compare AUC performance across datasets.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
Applications for access of the OMI-DB can be completed at https://medphys.royalsurrey.nhs.uk/omidb/getting-access/. The DDSM can be accessed at http://www.eng.usf.edu/cvprg/Mammography/Database.html. The remainder

of the datasets used are not currently permitted for public release by their respective Institutional Review Boards.

Code availability
Code to enable model evaluation for research purposes via an evaluation server has been made available at https://github.com/DeepHealthAI/nature_medicine_2020.

References
39. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
40. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, (2016).
41. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. in *The 3rd International Conference on Learning Representations (ICLR)* (2015).
42. Lin, T., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. in *The IEEE International Conference on Computer Vision (ICCV)*, 2999–3007 (2017).
43. Gallas, B. D. et al. Impact of prevalence and case distribution in lab-based diagnostic imaging studies. *J. Med. Imaging* **6**, 1 1 (2019).
44. Evans, K. K., Birdwell, R. L. & Wolfe, J. M. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. *PLoS ONE* **8**, e64366 (2013).
45. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).

Author contributions
W.L., B.H., G.R.V. and A.G.S. conceived of the research design. W.L., B.H., J.G.K., J.L.B., M.W., M.B., G.R.V. and A.G.S. contributed to the acquisition of data. W.L., A.R.D., B.H. and J.G.K. contributed to the processing of data. W.L. developed the deep learning models. W.L., A.R.D., B.H., J.G.K., G.G., E.W., K.W., Y.B., M.B., G.R.V. and A.G.S. contributed to the analysis and interpretation of data. E.W. and J.O.O. developed the research computing infrastructure. W.L., A.R.D., E.W., K.W. and J.O.O. developed the evaluation code repository. W.L., A.R.D., B.H., J.G.K., G.G. and A.G.S. drafted the manuscript.

Competing interests
W.L., A.R.D., B.H., J.G.K., G.G., J.O.O., Y.B. and A.G.S. are employees of RadNet, the parent company of DeepHealth. M.B. serves as a consultant for DeepHealth. Two patent disclosures have been filed related to the study methods under inventor W.L.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-01174-9.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-01174-9.

Correspondence and requests for materials should be addressed to W.L. or A.G.S.

Peer review information Javier Carmona was the primary editor on this article, and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.
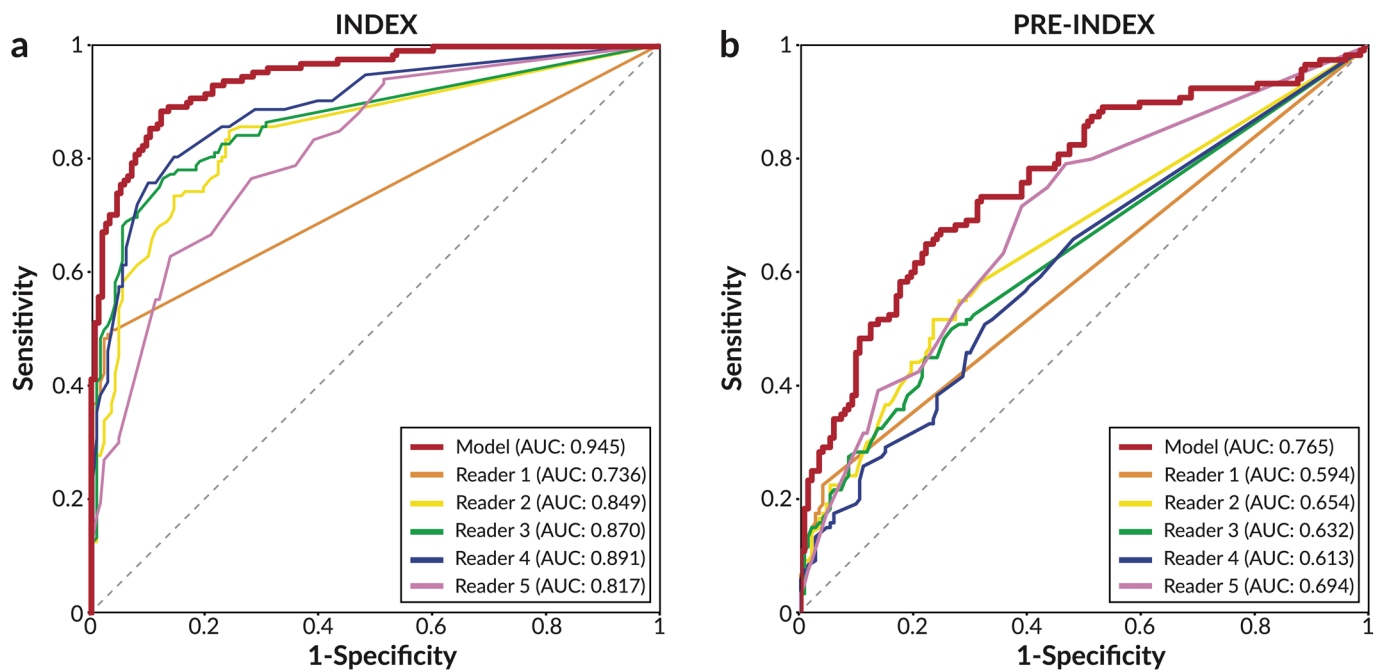
Extended Data Fig. 1 | Reader ROC curves using Probability of Malignancy metric. For each lesion deemed suspicious enough to warrant recall, readers assigned a 0–100 probability of malignancy (POM) score. Cases not recalled were assigned a score of 0. **a**, ROC curve using POM on the 131 index cancer cases and 154 confirmed n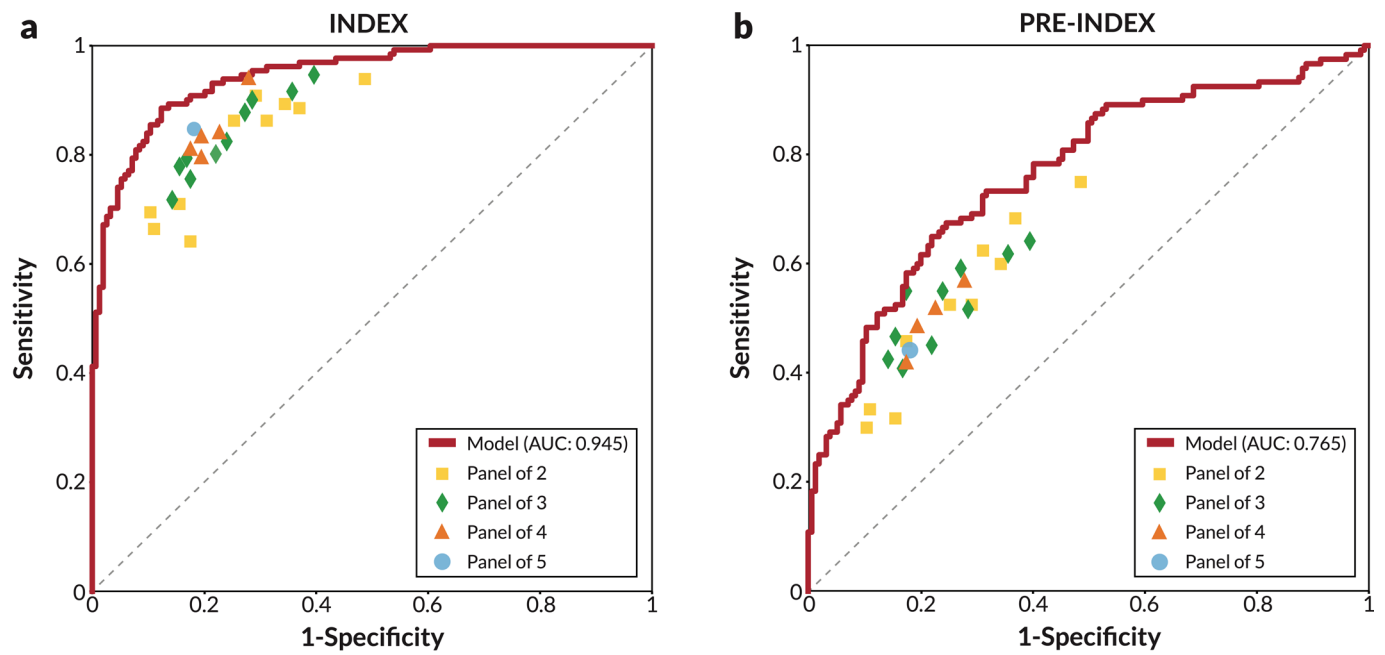egatives. In order of reader number, the reader AUCs are 0.736 ± 0.023, 0.849 ± 0.022, 0.870 ± 0.021, 0.891 ± 0.019, and 0.817 ± 0.025. **b**, ROC curve using POM on the 120 pre-index cancer cases and 154 confirmed negatives. In order of reader number, the reader AUCs are 0.594 ± 0.021, 0.654 ± 0.031, 0.632 ± 0.030, 0.613 ± 0.033, and 0.694 ± 0.031. The standard deviation for each AUC value was calculated via bootstrapping.

Extended Data Fig. 2 | Results of model compared to synthesized panel of readers. Comparison of model ROC curves to every combination of 2, 3, 4 and 5 readers. Readers were combined by averaging BIRADS scores, with sensitivity and specificity calculated using a threshold of 3. On both the **a**, index cancer exams and **b**, pre-index cancer exams, the model outperformed every combination of readers, as indicated by each combination falling below the model's respective ROC curve. The reader study dataset consists of 131 index cancer exams, 120 pre-index cancer exams and 154 confirmed negatives.

a

ROC AUC Comparison: Reader Study Data			
Model	**AUC (95% CI)**	**Delta to Proposed**	**P-value**
Proposed Model	**0.945 (0.919, 0.968)**	–	–
DH-DREAM [32]	0.895 (0.855, 0.931)	0.050	0.0006
Yala et al., 2019 [7]	0.858 (0.812, 0.900)	0.087	1.1E-06
Wu et al., 2019 [11]	0.841 (0.792, 0.887)	0.104	2.0E-06

b

Sensitivity of Models Compared to Readers			
Model	**Sensitivity (95% CI) @ Reader Specificity**	**Delta to Readers**	**P-value**
Proposed Model	**96.2 (91.7, 99.2)**	**14.2 (9.2, 18.5)**	**<0.0001**
DH-DREAM [32]	91.6 (84.3, 95.9)	9.6 (2.2, 14.8)	0.003
Yala et al., 2019 [7]	86.3 (75.7, 92.9)	4.3 (-6.8, 11.7)	0.205
Wu et al., 2019 [11]	84.0 (74.4, 90.2)	2.0 (-7.8, 8.9)	0.372

c

Specificity of Models Compared to Readers			
Model	**Specificity (95% CI) @ Reader Sensitivity**	**Delta to Readers**	**P-value**
Proposed Model	**90.9 (84.9, 96.1)**	**24.0 (17.4, 30.4)**	**<0.0001**
DH-DREAM [32]	85.7 (73.4, 92.9)	18.8 (6.4, 26.8)	0.003
Yala et al., 2019 [7]	71.4 (59.6, 84.5)	4.5 (-5.7, 13.0)	0.244
Wu et al., 2019 [11]	68.2 (48.5, 84.0)	1.3 (-18.5, 17.5)	0.369

d

ROC AUC Comparison: Site A – DM Dataset			
Model	**AUC (95% CI)**	**Delta to Proposed**	**P-value**
Proposed Model	**0.928 (0.912, 0.943)**	–	–
DH-DREAM [32]	0.904 (0.882, 0.924)	0.024	0.0047
Yala et al., 2019 [7]	0.909 (0.889, 0.929)	0.019	0.038
Wu et al., 2019 [11]	0.688 (0.656, 0.721)	0.240	1.7E-48

Extended Data Fig. 3 | Comparison to recent work – index cancer exams. The performance of the proposed model is compared to other recently published models on the set of index cancer exams and confirmed negatives from our reader study **a-c**, and the 'Site A – DM dataset' **d**. *P*-values for AUC differences were calculated using the DeLong method[45] (two sided). Confidence intervals for AUC, sensitivity and specificity were computed via bootstrapping. **a**, ROC AUC comparison: Reader study data (Site D). The Site D dataset contains 131 index cancer exams and 154 confirmed negatives. The DeLong method z-values corresponding to the AUC differences are, from top to bottom, 3.44, 4.87, and 4.76. **b**, Sensitivity of models compared to readers. Sensitivity was obtained at the point on the ROC curve corresponding to the average reader specificity. Delta values show the difference between model sensitivity and average reader sensitivity and the *P*-values correspond to this difference (computed via bootstrapping). **c**, Specificity of models compared to readers. Specificity was obtained at the point on the ROC curve corresponding to the average reader sensitivity. Delta values show the difference between model specificity and average reader specificity and the *P*-values correspond to this difference (computed via bootstrapping). **d**, ROC AUC comparison: Site A – DM dataset. Compared to the original dataset, 60 negatives (0.78% of the negatives) were excluded from the comparison analysis because at least one of the models were unable to successfully process these studies. All positives were successfully processed by all models, resulting in 254 index cancer exams and 7,637 confirmed negatives for comparison. The DeLong method z-values corresponding to the AUC differences are, from top to bottom, 2.83, 2.08, and 14.6.

a

ROC AUC Comparison: Reader Study Data			
Model	AUC (95% CI)	Delta to Proposed	P-value
Proposed Model	**0.765 (0.705, 0.820)**	–	–
DH-DREAM [32]	0.701 (0.639, 0.763)	0.064	0.009
Yala et al., 2019 [7]	0.702 (0.638, 0.762)	0.063	0.008
Wu et al., 2019 [11]	0.701 (0.637, 0.762)	0.064	0.040

b

Sensitivity of Models Compared to Readers			
Model	Sensitivity (95% CI) @ Reader Specificity	Delta to Readers	P-value
Proposed Model	**73.3 (62.0, 81.3)**	**17.5 (6.0, 26.2)**	**0.0009**
DH-DREAM [32]	60.8 (49.2, 76.5)	5.0 (-5.7, 20.5)	0.182
Yala et al., 2019 [7]	59.2 (49.1, 68.9)	3.4 (-7.5, 17.8)	0.204
Wu et al., 2019 [11]	66.7 (51.9, 75.4)	10.9 (-3.8, 19.5)	0.072

c

Specificity of Models Compared to Readers			
Model	Specificity (95% CI) @ Reader Sensitivity	Delta to Readers	P-value
Proposed Model	**83.1 (74.5, 91.0)**	**16.2 (7.3, 24.6)**	**0.0008**
DH-DREAM [32]	72.1 (61.5, 82.5)	5.2 (-5.4, 15.7)	0.179
Yala et al., 2019 [7]	73.4 (56.2, 86.9)	6.5 (-10.5, 20.3)	0.252
Wu et al., 2019 [11]	73.4 (64.9, 85.4)	6.5 (-2.5, 19.4)	0.070

d

ROC AUC Comparison: Site A – DM Dataset			
Model	AUC (95% CI)	Delta to Proposed	P-value
Proposed Model	**0.745 (0.710, 0.778)**	–	–
DH-DREAM [32]	0.683 (0.648, 0.719)	0.062	0.00066
Yala et al., 2019 [7]	0.708 (0.673, 0.742)	0.037	0.014
Wu et al., 2019 [11]	0.607 (0.567, 0.645)	0.138	1.0E-11

Extended Data Fig. 4 | Comparison to recent work – pre-index cancer exams. The performance of the proposed model is compared to other recently published models on the set of pre-index cancer exams and confirmed negatives from our reader study **a-c**, and the 'Site A – DM dataset' **d**. P-values for AUC differences were calculated using the DeLong method[45] (two sided). Confidence intervals for AUC, sensitivity and specificity were computed via bootstrapping. **a**, ROC AUC comparison: Reader study data (Site D). The Site D dataset contains 120 pre-index cancer exams and 154 confirmed negatives. The DeLong method z-values corresponding to the AUC differences are, from top to bottom, 2.60, 2.66, and 2.06. **b**, Sensitivity of models compared to readers. Sensitivity was obtained at the point on the ROC curve corresponding to the average reader specificity. Delta values show the difference between model sensitivity and average reader sensitivity and the P-values correspond to this difference (computed via bootstrapping). **c**, Specificity of models compared to readers. Specificity was obtained at the point on the ROC curve corresponding to the average reader sensitivity. Delta values show the difference between model specificity and average reader specificity and the P-values correspond to this difference (computed via bootstrapping). **d**, ROC AUC comparison: Site A – DM dataset. Compared to the original dataset, 60 negatives (0.78% of the negatives) were excluded from the comparison analysis because at least one of the models were unable to successfully process these studies. All positives were successfully processed by all models, resulting in 217 pre-index cancer exams and 7,637 confirmed negatives for comparison. The DeLong method z-values corresponding to the AUC differences are, from top to bottom, 3.41, 2.47, and 6.81.

a

	Sensitivity by Localization-Level – Index Cases								
	Case-Level Sensitivity			Laterality-Level Sensitivity			Quadrant-Level Sensitivity		
Reader #	Reader	Model	Difference	Reader	Model	Difference	Reader	Model	Difference
1	49.6	70.2	20.6 ± 10.1	42.0	67.2	25.2 ± 8.6	35.9	52.7	16.8 ± 7.1
2	85.5	96.2	10.7 ± 3.7	78.6	87.8	9.2 ± 4.3	64.1	67.2	3.1 ± 4.5
3	86.3	95.4	9.2 ± 3.3	80.9	87.8	6.9 ± 3.6	63.4	67.2	3.8 ± 4.0
4	94.7	97.7	3.1 ± 2.2	84.0	88.6	4.6 ± 3.0	64.1	67.2	3.1 ± 4.1
5	93.9	97.7	3.8 ± 2.5	86.3	88.6	2.3 ± 3.5	70.2	67.2	-3.1 ± 4.3
Average	**82.0**	**96.2**	**14.2 ± 2.4***	**74.4**	**87.8**	**13.4 ± 2.8***	**59.5**	**67.2**	**7.6 ± 3.3***

b

	Sensitivity by Localization-Level – Pre-Index Cases								
	Case-Level Sensitivity			Laterality-Level Sensitivity			Quadrant-Level Sensitivity		
Reader #	Reader	Model	Difference	Reader	Model	Difference	Reader	Model	Difference
1	22.5	29.2	6.7 ± 6.9	17.5	25.0	7.5 ± 5.2	16.7	20.8	4.2 ± 4.8
2	58.3	73.3	15.0 ± 6.7	45.8	51.7	5.8 ± 5.3	32.5	36.7	4.2 ± 4.4
3	52.5	69.2	16.7 ± 6.3	40.0	50.0	10.0 ± 5.0	36.7	36.7	0.0 ± 4.9
4	65.8	82.5	16.7 ± 6.9	46.7	55.8	9.2 ± 5.6	34.2	39.2	5.0 ± 5.0
5	80.0	86.7	6.7 ± 6.2	56.7	58.3	1.7 ± 6.1	42.5	40.0	-2.5 ± 5.1
Average	**55.8**	**73.3**	**17.5 ± 5.1***	**41.3**	**51.7**	**10.3 ± 4.2***	**32.5**	**36.7**	**4.2 ± 3.9**

Extended Data Fig. 5 | Localization-based sensitivity analysis. In the main text, case-level results are reported. Here, we additionally consider lesion localization when computing sensitivity for the reader study. Localization-based sensitivity is computed at two levels – laterality and quadrant (see Methods). As in Fig. 2 in the main text, we report the model's sensitivity at each reader's specificity (96.1, 68.2, 69.5, 51.9, and 48.7 for Readers 1–5 respectively) and at the reader average specificity (66.9). **a**, Localization-based sensitivity for the index cases (131 cases). **b**, Localization-based sensitivity for the pre-index cases (120 cases). For reference, the case-level sensitivities are also provided. We find that the model outperforms the reader average for both localization levels and for both index and pre-index cases (*$P < 0.05$; Specific P-values: index – laterality: $P < 1e − 4$, index – quadrant: $P = 0.01$, pre-index – laterality: $P = 0.01$, pre-index – quadrant: $P = 0.14$). The results in the tables below correspond to restricting localization to the top scoring predicted lesion for both reader and model (see Methods). If we allow localization by any predicted lesion for readers while still restricting the model to only one predicted bounding box, the difference between the model and reader average performance is as follows (positive values indicate higher performance by model): index – laterality: $11.2 ± 2.8$ ($P = 0.0001$), index – quadrant: $4.7 ± 3.3$ ($P = 0.08$), pre-index – laterality: $7.8 ± 4.2$ ($P = 0.04$), pre-index – quadrant: $2.3 ± 3.9$ ($P = 0.28$). P-values and standard deviations were computed via bootstrapping. Finally, we note that while the localization-based sensitivities of the model may seem relatively low on the pre-index cases, the model is evaluated in a strict scenario of only allowing one box per study and crucially, all of the pre-index effectively represent 'misses' in the clinic. Even when set to a specificity of 90%[36], the model still detects a meaningful number of the missed cancers while requiring localization, with a sensitivity of 37% and 28% for laterality and quadrant localization, respectively.

a

Sensitivity by Cancer Characteristics

		Count	Sensitivity			Model AUC
			Model	Reader	Difference (95% CI)	
Cancer Type	ILC or IDC	88	95.5	82.7	12.7 (8.5, 16.5)	0.944 ± 0.014
	DCIS	38	97.4	81.1	16.3 (10.9, 22.2)	0.947 ± 0.020
	Other	5	100	76.0	24.0 (8.0, 42.1)	0.969 ± NA
Cancer Size	0-1cm	45	95.6	83.6	12.0 (5.4, 17.3)	0.927 ± 0.021
	1-2cm	27	96.3	80.7	15.6 (9.7, 22.0)	0.966 ± 0.018
	2-5cm	11	100	81.8	18.2 (8.3, 28.8)	0.984 ± 0.012
	>5cm	3	100	93.3	6.7 (0.0, 22.2)	0.955 ± NA
	Unknown	45	95.6	80.4	15.1 (9.7, 21.0)	0.941 ± 0.02
Lesion Type	Soft Tissue	87	95.4	83.0	12.4 (8.1, 16.2)	0.942 ± 0.015
	Calcifications	53	98.1	81.9	16.2 (11.7, 20.9)	0.957 ± 0.015

b

Sensitivity by Breast Density

	Density	# Cancers	Sensitivity			Model AUC
			Model	Reader	Difference (95% CI)	
Index	Non-Dense (A & B)	81	97.5	82.2	15.3 (11.6, 19.5)	0.959 ± 0.013
	Dense (C & D)	50	90.0	81.6	8.4 (-1.1, 14.5)	0.918 ± 0.026
Pre-Index	Non-Dense (A & B)	74	77.0	58.1	18.9 (9.9, 30.5)	0.756 ± 0.038
	Dense (C & D)	46	69.6	52.2	17.4 (9.6, 27.6)	0.791 ± 0.046

c

Specificity by Breast Density

	Density	# Negatives	Specificity		
			Model	Reader	Difference (95% CI)
Index	Non-Dense (A & B)	96	91.7	62.3	29.4 (24.5, 35.8)
	Dense (C & D)	58	87.9	74.5	13.5 (-1.3, 19.6)
Pre-Index	Non-Dense (A & B)	96	80.2	62.3	17.9 (11.6, 23.7)
	Dense (C & D)	58	93.1	74.5	18.6 (9.5, 23.5)

Extended Data Fig. 6 | Reader study case characteristics and performance breakdown. The performance of the proposed deep learning model compared to the reader average grouped by various case characteristics is shown. For sensitivity calculations, the score threshold for the model is chosen to match the reader average specificity. For specificity calculations, the score threshold for the model is chosen to match the reader average sensitivity. **a**, Sensitivity and model AUC grouped by cancer characteristics, including cancer type, cancer size and lesion type. The cases correspond to the index exams since the status of these features are unknown at the time of the pre-index exams. Lesion types are grouped by soft tissue lesions (masses, asymmetries and architectural distortions) and calcifications. Malignancies containing lesions of both types are included in both categories (9 total cases). 'NA' entries for model AUC standard deviation indicate that there were too few positive samples for bootstrap estimates. The 154 confirmed negatives in the reader study dataset were used for each AUC calculation. **b**, Sensitivity and model AUC by breast density. The breast density is obtained from the original radiology report for each case. **c**, Specificity by breast density. Confidence intervals and standard deviations were computed via bootstrapping.
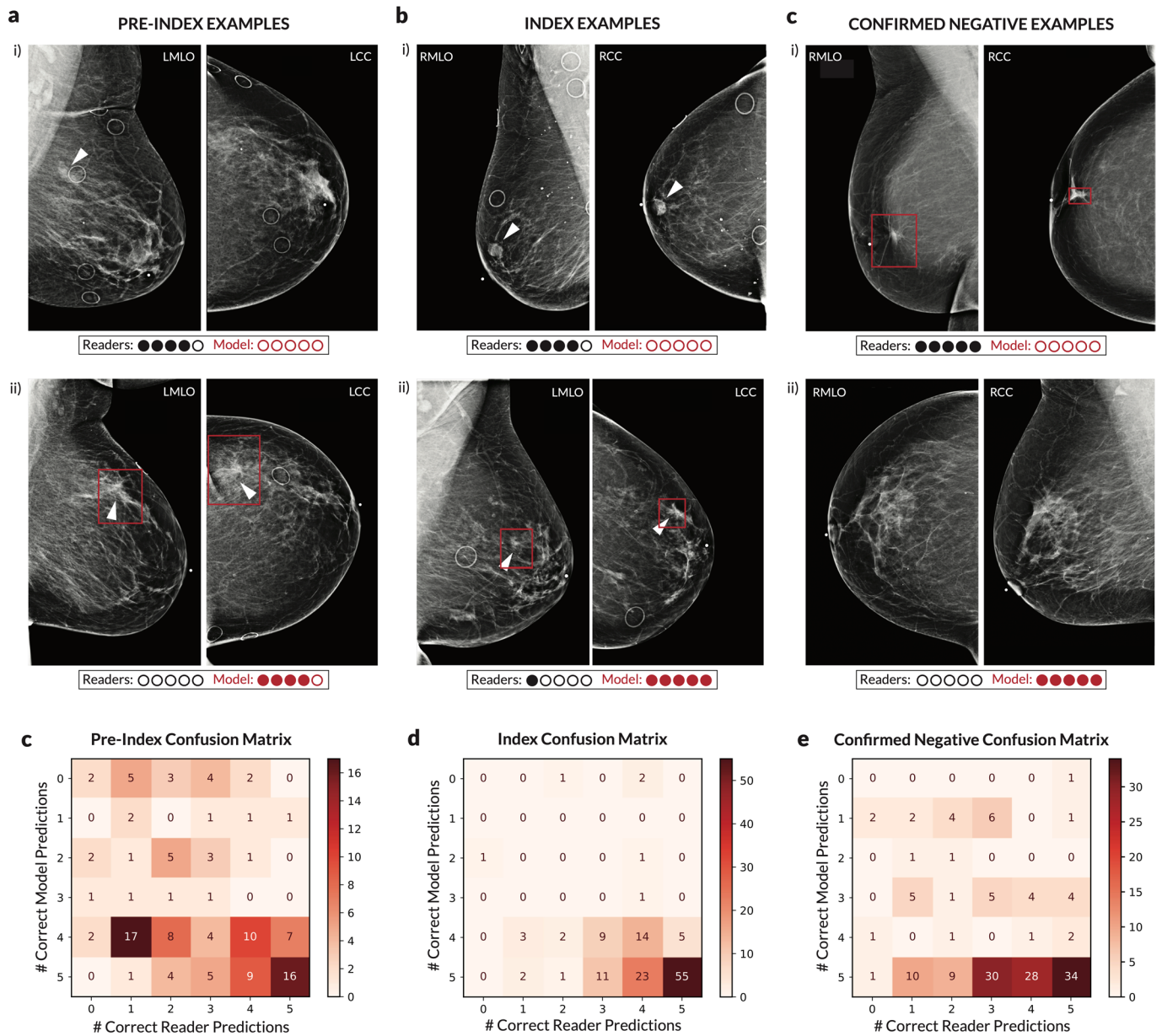
Extended Data Fig. 7 | Discrepancies between readers and the deep learning model. For each case, the number of readers that correctly classified the case was calculated along with the number of times the deep learning model would classify the case correctly when setting a score threshold to correspond to either the specificity of each reader (for index and pre-index cases) or the sensitivity of each reader (for confirmed negative cases). Thus, for each case, 0–5 readers could be correct, and the model could achieve 0–5 correct predictions. The evaluation of the model at each of the operating points dictated by each reader was done to ensure a fair, controlled comparison (that is, when analyzing sensitivity, specificity is controlled and vice versa). We note that in practice a different operating point may be used. The examples shown illustrate discrepancies between model and human performance, with the row of dots below each case illustrating the number of correct predictions. Red boxes on the images indicate the model's bounding box output. White arrows indicate the location of a malignant lesion. **a**, Examples of pre-index cases where the readers outperformed the model (i) and where the model outperformed the readers (ii). **b**, Examples of index cases where the readers outperformed the model (i) and where the model outperformed the readers (ii). **c**, Examples of confirmed negative cases where the readers outperformed the model (i) and where the model outperformed the readers (ii). For the example in **c**.i.), the patient previously had surgery six years ago for breast cancer at the location indicated by the model, but the displayed exam and the subsequent exam the following year were interpreted as BIRADS 2. For the example in **c**.ii.), there are posterior calcifications that had previously been biopsied with benign results, and all subsequent exams (including the one displayed) were interpreted as BIRADS 2. **d**, Full confusion matrix between the model and readers for pre-index cases. **e**, Full confusion matrix between the model and readers for index cases. **f**, Full confusion matrix between the model and readers for confirmed negative cases.

a

| AUC by Cancer Time Window – No Benigns			
Dataset	**0-3 Months**	**0-12 Months**	**12-24 Months**
OMI-DB – UK	0.963 ± 0.003	0.959 ± 0.003	0.744 ± 0.029
Site E – China (Original)	0.971 ± 0.005	–	–
Site E – China (Resampled)	0.956 ± 0.020	–	–
Site A – Oregon (DM)	0.927 ± 0.008	0.900 ± 0.009	0.745 ± 0.017
Site A – Oregon (DBT)			
Slices, no fine-tuning	0.865 ± 0.020	0.859 ± 0.021	–
Manufacturer synthetics, no fine-tuning	0.893 ± 0.022	0.885 ± 0.022	–
Manufacturer synthetics, with fine-tuning	0.922 ± 0.016	0.917 ± 0.016	–
Slices, final model	0.947 ± 0.012	0.936 ± 0.014	–
Slices & man. synthetics, final models	0.957 ± 0.010	0.945 ± 0.013	–

b

AUC by Cancer Time Window – With Benigns			
Dataset	**0-3 Months**	**0-12 Months**	**12-24 Months**
OMI-DB – UK	0.951 ± 0.004	0.947 ± 0.004	0.712 ± 0.029
Site E – China (Original)	0.966 ± 0.006	–	–
Site E – China (Resampled)	0.949 ± 0.020	–	–
Site A – Oregon (DM)	–	–	–
Site A – Oregon (DBT)			
Slices, no fine-tuning	0.863 ± 0.020	0.856 ± 0.021	–
Manufacturer synthetics, no fine-tuning	0.887 ± 0.023	0.879 ± 0.023	–
Manufacturer synthetics, with fine-tuning	0.915 ± 0.017	0.910 ± 0.017	–
Slices, final model	0.941 ± 0.012	0.930 ± 0.014	–
Slices & man. synthetics, final models	0.950 ± 0.010	0.938 ± 0.013	–

Extended Data Fig. 8 | Performance of the proposed models under different case compositions. Unless otherwise noted, in the main text we chose case compositions and definitions to match those of the reader study, specifically index cancer exams were mammograms acquired within 3 months preceding a cancer diagnosis and non-cancers were negative mammograms (BIRADS 1 or 2) that were 'confirmed' by a subsequent negative screen. Here, we additionally consider **a**, a 12-month definition of index cancers, meaning mammograms acquired within 0–12 months preceding a cancer diagnosis, as well as **b**, including biopsy-proven benign cases as non-cancers. The 3-month time window for cancer diagnosis includes 1,205, 533, 254 and 78 cancer cases for OMI-DB, Site E, Site A – DM, and Site A – DBT, respectively. The number of additional cancer cases included in the 12-month time window is 38, 46 and 7 for OMI-DB, Site A – DM, and Site A – DBT, respectively. A 12–24 month time window results in 68 cancer cases for OMI-DB and 217 cancer cases for Site A – DM. When including benign cases, those in which the patient was recalled and ultimately biopsied with benign results, we use a 10:1 negative to benign ratio to correspond with a typical recall rate in the United States.[36] For a given dataset, the negative cases are shared amongst all cancer time window calculations, with 1,538, 1,000, 7,697 and 518 negative cases for OMI-DB, Site E, Site A – DM, and Site A – DBT, respectively. For all datasets except Site E, the calculations below involve confirmed negatives. Dashes indicate calculations that are not possible given the data and information available for each site. The standard deviation for each AUC value was calculated via bootstrapping.

Aggregate Testing Summary										
Dataset	**Location**	**Modality**	**Manufacturer**	**Screening/ Diagnostic**	**# Negative Patients**	**# Confirmed Negatives**	**# Cancer Patients**	**Cancer %**	**AUC - All Negatives**	**AUC – Confirmed Negatives**
OMI-DB	UK	DM	Hologic	Screening	1,967	1,538	1,205	38.0	0.961 ± 0.003	0.963 ± 0.003
Site A – DM	Oregon	DM	GE	Screening	16,369	7,697	254	1.6	0.931 ± 0.008	0.927 ± 0.008
Site A - DBT	Oregon	DBT	Hologic	Screening	11,609	518	78	0.67	0.959 ± 0.008	0.957 ± 0.010
Site D	Mass.	DM	Hologic	Screening	154	154	131	46.0	0.945 ± 0.012	0.945 ± 0.012
Site E	China	DM	Hologic	Diagnostic	1,000	-	533	34.8	0.971 ± 0.005	-

Extended Data Fig. 9 | Aggregate summary of testing data and results. Results are calculated using index cancer exams and both confirmed negatives and all negatives (confirmed and unconfirmed) separately. While requiring negative confirmation excludes some data, similar levels of performance are observed across both confirmation statuses in each dataset. Across datasets, performance is also relatively consistent, though there is some variation as might be expected given different screening paradigms and population characteristics. Further understanding of performance characteristics across these populations and other large-scale cohorts will be important future work. The standard deviation for each AUC value was calculated via bootstrapping.
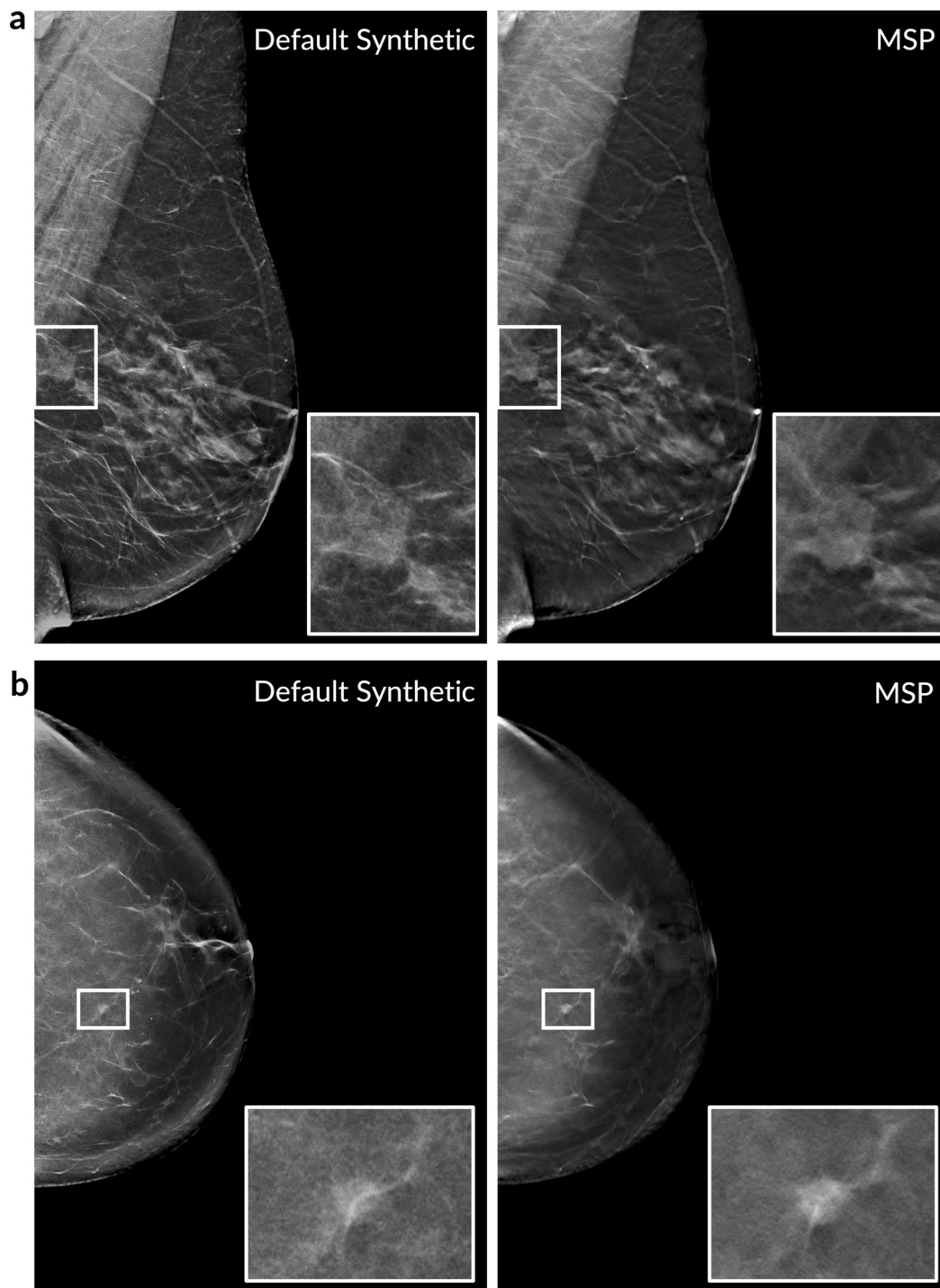
Extended Data Fig. 10 | Examples of maximum suspicion projection (MSP) images. Two cancer cases are presented. Left column: Default 2D synthetic images. Right column: MSP images. The insets highlight the malignant lesion. In both cases, the deep learning algorithm scored the MSP image higher for the likelihood of cancer (**a**: 0.77 vs. 0.14, **b**: 0.87 vs. 0.31). We note that the deep learning algorithm correctly localized the lesion in both of the MSP images as well.

nature research

Corresponding author(s): William Lotter

Last updated by author(s): Oct 25, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The open source libraries DCMTK (https://dicom.offis.de/dcmtk.php.en, version 3.6.1) and Pydicom (https://pydicom.github.io/, version v1.3.0) were used for processing DICOM files. Custom Python (version 3.7.4) scripts were developed for data retrieval and de-identification.
Data analysis	Tensorflow (https://www.tensorflow.org/, version 1.14.0), Keras (https://keras.io/, version 2.1.4), and keras-retinanet (https://github.com/fizyr/keras-retinanet, version 0.3.1) were used for model training and testing. Data analysis was performed using Python (version 3.7.4) with the numpy (version 1.17.3), pandas (version 0.25.2) , scipy (version 1.2.1), and sklearn (version 0.21.3) packages. Code to enable model evaluation for research purposes via an evaluation server is made available at https://github.com/DeepHealthAI/nature_medicine_2020.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Applications for access of the OPTIMAM Mammography Imaging Database (OMI-DB) can be completed at https://medphys.royalsurrey.nhs.uk/omidb/getting-access/. The Digital Database for Screening Mammography (DDSM) can be accessed at http://www.eng.usf.edu/cvprg/Mammography/Database.html. The remainder of the datasets used are not currently permitted for public release by their respective Institutional Review Boards.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The number of mammograms used in the reader study was selected based on time and budgetary constraints. The study involved 154 negative exams, 131 index cancer exams, and 120 pre-index cancer exams. These sample sizes are similar to those used in other mammography reader studies (e.g., FDA approvals) where the statistical significance of new technologies has been tested. Exams were selected for the study such that there was no overlap in patients in the negative and index/pre-index populations. Due to exclusion criteria (see below), there were 13 index cancer exams without a corresponding pre-index cancer exam, and 2 pre-index cancer exams without a corresponding index cancer exam. For the standalone model testing, the percentage of patients that was reserved for testing for each dataset is as follows: OMI-DB - 20%, Site A (Oregon) DM - 40%, Site A (Oregon) DBT - 100%, Site E (China) - 100%. Cross validation splits were created at the patient level. For Site A - DBT and Site E, 100% of the data was reserved for testing to maximally test generalization. The test percentages for OMI-DB and Site A - DM were chosen to provide sufficient positives to ensure reasonable standard deviations of AUC performance. The number of positive/negative exams in the test set for each dataset is as follows: OMI-DB - 1205/1967, Site A (Oregon) DM - 254/16369, Site A (Oregon) DBT - 78/11609, Site D (China) - 533/1000.
Data exclusions	From a total of 462 mammograms that were originally collected for the reader study, 36 exams were excluded from analysis because they were either diagnostic exams, as indicated by containing magnification views, or did not contain all four standard screening views, which could bias the readers. Additionally, 21 studies were excluded due to PACS availability errors. The final analysis dataset thus consisted of 405 studies from 285 women. The standalone testing performances on Site A (Oregon, US) and OMI-DB (UK) "were calculated using screening mammograms. Mammograms obtained in Site A were classified as screening based on the "Study Description DICOM tag. For all training and testing, only craniocaudal (CC) and mediolateral oblique (MLO) mammographic views were utilized. Magnification views, indicated by having a value greater than 1.073 in the "Estimated Radiographic Magnification Factor" DICOM tag, were additionally excluded. Mammograms from patients with breast implants, indicated by the corresponding DICOM tag, were also excluded. All exclusion criteria were established prior to model training and analysis.
Replication	All attempts at replication were successful. In the reader study, the differential in performance between the deep learning system and the human readers was consistent in both the index and pre-index cancer exams. The standalone performance of the deep learning system in the reader study was also consistent with the standalone performances in the additional tested datasets. In particular, the performance of the deep learning system was consistent across populations (from the US, UK, and China), equipment manufacturers (GE and Hologic), and modalities (DM and DBT). Furthermore, the deep learning system exhibited high performance on the reader study dataset and the Chinese dataset despite never being trained on data from these sources. Thus, even though all of the experimental evaluations were performed once and independently for each condition, performance was consistent across the different assessments.
Randomization	For datasets that were used for both training and testing, patients were randomly assigned into training, model selection (validation), and testing splits. All mammograms for a given patient were thus assigned to the same split.
Blinding	The testing data partitions were never used for model development. In particular, the entire datasets from Site A (Oregon) DBT, Site D (Massachusetts), and Site E (China) were excluded from model development. Model selection was performed based on AUC performance on the held-out model selection (validation) data partition. Given the geographic proximity between Site C and Site A, we exclude 14 studies from the test set of Site A where the patient also appears in Site C. Overlap between these two sets was assessed using medical record number and a fuzzy matching algorithm based on patient name and date of birth. The reader study involved selecting cases based on known ground truth status, but besides case selection, the investigators (and readers) involved in the study were blinded to group allocation during data collection of the study. The data collection was additionally performed by a separate set of investigators from those involved in the deep learning model development. For the deep learning model evaluation, the investigators involved in developing the deep learning model were blinded to group allocation during model evaluation. Specifically, the model was evaluated to produce a set of scores, which was then compared to the ground truth group allocation by another investigator who was not involved in the deep learning model development or evaluation. Similarly, for the Site E (China) dataset, the data collection and model assessment were performed by a different set of investigators than those involved in the development of the deep learning model. Given that the investigators involved in the deep learning model development had historically been involved with the curation of the Site A (Oregon) and OMI-DB datasets, such complete separation was not possible for these datasets, but the test sets nonetheless involved data that had never been used for model development.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒ ☐	Antibodies
☒ ☐	Eukaryotic cell lines
☒ ☐	Palaeontology
☒ ☐	Animals and other organisms
☐ ☒	Human research participants
☒ ☐	Clinical data

Methods

n/a	Involved in the study
☒ ☐	ChIP-seq
☒ ☐	Flow cytometry
☒ ☐	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

The results of this paper are calculated using screening mammograms from the reader dataset (Site D; Massachusetts, US), screening mammograms from Site A (Oregon, US), screening mammograms from the OMI-DB dataset (UK), and diagnostic mammograms from Site E (Henan Province, China). The training data included screening and diagnostic mammograms from Site A (Oregon, US), Site B (Rhode Island, US), Site C (Oregon, US), OMI-DB (UK), and the Digital Database for Screening Mammography (DDSM; US). All mammograms used in the study were from female patients.

In the United States, screening mammography is recommended annually or biannually starting at age 40 or 50 depending on specific guidelines. For example, the American College of Radiology (ACR) and Society of Breast Imaging (SBI) recommend general screening start at age 40, the American Cancer Society (ACS) at age 40-45, and the US Preventive Services Task Force (USPSTF) at age 50. The Site A dataset was collected from several clinics in the Medical Radiology Group located in Medford, OR between 2010-2017. The Site B dataset was collected from Rhode Island Medical Imaging, an inpatient medical center with affiliated mammography centers located in Rhode Island, between 2016-2017. The Site C dataset was collected from multiple imaging facilities in the Providence health system located in Oregon between 2017-2018. The Site D dataset was collected from multiple imaging facilities within the University of Massachusetts Memorial Medical system located in Massachusetts between 2011-2014.

In the United Kingdom, the National Health Service Breast Screening Programme (NHSBSP) provides screening to women age 50-70 who are registered with a general practitioner (GP). Screening is recommended every 3 years. The OMI-DB dataset contains data collected in several locations in the UK.

In China, screening mammography is much less prevalent than the US or UK. The Site E dataset was collected from Henan Provincial People's Hospital in Zhengzhou, China between 2012-2017. Due to the limited screening mammography in China, the mammograms collected at this site were diagnostic mammograms. The dataset generally had larger lesions (64% larger than 2 cm) than would be expected in a screening population due to the diagnostic nature. To approximately match the distribution of tumor sizes found in a US population, a bootstrap method was used to sample the data towards a target distribution with 49% of tumors between 0 cm to 1 cm, 25% between 1 cm to 1.5 cm, 11% between 1.5 cm to 2 cm, and 15% greater than 2 cm. These binning proportions were chosen to approximately align with US statistics as estimated by the National Radiology Data Registry (https://nrdr.acr.org/Portal/NMD/Main/page.aspx).

Pathology confirmation of malignancy was used as the gold standard for cancer status throughout the study. Cancer subtype information was also collected for the reader study dataset (Site D). Out of the 131 cancers in this dataset, 88 were invasive (invasive lobular carcinoma (ILC) or invasive ductal carcinoma (IDC)), 38 were non-invasive (ductal carcinoma in situ (DCIS)), and 5 had an "other" classification (e.g., adenosquamous carcinoma). Forty five of the cancers had a size less than 1 cm, 27 had a size between 1-2 cm, 11 were between 2-5 cm, 3 cancers were larger than 5 cm, and the cancer size was unknown for 45 cancers. A soft tissue lesion was present in 87 of the cancers and calcifications were present in 53 of the cancers. The cancer patients had a mean age of 64.1 years with a standard deviation of 9.8, minimum of 43, and maximum of 88. The non-cancer patients in the study had a mean age of 63.4 years with a standard deviation of 9.9 years, minimum of 43, and maximum of 89.

Recruitment

All patient data was collected retrospectively from clinical centers in the United States, United Kingdom, and China. The majority of data from the US and UK sites correspond to screening mammograms. Self-selection biases corresponding with the choice to participate in screening may be present, but are likely to be consistent with the real-world patient population.

In the United States, screening mammography is recommended annually or biannually starting at age 40 or 50 depending on specific guidelines. As examples, the ACR and SBI recommend general screening start at age 40, the ACS at age 40-45 and the USPSTF at age 50. The Affordable Care Act requires that most insurance plans pay for 100% of screening mammography costs for women ages 40 and older. Patients do not need a physician referral to receive a screening mammogram. Mammograms in the US must be performed at a Mammography Quality Standards Act (MQSA) certified facility and must be read by an MQSA-qualified radiologist as regulated by the US Food and Drug Administration (FDA). Mammograms are typically only read by one radiologist in the US.

The Site A dataset was collected from several clinics served by the Medford Radiology Group located in Medford, OR between 2010-2017. From 2010-2015, full-field digital mammography was used for screening, with a switch to digital breast tomosynthesis in 2016. For 2010-2015, a list of all mammograms acquired in this time period was obtained by querying the PACS system, and then all of the mammograms in this list were subsequently retrieved. For 2016-2017, a list of all screening mammograms acquired over this time period was provided by the clinic, which were then subsequently retrieved.

The Site B dataset was collected from Rhode Island Medical Imaging, an inpatient medical center with affiliated mammography

centers located in Rhode Island, between 2016-2017. The collected set of mammograms (almost entirely DBT) corresponded to a list provided by the clinic consisting of all screening and diagnostic mammograms acquired over this time period.

The Site C dataset was collected from multiple imaging facilities in the Providence health system located in Oregon between 2017-2018. All biopsy-confirmed cancer cases over this period were collected, along with a random sample (approximately 20%) of all non-cancer mammograms over this period, where the list of all exams was obtained by querying the PACS system. The data from Site C consists of almost entirely DBT.

The reader study (Site D) data was collected from multiple imaging facilities within the University of Massachusetts Memorial Medical system from 2011-2014. The positive exams in the study were obtained using a list of screen-caught cancer patients with biopsy proven pathology, who additionally had received a screening mammogram interpreted as BIRADS 1 or 2 within the 12-24 months preceding the cancer-detected mammogram. Negative exams, consisting of a BIRADS 1 or 2 screening mammogram followed by a subsequent BIRADS 1 or 2 screening mammogram, from a different set of patients were chosen to have a similar distribution in patient age and breast density compared to the cancer cases.

Five readers who were fellowship trained in breast imaging and MQSA-qualified were recruited for the study. The readers had an average of 5.6 years in practice post-fellowship with a range of 2 to 12 years. The readers read an average of 6,969 mammograms in the past year with a range of 2,921 to 9,260 exams per year.

In the United Kingdom, the National Health Service Breast Screening Programme (NHSBSP) provides screening to women 50 to 70 years of age who are registered with a general practitioner. Women who do not meet these criteria can self-refer to a screening program. Screening is recommended every 3 years in the UK. The OMI-DB dataset contains data exclusively from the UK.

The Site E dataset was collected from Henan Provincial People's Hospital in Zhengzhou, China between 2012-2017. Due to the limited screening mammography in China, the mammograms collected at this site were diagnostic mammograms.

Ethics oversight

All non-public datasets (data from Sites A, B, C, D, and E) were collected under IRB approved protocols. The following review boards were used for each dataset: Site A: Southern Oregon IRB, Site B: Rhode Island Hospital IRB, Site C: Providence IRB, Site D: Advarra IRB, and Site E: Henan Provincial People's Hospital IRB. All data were de-identified prior to model training and testing.

Note that full information on the approval of the study protocol must also be provided in the manuscript.