Radiology: Artificial Intelligence

High-performance Open-source AI for Breast Cancer Detection and Localization in MRI

Lukas Hirsch, PhD[1] (https://orcid.org/0000-0003-3173-9142)

Elizabeth J. Sutton, MD[2] (http://orcid.org/0000-0002-1407-7012)

Yu Huang, PhD[1] (http://orcid.org/0000-0003-4178-0739)

Beliz Kayis, BS[1] (http://orcid.org/0009-0001-8387-877X)

Mary Hughes, MD[2] (http://orcid.org/0000-0002-4168-677X)

Danny Martinez, MS[2] (http://orcid.org/0000-0002-8564-9049)

Hernan A. Makse, PhD[3]

Lucas C. Parra, PhD[1] (http://orcid.org/0000-0003-4667-816X)

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

https://doi.org/10.1148/ryai.240550

Purpose: To develop and evaluate an open-source deep learning model for detection and localization of breast cancer on MRI.

Materials and Methods: In this retrospective study, a deep learning model for breast cancer detection and localization was trained on the largest breast MRI dataset to date. Data included all breast MRIs conducted at a tertiary cancer center in the United States between 2002 and 2019. The model was validated on sagittal MRIs from the primary site (n = 6,615 breasts). Generalizability was assessed by evaluating model performance on axial data from the primary site (n = 7,058 breasts) and a second clinical site (n = 1,840 breasts).

Results: The primary site dataset included 30,672 sagittal MRI examinations (52,598 breasts) from 9,986 female patients (mean [SD] age, 53 [11] years). The model achieved an area under the receiver operating characteristic curve (AUC) of 0.95 for detecting cancer in the primary site. At 90% specificity (5717/6353), model sensitivity was 83% (217/262), which was comparable to historical performance data for radiologists. The model generalized well to axial examinations, achieving an AUC of 0.92 on data from the same clinical site and 0.92 on data from a secondary site. The model accurately located the tumor in 88.5% (232/262) of sagittal images, 92.8% (272/293) of axial images from the primary site, and 87.7% (807/920) of secondary site axial images.

Conclusion: The model demonstrated state-of-the-art performance on breast cancer detection. Code and weights are openly available to stimulate further development and validation.

An open-source deep learning model developed and trained on the largest breast MRI dataset to date achieved state-of-the-art in performance in breast cancer detection and localization.

Abbreviations

AI = Artificial Intelligence, AUC = Area Under the Receiver Operating Characteristic Curve, CNN = Convolutional Neural Network, CI = Confidence Interval

Key Points:

- A 2D CNN trained on a uniquely-large breast MRI dataset achieved state-of-the-art performance (AUC = 0.95), which is comparable to radiologist performance.
- The model generalized well across acquisition orientations and clinical sites.
- Open-source code and pretrained model weights are made publicly available.

Breast cancer remains a leading cause of cancer-related deaths among women in the United States (1). Early detection is crucial for successful treatment and improved patient outcomes (2). For women at high risk of developing breast cancer, annual MRI screening is recommended in addition to mammography (3). Breast MRI is also used diagnostically when a tumor is suspected based on clinical findings, mammography, or ultrasound. It is effective for early breast cancer detection (4), including in women with dense breast tissue where mammography may be less reliable. Supplemental breast MRI use is expected to rise following recent recommendations to screen women with extremely dense breasts (5).

Breast MRI interpretation is time-consuming and requires specialized training, as radiologists must review multiple slices in each volume. Automated reading has the potential to assist radiologists by identifying MRI slices most likely to contain a tumor (6) or triaging low-probability images that do not require reading (7). Given that three out of four biopsies are negative in routine clinical practice (8,9), reliable prediction of negative outcomes through automation could also help reduce the biopsy burden.

There has been substantial progress with deep learning models in radiology (10–18), with the ability to detect subtle patterns and abnormalities by analyzing large datasets. This progress is most evident in mammography, where population-wide screening programs have generated large datasets containing hundreds of thousands of images (19–22), enabling the development and validation of accurate models and fueling a growing interest in automation. In contrast, breast MRI screening reaches a smaller population, resulting in more limited datasets. Therefore, model performance in breast MRI has yet to consistently match that of radiologists and often excludes complex cases, such as those involving implants or postsurgical changes (23,24). Moreover, most existing models provide only a probability of malignancy (10) without localizing the area of concern on MRI, limiting their practical utility for radiologists.

Modern deep learning models contain a large number of parameters, making them prone to overfitting when trained on small, single-site datasets, which may limit generalizability to other sites (25–27). This is particularly relevant for MRI, where image acquisition parameters can vary significantly between institutions. Even within a single institution, temporal changes, such as a shift from sagittal to axial-plane acquisition can introduce variability (28). To date, only two published studies have included multisite validation for breast cancer detection in MRI (10,29). Cross-site validation remains challenging, largely because most trained models have not been publicly released, with one recent exception (10). Additionally, performance may be constrained by the relatively small size of available datasets, typically limited to a few hundred examinations (30–33).

This study aimed to develop an open-source deep learning model for MRI-based breast cancer detection trained on a large dataset comprising tens of thousands of examinations from Memorial Sloan-Kettering Cancer Center in New York (MSKCC). The model was validated across different imaging planes and clinical sites, including an external dataset from Duke University (34). The model was designed to both detect and localize the cancer, thus aiding radiologists during

interpretation. By making the model and its parameters openly available, we aim to foster further research and development in this field.

Materials and Methods

Study Sample and Data Partition

The use of this retrospective data were approved by the institutional review board with a waiver of informed consent, and all procedures were HIPAA compliant. Identifiable patient information was removed, and MRIs were saved with anonymized identifiers before analysis.

We used three distinct datasets for training and testing from a primary and secondary site (Fig 1).

Primary site.—

Data included all breast MRIs conducted at a MSKCC between January 2002 and December 2019 in women. The inclusion criterion was a complete sequence of DCE-MRI and available pathology or clinical follow-up of two years. Data included both screening and diagnostic imaging with multiple examinations for each screening patient. Data were excluded for benign breast images from 332 screening patients who eventually developed cancer to avoid potential false negatives. Primary site data were then separated into sagittal and axial examinations (details in Table 1 and below). To assign labels for each breast, bilateral examinations are separated into right and left breasts. MRIs were labeled as "malignant" when there was biopsy-proven cancer and "benign" if there was no cancer diagnosis within two years of clinical follow-up. Two years is the standard follow-up period for treatment studies, e.g (36,37). and more stringent than previous deep-learning studies when labeling healthy breasts (10). Model performance was evaluated on individual breasts as well as examinations.

Primary site.—, sagittal.—-plane examinations.— (training.—, validation, testing)

Sagittal data included 38,005 examinations from 2002 and 2014 (31,564 screening, 6,015 diagnostic, 426 unknown or N/A) from 12,329 patients. Counting each breast individually yielded 65,105 sagittal breast images with 2,690 malignant images. This dataset was drawn from the same patient cohort reported in previous work (35), used to develop a lesion segmenter, and is otherwise independent from this study. This data were randomly divided by patient into training, validation, and test sets (90/10 for training and test, and subsequently 90/10 for training and validation patients). This dataset included radiologist segmentation in 2D for the slice containing the largest (index) cancer for all 2,690 malignant breast images (termed the "index slice"). Radiologists selected a single index-slice per breast showing the largest tumor extent.

Primary site.—, axial.—-plane examinations.— (testing)

To evaluate the model's performance on a different imaging protocol, axial MRIs, excluding patients from the sagittal cohort were used. This dataset comprised 3,873 examinations from 2013 and 2019 (3,069 screening, 720 diagnostic, 84 unknown or N/A) from 3,219 patients and 7,058 breasts. The dataset contained 688 malignant and 6,370 benign breast images. Volumetric segmentation generated by radiologists for the index lesions was available for 293 of the malignant breasts. Cancer localization could only be evaluated on this subset.

Secondary site.—, axial.—-plane examinations.— (testing)

Performance was also evaluated on axial MRIs from a secondary clinical site, using a public dataset released by Duke University (34). This dataset included 922 axial examinations from patients with confirmed breast cancer, excluding two examinations due to issues identifying pre-and postcontrast images (cases 120 and 596). This dataset included pathology information and radiologist annotations on the extent of the index lesion in one breast. This was provided as a 3D bounding box, only for one lesion, even if there was multicentric or bilateral breast cancer present. “Malignant” breast images were those with malignant pathology (n = 948, as some examinations had malignancy in both breasts). Contralateral breast images without malignant pathology were labeled as “Benign” (n = 892).

Model Architecture

We used a conventional 2D convolutional network designed to detect breast cancer in the 3D volume, by assigning a probability of containing cancer to each 2D sagittal slice of a breast (Fig 2). The maximum probability across all slices is used as the prediction for the whole breast. Localization performance is evaluated in terms of the distance of the maximum probability slice from the reference standard provided by radiologists on the location of the cancer. The inputs to the network are three input channels capturing dynamic contrast enhancement (T1w postcontrast, DCE-in, DCE-out) for the corresponding 2D slice. The model consisted of thirteen 2D-convolutional layers (3 × 3 kernel size), each followed by batch normalization, ReLU activation. A Max Pooling layer was added after every two convolutional layers for a total of 5 downsampling steps. This mapped the input (of dimensions 512,5,12,3) to image features (of dimension 16,16,252), which were flattened, processed by one dense layer, and then concatenated with clinical features, ending in two final dense layers and a Softmax activation. In total, this network had 4,126,294 trainable parameters. Code and model weights can be accessed at: https://github.com/lkshrsch/BreastCancerDiagnosisMRI.

Data Pre-processing and Harmonization

Preprocessing followed our previous work in segmentation (35). Briefly, pre-and postcontrast T1-weighted images were coregistered using NiftyReg (38), and dynamic contrast enhancement was summarized into images capturing initial contrast uptake (DC-in) and washout (DC-out), alongside the first postcontrast T1-weighted image (T1-post). These three channels were normalized by dividing by the 95th percentile of the precontrast T1-weighted image in each examination. To adjust for interchannel differences, each channel was divided by its 95th percentile across the training set. The sagittal MRI data had varying in-plane resolutions (0.4 mm to 0.8 mm). Low-resolution images were upsampled by a factor of two for harmonization. Axial images were resampled to match a sagittal in-plane resolution of 0.4 mm and separated into left and right breasts. All images were cropped to 512 × 512 pixels, ensuring the breast was centered. For the axial examinations, the network operated on these resampled sagittal images.

Demographic Data

Demographic information included age and 11 categorical variables with one-hot-encoding: Family history of breast cancer (yes, no), ethnicity (Hispanic or Latino, not Hispanic, unknown); and race (Asian-far east/Indian subcontinental, black or African American, native American/American islander, native Hawaiian or Pacific islander, white, unknown). All information was self-reported, with missing ethnicity and race imputed as "unknown."

Training

The model was trained using index slices from malignant images as positive examples. As negative examples, we selected the center slice and one randomly selected slice from benign images. All models were trained using a focal loss (39) with alpha = 5, using the" Adam" optimizer (40) with learning rate of 1e-5. All models were trained for 100 epochs with early stopping (Fig S1), and the network weights with the lowest validation loss were saved for evaluation. Unless otherwise specified, all models were trained with data augmentation, consisting of random rotation within 60 degrees, random shear of scale 0.1, random horizontal and vertical flips, and random intensity scaling in range 0.8–1.2, all implemented using the TensorFlow preprocessing ImageDataGenerator library.

Validation and Model Selection

Various model architectures and hyper parameters were compared based on their AUC performance on the sagittal validation set (Fig S2). Comparisons included different loss functions (binary-cross entropy vs focal loss; Fig S2A), the effect of data augmentation (Fig S2B), and training data sizes (10%, 50%, 100% of the whole data (Fig S2C). Adding the contralateral breast as input (Fig S2D) and comparing the architecture to ResNet50 (with and without ImageNet pretrained weights) were also tested (Fig S2A). Fine-tuning the pretrained ResNet50 outperformed training it from scratch, but the CNN, trained from scratch with demographic information, achieved the best validation-set performance. Adding the contralateral breast did not significantly improve performance and was excluded from the final model. Data augmentation substantially boosted validation set performance and was included in the final model training.

Categorization of Image Quality in Sagittal Test Set

The sagittal test set was visually inspected (blinded to labels and predictions by LAH, 6 years of experience with breast MRI) to categorize image quality and determine its effect on performance. Categories that could occur simultaneously included biopsy clips, implants, large postsurgery changes, and poor image quality (blurriness, movement or fat-saturation artifacts, enhancing nipple tissue).

Statistical Analysis

Performance was evaluated using AUC. Bootstrapped confidence intervals for AUC were obtained by resampling with replacement subjects and the predicted probabilities 1,000 times. Confidence intervals were then computed from the 2.5 and 97.5 percentiles of the AUC values derived from each data drawn during bootstrapping.

AUC differences between examination categories were assessed using a bootstrap sample. All images were combined, ignoring categories, and randomly drawn with replacements to match benign/malignant numbers in categories. AUC differences from this bootstrap sample were used to compute *P* values in a one-sided test, assuming conservatively that the presence of these clinical and imaging abnormalities worsenes model performance (Fig S3). Wilcoxon signed-rank tests and Pearson correlation tests were performed with the scipy.stats package in Python (v3.1). A *P* value of less than 0.05 was considered statistically significant.

Results

We trained deep networks with various configurations using the sagittal examinations from the primary site. The primary site dataset included 30,672 sagittal MRI examinations (52,598 breasts) from 9,986 female patients (mean [SD] age, 53 [11] years; range 13–93, SD 11, Fig 2 and Table 1 for detailed clinical characteristics). The AUC measured on a validation set showed the benefit of the large data size, 2D data augmentation, and superiority of the CNN over a ResNet50 (Fig S2). Based on this, we selected for final testing a 2D deep CNN (Fig 2) trained with data augmentation, focal loss function, and demographic information.

Detection Performance on Sagittal Images

Evaluation on a random subset of patients not included in the training set (6,615 breast sagittal images, 262 with biopsy-confirmed cancers) demonstrated an AUC of 0.95 (CI: 0.93–0.96) (Fig 3A). Evaluating results by the outcome of the examination instead of each individual breast achieves an AUC of 0.94 (Fig S4A). We also evaluated results using only the first examination for screening patients to rule out repeated measures and obtain the same AUC of 0.94 (Fig S5A). All examinations underwent routine clinical Breast Imaging Reporting & Data System (BI-RADS) assessment by radiologists, and the estimated cancer probability generally increased with the BI-RADS score (Fig S6), supporting internal validity of the model's prediction. A network trained on demographic information alone achieved an AUC of 0.60 (CI:0.56–0.63) (Fig S2A). Of the demographic variables only family history and unknown race and ethnicity were individually associated with outcome (Table S2).

Generalization of Detection Performance to Axial Images

MRI volumes have a higher in-plane resolution. While the training dataset primarily consisted of sagittal images (higher resolution in depth and lateral directions), current clinical practice often uses axial acquisition. To assess the model's generalization to axial images (higher lateral and vertical resolution), we resampled axial images in the sagittal plane to match the resolution of the training data, and processed them with the same trained model. Without fine-tuning, the model achieved an AUC of 0.92 (CI: 0.91–0.93) on the primary site axial data (n = 7,058 breasts, 688 cancers) (Fig 3B) and an AUC of 0.92 (CI: 0.91–0.93) on axial scans from a secondary site (n = 948 malignant, n = 892 benign) (Fig 3C). Because the axial data from the primary site includes screening data, there are multiple samples from the same patient. Removing this correlation and evaluating results only considering the first examination, the model achieves a performance of 0.90 (Fig S4B). Similarly, when only evaluating performance based on examination outcomes instead of individual breasts, the model achieves an AUC of 0.92 (Fig S5B). Data from the secondary site already consist of single patients, and the outcome of all examinations is positive, therefore these evaluations cannot be applied.

Cancer Localization

The model estimates cancer probability for each 2D slice in the MRI volume (Fig 2). The slice with the maximum probability localizes the tumor. To determine the accuracy of this localization for the sagittal data from the primary site, we used the 2D segmentation provided by radiologists for the index lesion, and extended these to 3D using volumetric automatic segmentation (35). The model's maximum probability slice intersected the lesion volume in 88.5% (232/262) of breasts (Fig 4A).

Additionally, the maximum probability slice correlated with the index slice provided by the radiologist (Pearson correlation r = 0.87, $P < .001$, $n = 262$, Fig S7).

For axial data, the maximum probability slice intersected the 3D segmentation provided by radiologists in 92.8% (272/293) of breasts from the primary site (Fig 4C) and the bounding box around the cancers in 87.7% (807/920) of breasts from the secondary site (Fig 4B).

Post Hoc Exploratory Analyses of Test Case Subsets

We conducted exploratory analyses on subsets of the sagittal test data to further characterize the model's performance. To assess its ability to predict biopsy outcomes, we evaluated the performance on the subset that received biopsies (BI-RADS 4&5, $n = 578$, 94 cancers), achieving an AUC of 0.86 (Fig S8). Given that this subset consisted only of suspicious cases requiring biopsies, a lower performance is expected.

To determine the robustness of the model, we analyzed performance on cases typically excluded from previous studies, such as implants and imaging artifacts (10,41). All images were categorized into one of 4 categories (Table 2). For categories with both benign and malignant images, we evaluated AUC and assessed if performance differed significantly between breasts with or without the category (eg, presence or absence of biopsy clips). We found that performance was comparable across all these categories.

Discussion

We demonstrated that training a 2D CNN from scratch with an uniquely large MRI dataset can lead to a new state-of-the-art performance in breast cancer detection (Table 3). Performance benefitted also from efficient implementations of 2D data augmentation methods. Notably, we leveraged information from radiologists regarding the location of the cancer along one dimension (slice number with the index lesion). This information is considerably more informative than a single overall diagnostic label for the entire volume. This allowed us to design a network that highlights the image most likely to contain a tumor. In doing so, the network provides interpretable results that may assist radiologists in their diagnostic workflow.

This work addresses several limitations of previous studies: data size, case exclusions, out-of-plane and cross-site validation, interpretability, comparison with radiologist performance, and public release of the trained model. We will discuss each aspect in turn, while providing an overview of published results on deep learning-based breast cancer detection in MRI (Table 3).

A key distinguishing factor of this study is the size of the dataset used. Recently, curated MRI datasets have grown from a few hundred (30,32,33,42) to thousands of images (10,29,41). Larger datasets offer greater diversity, enabling the training of large models from scratch, as we have shown here (Fig 5).

In contrast, previous studies (23,30,32,33,41,43) with smaller datasets relied on fine-tuning pretrained models from unrelated tasks, like ImageNet or Kinetics-400 (44). However, features extracted by such models may differ significantly from those needed for analyzing 3D MRI images. Some studies have shown that models trained on just over a thousand examinations can outperform those pretrained on ImageNet (23). Others have shown that fine-tuning a pretrained model can improve performance as compared with training from scratch, even when a larger MRI dataset is available (10). Our validation set analysis found that fine-tuning ResNet50 trained on ImageNet numerically outperformed training it from scratch, supporting the view that pretrained models do

help, even if the imaging domains are quite distinct. However, the simpler CNN model, trained from scratch, outperformed a ResNet50, showing that with enough data the model architecture becomes less significant.

Unlike many studies that exclude difficult cases like breast implants, postoperative changes, or imaging artifacts (10,23,24), our training included such examples. We hypothesized that the larger training set would encompass enough of these anomalies to avoid exclusions. Indeed, our overall performance (AUC = 0.95, CI: 0.93–0.96) exceeded the top-performing AI study to date (10) (AUC = 0.92, CI: 0.92–0.93), despite including previously excluded cases. Beyond the obvious benefit of providing detection for all cases, avoiding exclusions is essential for handling tens of thousands of examinations, as manual visual inspection is impractical. Full automation also eliminates the need for manual region selection, as required in previous studies (23,43).

This study trained a model using sagittal MRI scans because this was the largest dataset available to us at this time. Since MRI resolution is higher in-plane, axial test set images were upsampled vertically. Nonetheless, the model performed well on both primary and secondary site axial scans without fine-tuning, demonstrating its ability to generalize beyond the training data. Notably, it performed well in both cancer classification and localization on axial examinations.

Trained on individual slices without global position knowledge, the model accurately selected slices containing index lesions. This is evidenced by the high hit rates and correlation with the index slice from radiologists. While some previous models estimated the location of the tumor (29,30,33,43), many did not (10,23,32,41). Providing such information is crucial for integrating AI into clinical workflows, building confidence in its detection, and potentially guiding abbreviated radiologist re-evaluations.

The network's detection performance appears to be comparable to historical data on radiologist performance. Among the four studies with such measures, Witowski et al (2022) (10) reported an average AUC of 0.89 (CI: 0.85–0.95) for five readers on 100 examinations, numerically lower than our model's AUC (0.95, CI: 0.93–0.96). Other studies reported varying radiologist sensitivity and specificity, with our model demonstrating numerically superior specificity at those sensitivities. Truhn et al (2019). Dalmis et al (2019) report for the radiologist sensitivity of 98% (specificity of 28%), and Zhou et al (2019) sensitivity of 59% (specificity of 86%). At these sensitivities, our model achieves a numerically superior specificity (62% and 99%, respectively). One outlier is the study by Truhn 2019 (43), reporting a higher radiologist (AUC 0.98; 95% CI:0.96, 0.99) likely due to an easier curated dataset (eg, high prevalence, only large and enhancing lesions, excluding high-risk benign lesions). However, direct statistical comparison is impossible as datasets are unavailable and potential heterogeneity. While our model outperformed the existing state-of-the-art (10) when tested on sagittal data, performance matched that on axial test data, using a much simpler model. We expect performance gains when fine-tuning the model on such axial MRIs.

This study has several limitations. First, while our dataset is uniquely large, it primarily comprises sagittal MRI scans from a single institution. Although the model generalized well to external axial scans, broader multi-institutional training would likely improve performance further. Second, the training approach relies on slice-level annotations, which may not be routinely available. Third, while we did include demographic information, we did not explore its added value in detail. Finally, while model performance is comparable to radiologists on retrospective datasets, a prospective reader study on the same data are needed for a direct comparison.

Training a 2D CNN from scratch on a large and diverse breast MRI dataset enabled state-of-the-art cancer detection, even in challenging clinical cases. Incorporating slice-level lesion annotations during training improved both classification performance and interpretability by highlighting relevant slices. The model demonstrates generalizability across different MRI protocols and institutions. Future studies should explore prospective clinical validation. To this end, we are openly releasing source code and trained weights. We hope this will enhance reproducibility in AI for radiology and encourage further technical development.

Author affiliations:

[1] Department of Biomedical Engineering, City College of the City University of New York, 160 Convent Ave, New York, NY 10031

[2] Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY

[3] Levich institute and Department of Physics, City College of the City University of New York, New York, NY

Received XXX; revision requested XXX; revision received XXX; accepted XXX

Address correspondence to: L.C.P. (email: parra@ccny.cuny.edu).

Funding: This work was supported by NIH grant R01CA247910 with additional support from NIH/NCI P30 CA008748 grant and NIH BRAIN Initiative Grant R01 EB028157.

Author contributions: Guarantors of integrity of entire study, **L.H., L.C.P.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **L.H., E.J.S., M.H., H.A.M., L.C.P.**; clinical studies, **E.J.S., B.K., M.H., D.M.**; experimental studies, **L.H., E.J.S., B.K., D.M., L.C.P.**; statistical analysis, **L.H., H.A.M., L.C.P.**; and manuscript editing, **L.H., E.J.S., Y.H., M.H., H.A.M., L.C.P.**

Disclosures of conflicts of interest: L.H. No relevant relationships. **E.J.S.** No relevant relationships. **Y.H.** No relevant relationships. **B.K.** No relevant relationships. **M.H.** No relevant relationships. **D.M.** No relevant relationships. **H.A.M.** No relevant relationships. **L.C.P.** No relevant relationships.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Jemal, A. Cancer statistics, 2022. *CA*. CA A Cancer J Clinicians 2022;72(1):7–33.

2. Pinsky PF. Principles of Cancer Screening. Surg Clin North Am 2015;95(5):953–966.

3. Mann RM, Kuhl CK, Moy L. Contrast-enhanced MRI for breast cancer screening. J Magn Reson Imaging 2019;50(2):377–390.

4. Hylton NM, Gatsonis CA, Rosen MA, et al; ACRIN 6657 Trial Team and I-SPY 1 TRIAL Investigators. Neoadjuvant Chemotherapy for Breast Cancer: Functional Tumor Volume by MR Imaging Predicts Recurrence-free Survival-Results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL. Radiology 2016;279(1):44–55.

5. Mann RM, Athanasiou A, Baltzer PAT, et al; European Society of Breast Imaging (EUSOBI). Breast cancer screening in women with extremely dense breasts recommendations of the European Society of Breast Imaging (EUSOBI). Eur Radiol 2022;32(6):4036–4045.

6. Sheth D, Giger ML. Artificial intelligence in the interpretation of breast cancer on MRI. Eur J Magn Reson Imaging 2020;51(5):1310–1324.

7. Bhowmik A, Eskreis-Winkler S. Deep learning in breast imaging. BJR Open 2022;4(1): 20210060.

8. Gavenonis SC, Lee JM, Halpern EF, Rafferty EA. Positive predictive value of breast MRI in cancer detection. Cancer Res 2009;69(2_Supplement):4007.

9. Han BK, Schnall MD, Orel SG, Rosen M. Outcome of MRI-Guided Breast Biopsy. AJR Am J Roentgenol 2008;191(6):1798–1804.

10. Witowski J, Heacock L, Reig B, et al. Improving Breast Cancer Diagnostics with Artificial Intelligence for MRI. Science Translational Medicine 2022;14(664):eabo4802.

11. Wang X, Yang W, Weinreb J, et al. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning vs non-deep learning. Sci Rep 2017;7(1):15415.

12. Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification. IEEE J Biomed Health Inform 2017;21(1):31–40.

13. Zhu Z, Albadawy E, Saha A, et al. Deep learning for identifying radiogenomic associations in breast cancer. Comput Biol Med 2019;109:85–90.

14. Suk HI, Lee SW, Shen D; Alzheimer's Disease Neuroimaging Initiative. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal 2017;37:101–113.

15. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018; 73(5):439–445.

16. Yoo Y, Tang LYW, Brosch T, et al. Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls. Neuroimage Clin 2018;17:169–178.

17. Li N, Haopeng L, Bin Q, et al. Detection and Attention: Diagnosing Pulmonary Lung Cancer from CT by Imitating Physicians. arXiv 2017. Preprint posted online December 14, 2017; https://arxiv.org/abs/1712.05114.

18. Saiz F, Barandiaran, I. COVID-19 Detection in Chest X-ray Images using a Deep Learning Approach. https://doi.org/10.9781/ijimai.2020.04.003. Published June 2020. Accessed DATE.

19. Halling-Brown MD, Warren LM, Ward D, et al. OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data. Radiol Artif Intell 2021;3(1):e200103.

20. Cai H, Wang J, Dan T, et al. An Online Mammography Database with Biopsy Confirmed Types. Sci Data 2023;10(1).

21. Lopez MG, Posada N, Moura DC, et al. BCDR: a breast cancer digital repository. In: 15th International conference on experimental mechanics, volume 1215: 113–120. https://www.researchgate.net/profile/Jose-Franco-Valiente/publication/258243150_BCDR_A_BREAST_CANCER_DIGITAL_REPOSITORY/links/59afe98a0f7e9bf3c72930e5/BCDR-A-BREAST-CANCER-DIGITAL-REPOSITORY.pdf.

22. Heath M, et al. Current Status of the Digital Database for Screening Mammography. In: Karssemeijer N, Thijssen M, Hendriks J, Van Erning L, eds. Digital Mammography. Springer Netherlands, Dordrecht, 1998: 457–460.

23. Amit G, Ben-Ari R, Hadad O, Monovich E, Granot N, Hashoul S. Classification of breast MRI lesions using small-size training sets: comparison of deep learning approaches. In: Medical Imaging, Computer-Aided Diagnosis; SPIE, 2017: 374–379.

24. Zhou J, Luo LY, Dou Q, et al. Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images. J Magn Reson Imaging 2019;50(4):1144–1151.

25. McDermott MBA, Wang S, Marinsek N, et al. Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med 2021;13(586):eabb1655.

26. Burns ML, Kheterpal S. Machine Learning Comes of Age: Local Impact vs National Generalizability. Anesthesiology 2020;132(5):939–941.

27. Barak-Corren Y, Chaudhari P, Perniciaro J, et al. Prediction across healthcare settings: a case study in predicting emergency department disposition. npj Digit Med 2021;4(1):1–7.

28. Mann RM, Cho N, Moy L. Breast MRI: State of the Art. Radiology 2019;292(3):520–536.

29. Verburg E, van Gils CH, van der Velden BHM, et al. Deep Learning for Automated Triaging of 4581 Breast MRI Examinations from the DENSE Trial. Radiology 2022;302(1):29–36.

30. Herent P, Schmauch B, Jehanno P, et al. Detection and characterization of MRI breast lesions using deep learning. Diagn Interv Imaging 2019;100(4):219–225.

31. Dalmiş MU, Gubern-Mérida A, Vreemann S, et al. Artificial Intelligence–Based Classification of Breast Lesions Imaged With a Multiparametric Breast MRI Protocol With Ultrafast DCE-MRI, T2, and DWI. Invest Radiol 2019;54(6):325–332.

32. Li J, Fan M, Zhang J, Li L. Discriminating between benign and malignant breast tumors using 3D convolutional neural network in dynamic contrast enhanced-MR images. In: Cook TS, Zhang J, eds. Proceedings of the SPIE 2017, vol 10138; 1013808.

33. Zhang Y, Liu YL, Nie K, et al. Deep Learning-based Automatic Diagnosis of Breast Cancer on MRI Using Mask R-CNN for Detection Followed by ResNet50 for Classification. Acad Radiol 2023;30 Suppl 2(Suppl 2):S161–S171.

34. Saha A, Harowicz MR, Grimm LJ, et al. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. Br J Cancer 2018;119(4):508–516.

35. Hirsch L, Huang Y, Luo S, et al. Radiologist-Level Performance by Using Deep Learning for Segmentation of Breast Cancers on MRI Scans. Radiol Artif Intell 2022;4(1):e200231.

36. Smith I, Procter M, Gelber RD, et al; HERA study team. 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. Lancet 2007;369(9555):29–36.

37. Emens LA, Davidson NE. The follow-up of breast cancer. Semin Oncol 2003;30(3):338–348.

38. Modat M, Ridgway GR, Taylor ZA, et al. Fast free-form deformation using graphics processing units. Comput Methods Programs Biomed 2010;98(3):278–284.

39. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy: 2999–3007.

40. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv 2014. Preprint posted online December 22, 2014; https://arxiv.org/abs/1412.6980.

41. Hu Q, Whitney HM, Li H, et al. Improved Classification of Benign and Malignant Breast Lesions Using Deep Feature Maximum Intensity Projection MRI in Breast Cancer Diagnosis Using Dynamic Contrast-enhanced MRI. Radiol Artif Intell 2021;3(3):e200159.

42. Dalmış MU, Litjens G, Holland K, et al. Using deep learning to segment breast and fibroglandular tissue in MRI volumes. Med Phys 2017;44(2):533–546.

43. Truhn D, Schrading S, Haarburger C, et al. Radiomic vs Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. Radiology 2019;290(2):290–297.

44. Kay W, Carreira J, Simonyan K, et al. The Kinetics Human Action Video Dataset. arXiv 2017. Preprint posted online May 19, 2017; https://arxiv.org/abs/1705.06950.

45. Leithner D, Helbich TH, Bernard-Davila B, et al. Multiparametric 18F-FDG PET/MRI of the Breast: Are There Differences in Imaging Biomarkers of Contralateral Healthy Tissue Between Patients With and Without Breast Cancer? J Nucl Med 2020;61(1):20–25.

46. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv 2015. Preprint posted online December 10, 2015; https://arxiv.org/abs/1512.03385.

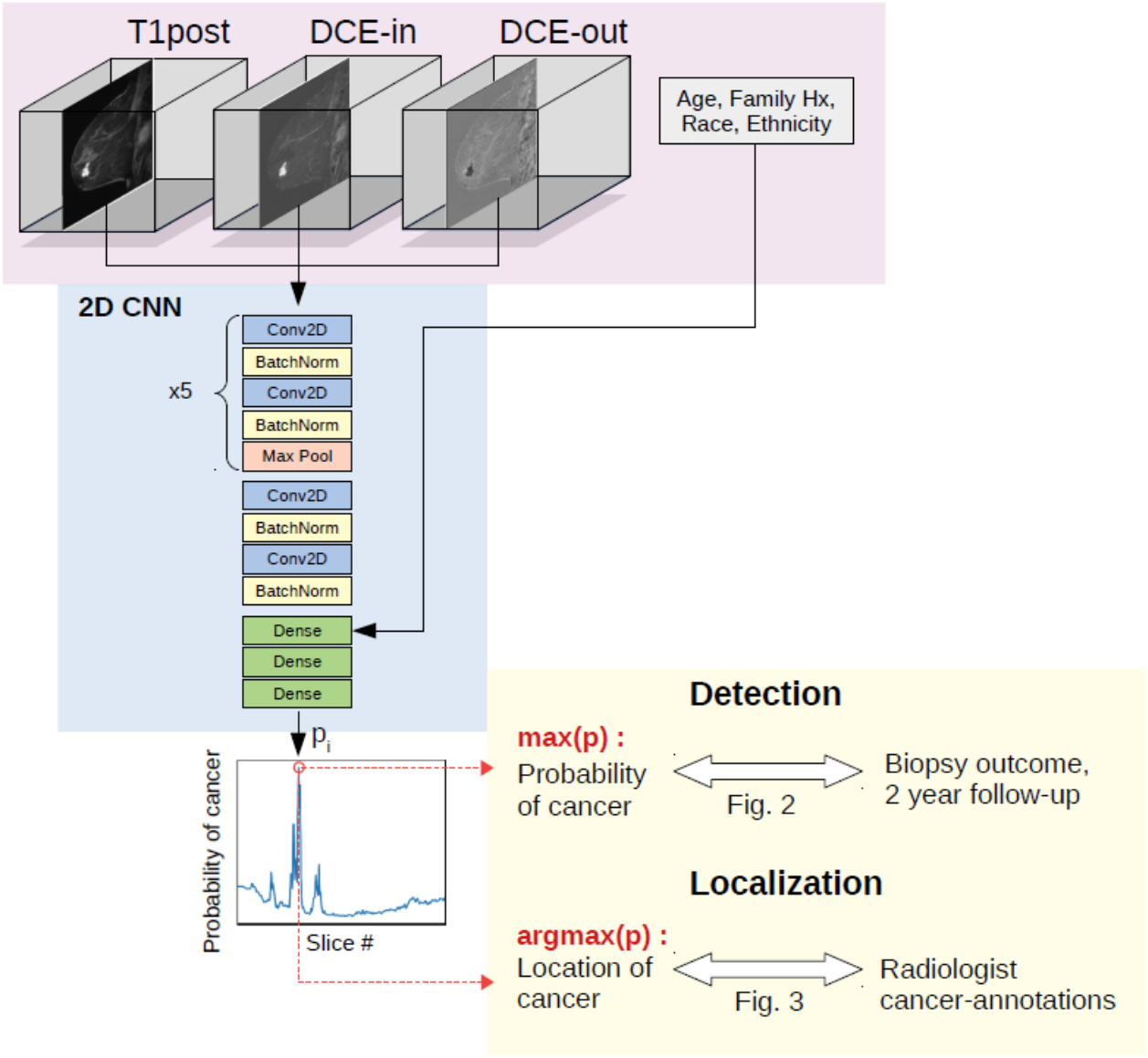


Figure 1: Data overview for breast cancer detection model development and evaluation. Three datasets were utilized: sagittal data from the primary clinical site (training, validation, testing), axial data from the primary site (testing), and axial data from a secondary clinical site (testing). The sagittal dataset was partitioned at the patient level (indicated as % for each partition). No patient overlap existed between the primary site's sagittal and axial data. Each box displays the number of patients, examinations, individual breasts (accounting for bilateral and unilateral examinations), malignant breasts, and malignant breasts with radiologist annotations indicating cancer location.

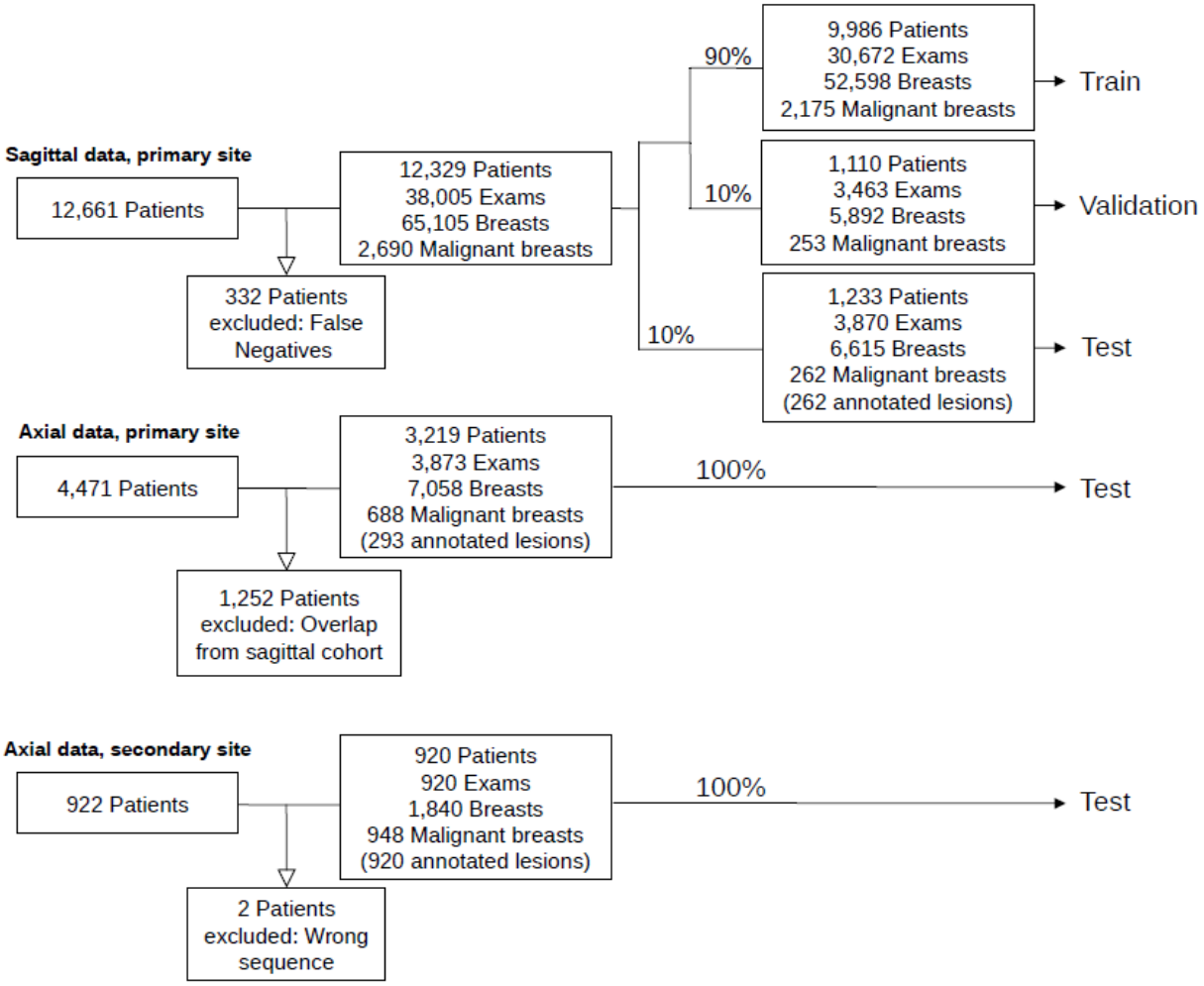


Figure 2: Cancer detection and localization using AI. The input of the model (purple box) consists of a sagittal slice from the full 3D MRI with three channels (T1 post, Dynamic Contrast Enhancement (DCE)-in, Dynamic Contrast Enhancement (DCE)-out). The model (blue box) is a 2D deep convolutional neural network (CNN) that outputs a probability of cancer for each slice. Model output is evaluated in two tasks, detection and localization (yellow box). The maximum probability (max) is used as the prediction for the whole volume and evaluated in the detection task against the reference standard (pathology from the biopsy of clinical follow-up). Similarly, the slice index with maximum probability (argmax) serves as the estimate of tumor location and is evaluated against the reference standard provided as annotations on cancer location made by radiologists.

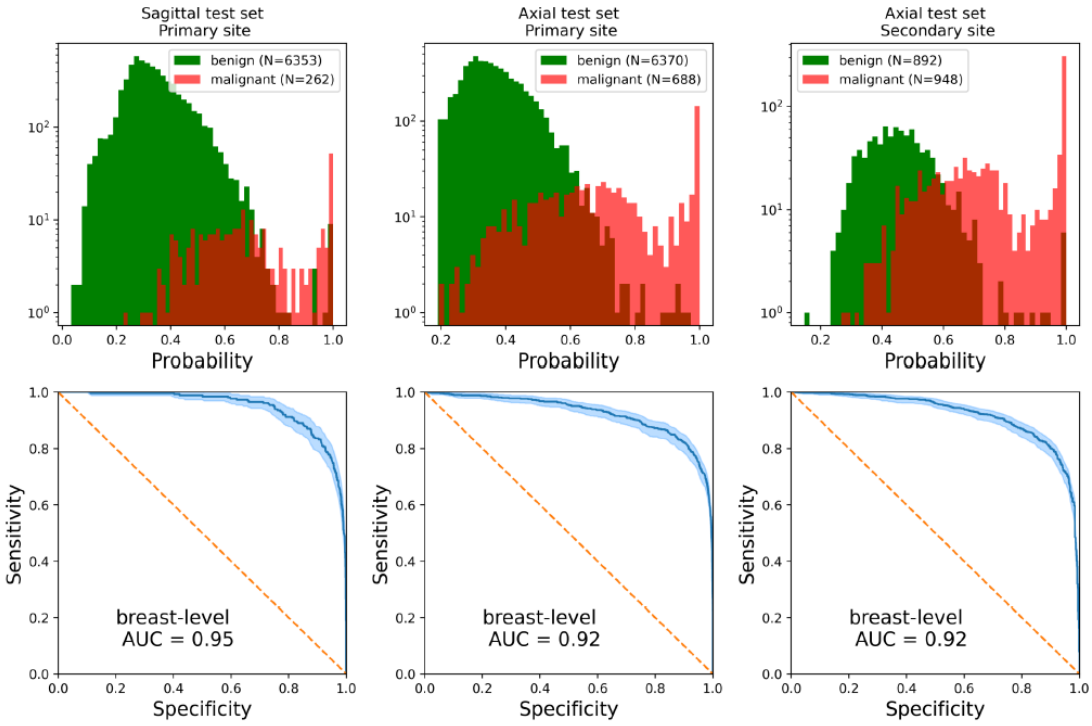


Figure 3: Test set performance on sagittal data and generalization to axial data across sites. **(A)** Top: Histogram of predicted probability of cancer for all breasts in the sagittal test set from the primary site, color-coded by breast outcome. Bottom: ROC curve and AUC. 95 percentile bootstrap confidence interval shown in shaded blue. **(B)** Same as A but for the axial test set from the primary site. **(C)** Same as A but the axial test set from the secondary site

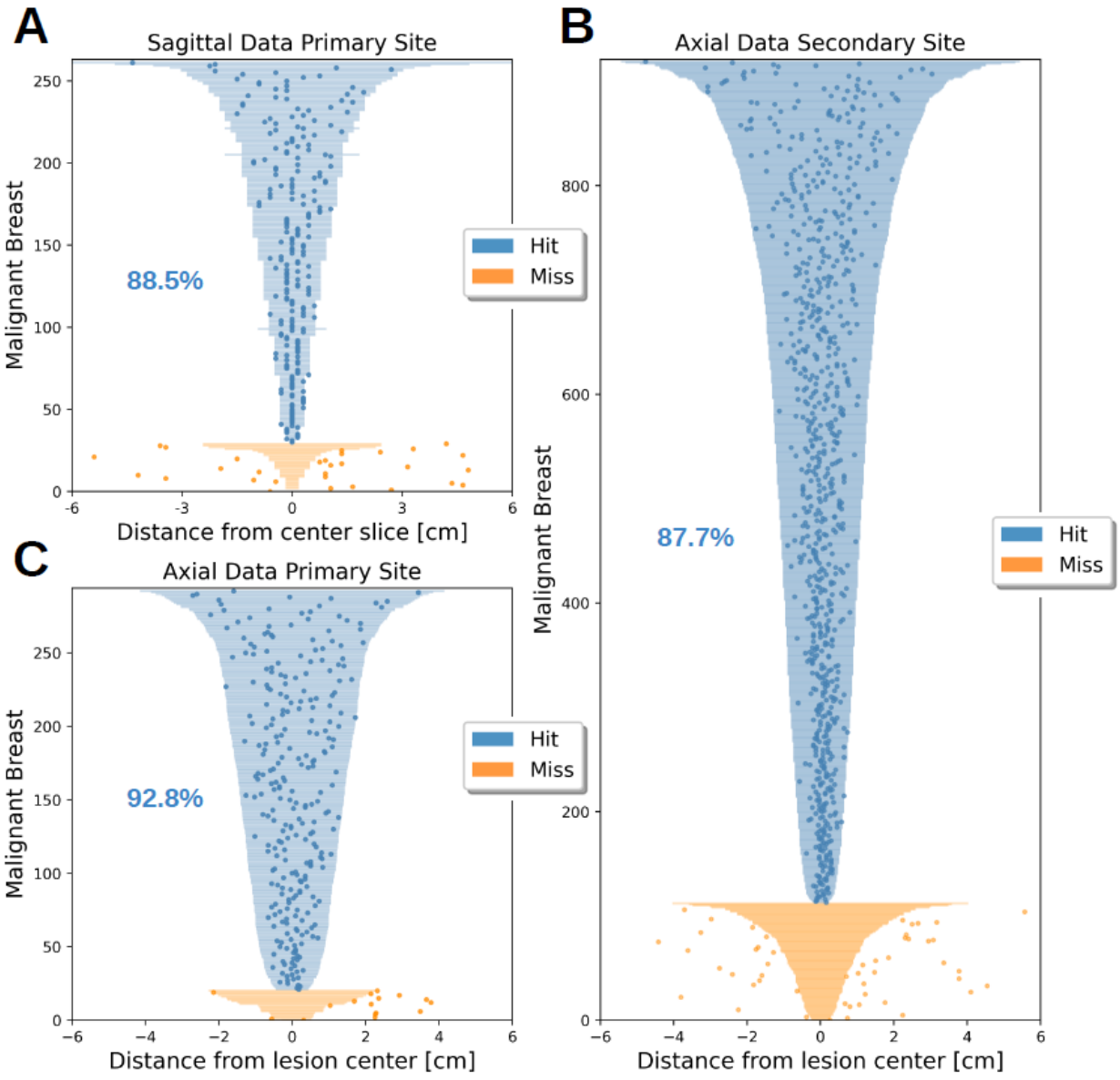


Figure 4: Predicted location and lesion lateral extent: Comparison of the location predicted by the network (dot) relative to the lateral extent of the cancer provided by radiologists (horizontal lines/shading). Lateral extent indicates the range of sagittal slices that contain cancer. Vertical axis indicates different breasts sorted by extent. **(A)** Sagittal test data of the primary site (n = 262). For this data, lateral extent for each cancer is based on a semiautomatic segmentation. A "hit" indicates that the location predicted by the network falls within the lesion (blue); a "miss" indicates no overlap (orange). The percentage of hits is shown in blue text. Breasts are sorted by decreasing lesion size for both hits and misses. **(B)** Same as panel A but uses axial data from the secondary site (n = 920). Here, cancer lateral extent is determined from bounding boxes provided by radiologists. **(C)** Same as panel A but uses axial data from the primary site (n = 293). Here, cancer lateral extent is based on radiologist's volumetric segmentations. Most of the large cancers missed in the Axial Primary Site correspond to cancers found in the left breast, which explains the rightward shift of the orange dots (this effect, however, is not statistically significant (binomial test: cancers = 21, left-sided = 14, population frequency of left cancers = 0.53 (156 of a total of 293), P = .15).

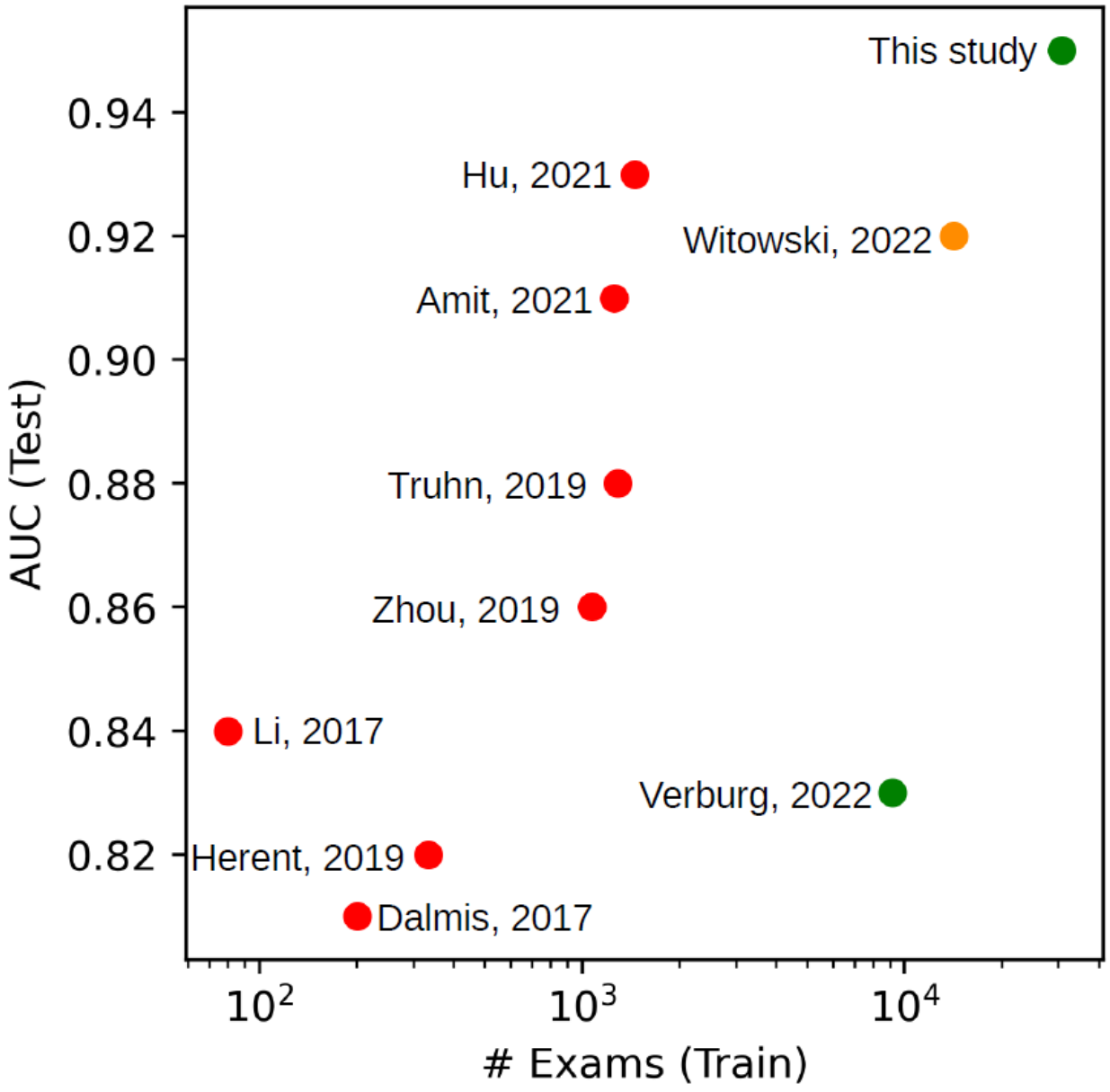


Figure 5: Comparison of related studies to date. Each study is shown in terms of the size of the training data (number of examinations) versus performance on the internal test set of each study (AUC). Color indicates if studies make code and trained weights openly available (red: not available, orange: upon request, green: no restriction).

Table 1: Summary of Patient Demographics, examination Counts, Imaging Protocol Distribution, Cancer Labels, and BI-RADS Categories for the Sagittal Training, Validation, and Test Sets, as well as the External Axial Test Set

		Sagittal Train	Sagittal Val	Sagittal Test	Sagittal Total	Axial Total	Total
Patients		9986 (81%)	1110 (9%)	1233 (10%)	12329	3219	15548
Race	Asian Far East/ Indian Subcont	376 (3,77%)	30 (2,7%)	52 (4,22%)	458	60 (1,86%)	458
	Black / African american	533 (5,34%)	67 (6,04%)	74 (6%)	674	62 (1,93%)	674
	native american/alaska	1 (0,01%)	0 (0%)	0 (0%)	1	0 (0%)	1
	Native hawaiian/ pacific islander	3 (0,03%)	0 (0%)	0 (0%)	3	0 (0%)	3
	White	6744 (67,53%)	753 (67,84%)	832 (67,48%)	8329	732 (22,74%)	8329
	Unknown/Missing	2329 (23,32%)	260 (23,42%)	275 (22,3%)	2864	107 (3,32%)	2864
Ethnicity	Hispanic	459 (4,6%)	55 (4,95%)	76 (6,16%)	590	56 (1,74%)	590
	Not Hispanic	7252 (72,62%)	797 (71,8%)	886 (71,86%)	8935	823 (25,57%)	8935
	Other/Unknown	2275 (22,78%)	258 (23,24%)	271 (21,98%)	2804	81 (2,52%)	2804
examinations		30672	3463	3870	38005	3873	41878
Age, (mean, STD)		52.1 ± 11.2	51.9 ± 11.1	52.3 ± 10.8		54.4 ± 10.6	
examination Label	Malignant	2114 (6,89%)	238 (6,87%)	255 (6,59%)	2607	688 (17,76%)	3295
	Benign/Negative	28558 (93,11%)	3225 (93,13%)	3615 (93,41%)	35398	3178 (82,06%)	38576
BI-RADS Category	BI-RADS 0	347 (1,13%)	40 (1,16%)	44 (1,14%)	431	4 (0,1%)	435
	BI-RADS 1	3949 (12,87%)	545 (15,74%)	534 (13,8%)	5028	870 (22,46%)	5898
	BI-RADS 2	17712 (57,75%)	1913 (55,24%)	2182 (56,38%)	21807	2054 (53,03%)	23861
	BI-RADS 3	3792 (12,36%)	409 (11,81%)	489 (12,64%)	4690	248 (6,4%)	4938
	BI-RADS 4	2455 (8%)	287 (8,29%)	321 (8,29%)	3063	195 (5,03%)	3258
	BI-RADS 5	446 (1,45%)	46 (1,33%)	72 (1,86%)	564	10 (0,26%)	574
	BI-RADS 6	1943 (6,33%)	221 (6,38%)	224 (5,79%)	2388	492 (12,7%)	2880
Lesion size cm (mean, STD)		1.8 (± 1.2)	1.8 (± 1.2)	1.8 (± 1.2)		2.7 (± 1.2)	

Data include race, ethnicity, age, and lesion size statistics. Data are presented as numbers (percentage) unless otherwise indicated. Age and lesion size are reported as mean ± SD (SD).

Table 2: Model Performance with Challenging Cases

	No			Yes			
Attribute	Benign	Malignant	AUC	Benign	Malignant	AUC	AUC Difference
Implant	5913	260	0.95	393	2	>0.99	*P* = .79
Biopsy Clip	4347	173	0.94	1940	89	0.95	*P* = .73
Postsurgery change	6184	258	0.96	122	4	0.91	*P* = .20
Poor image quality	5091	207	0.94	1215	55	0.96	*P* = .75

Model performance did not significantly change for cases considered challenging and excluded in previous studies.[10,23,24] "Poor image quality" encompasses various issues (see Methods)

Table 3: Summary of Studies to date, Involved in Detection of Breast Cancer in MRI using Deep Neural Networks

Publication	Pretrained Model	AUC	N_Train	N_Test	Comments	Published Model	Use Contralateral	Multisite	MRI Protocol
(Herent et al 2019)	ResNet50 (ImageNet)	0.82	335	168	Evaluation on single slice. Generates attention heatmaps through initial segmentation	no	yes (axial)	no	axial
(Truhn et al 2019)	ResNet18 (ImageNet)	0.88	1294	647	Manual cropping of lesions. Radiologist performance of 0.98. Post contrast sequence as RGB channels	no	unclear (no)	no	axial
(Amit et al 2017)	CIFAR-10, VGGNet	0.91	1256	1256	Requires manual selection of ROIs on lesions. Only BI-RADS 2 benign and BI-RADS 5 malignants. Only single lesions. No asymmetric BPE. Pretrained models performed worse than model trained from scratch	no	not specified	no	not specified
(Hu et al 2021)	VGG16 (ImageNet)	0.93	1455	535	DCE time sequence as RGB channels. Manual cropping of lesions before input to CNN	no	no	no	sagittal
(Zhou et al 2019)	No	0.86	1073	307	Only single lesions. No asymmetric BPE. Only evident BPE	no	yes (axial)	no	axial
(Verburg et al 2022)	No	0.83	9162	4581	Triages 40% of normal breasts at NPV = 100% in women with extremely dense breasts (DENSE trial)	no	yes (axial)	yes	axial
(Dalmış et al 2017)	No	0.81	201	160	No spatial registration of contralateral. No use of postcontrast images	no	yes	no	transverse and coronal
(Li et al 2017)	No	0.84	80	43	Manual cropping of lesions	no	no	no	axial
(Witowski et al 2022)	3D ResNet-18 (Kinetics-400)	0.92	14198	3936	No interpretability of results.	under request case-by-case	yes (axial)	yes	Sagittal and axial
(Zhang et al 2023)	Mask *R-CNN* and ResNet50	Sens 96% Spec 70%	241	176	Outputs bounding box on lesion.	no	yes (axial)	no	axial
This study	No	0.95 (axial 0.92)	52598 30672 examinations	6615	Identifies relevant slice and tumor location in slice.	yes	yes	yes	Sagittal and axial

Addition of the contralateral breast:

Radiologists normally will look at the contralateral breast whenever possible as a reference for symmetry in respect to breast density, amount of fibroglandular tissue and background parenchymal enhancement. These are some of the important biomarkers related to risk of breast cancer.[45] Several previous works have not included the contralateral breast, or did so only indirectly through use of axial exams that include both breasts [10,24,29,30] (Table 2). Despite the importance of having the contralateral breast available for radiologists, we don't observe any increase in performance in the validation set when adding to the model (using the validation set where the contralateral breast was available: ROC of 0.943 vs ROC of 0.943 with contralateral). We do observe a slight increase in the precision-recall curve (average precision without contralateral = 0.66 vs 0.68, see Table S2)

We tested three different ways to add contralateral breast. 1) As three additional input channels (Table S2: CNN + Contralateral 1); 2) as a parallel pathway merched by concatenating in the 8th convolutional layer. This was trained by freezing the entire model and only training on the parallel pathway (Table S2: CNN + Contralateral 2), and 3) same as 2, but training also the layers after the concatenation point (Table S2: CNN + Contralateral 3).

Model prediction by BI-RADS

BI-RADS 1 and 2 indicate a "Negative" and "Benign" exam; BI-RADS 3 is a "probably benign" assessment. "Suspicious" or "highly suggestive of malignancy" findings are BI-RADS 4 and 5, respectively. Fig. S6-Left indicates the numbers of cases broken down by BI-RADS and detection for the sagittal test set from the primary site. As expected, the predicted probability of cancer increases with BI-RADS score (Fig. S6-Right).

Performance on biopsied breasts

Here we report the performance of the model in predicting the outcome of a biopsy (that was prompted by the MRI). To do this we included only BI-RADS 4 and 5 and obtained an AUC of 0.86. Other exams are not called biopsy (BI-RADS 1-3) or may already have a biopsy (BI-RADS 6). Performance is lower here because the exams that are excluded are easier cases, i.e. BI-RADS 2-3, which are easier to identify as benign, and BI-RADS 6, are known cancer that had been detected previously, e.g. with mammography.

Implementation of ResNet50

The ResNet50[46] model was implemented with the tf.keras.applications library. The model has 23,538,690 trainable parameters. Every layer has a L2 weight penalty of 1e-4 for regularization purposes. The last layer was not included and the input shape was set to 512x512x3 to match our data. Inputs are specifically pre-processed for this model by passing the data through the preprocessing pipeline provided by the function keras.applications.resnet.preprocess_input. For the ImageNet pre-trained model, the input size is constrained to 221x221x3, so input data is appropriately resized. See Fig. S9 for examples of predictions in the re-formatted breast image.

Focal loss and Binary Cross-entropy:

Binary cross-entropy loss function is the standard metric used in many classification problems. Others have proposed the "focal-loss" function [39] for classification in the presence of class-imbalance, which is the situation in this study. We compared the two loss functions on the validation set and found better performance for the focal-loss. It is interesting to compare the actual distributions between the two cost functions (compare Fig. 3A to Fig. S10). It is evident that the focal loss does not force the distributions to extreme probabilities, as more uncertain predictions are less heavily penalized. This is advantageous in problems with class imbalance. For the present data, the binary cross-entropy focuses on reducing the predicted probabilities of the majority class (benign), even when they are already low. Instead, the focal loss focuses on correcting a few examples of incorrect minority classes (malignant). Focal loss is more tolerant of uncertainty, whereas binary cross-entropy emphasizes confidence.

Demographic and Clinical Factors Associated with Pathology Outcome

We evaluated whether any of the demographic and clinical information used correlated with a pathology outcome. For this we performed a chi-squared test on each categorical variable (Table S3). Taking into account multiple comparisons, only three variables are significantly correlated: Family History, Ethnicity Unknown and Race Unknown. Each of these variables result in the following point in a ROC-curve: Family History: Specificity = 0.777, Sensitivity = 0.012. Ethnicity Unknown: Specificity = 0.793, Sensitivity = 0.175. Race Unknown: Specificity = 0.793, Sensitivity = 0.176. These results point out that in this specific high-risk population, demographic information alone results in poor classification and proper evaluation of the MR images is essential for this task.

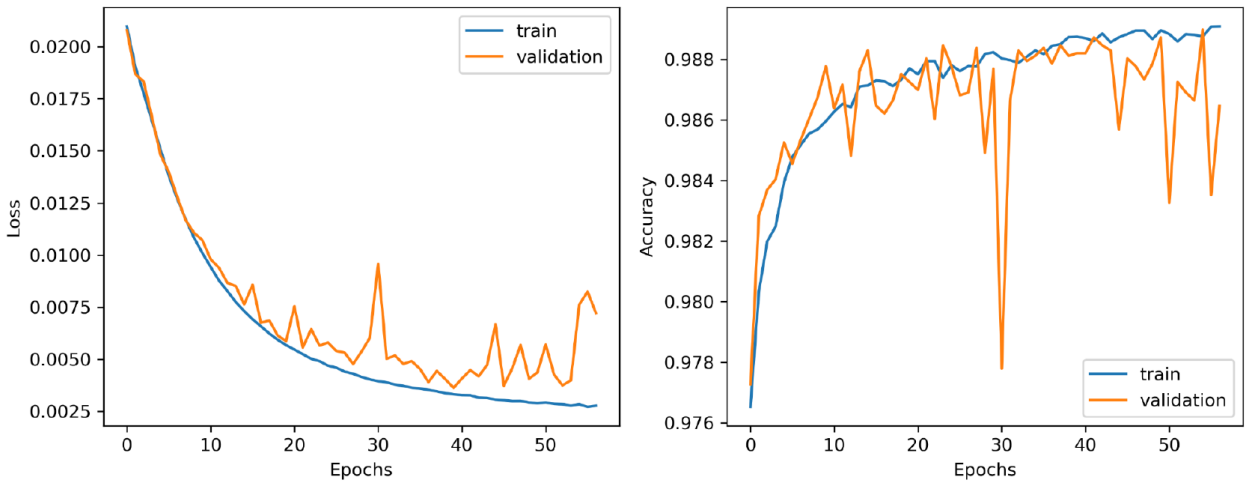


Figure S1: Training curves for monitoring overfitting and selecting the best validation-set performance with early stopping. Left: Loss function per training epoch evaluated on the training data (blue) and validation data (orange). Training was stopped early if validation loss did not drop further after 20 epochs (here lowest loss at 38th epoch). Right: Accuracy (at a threshold of 0.5 for the probability) during training shows decrease as overfitting starts occurring.

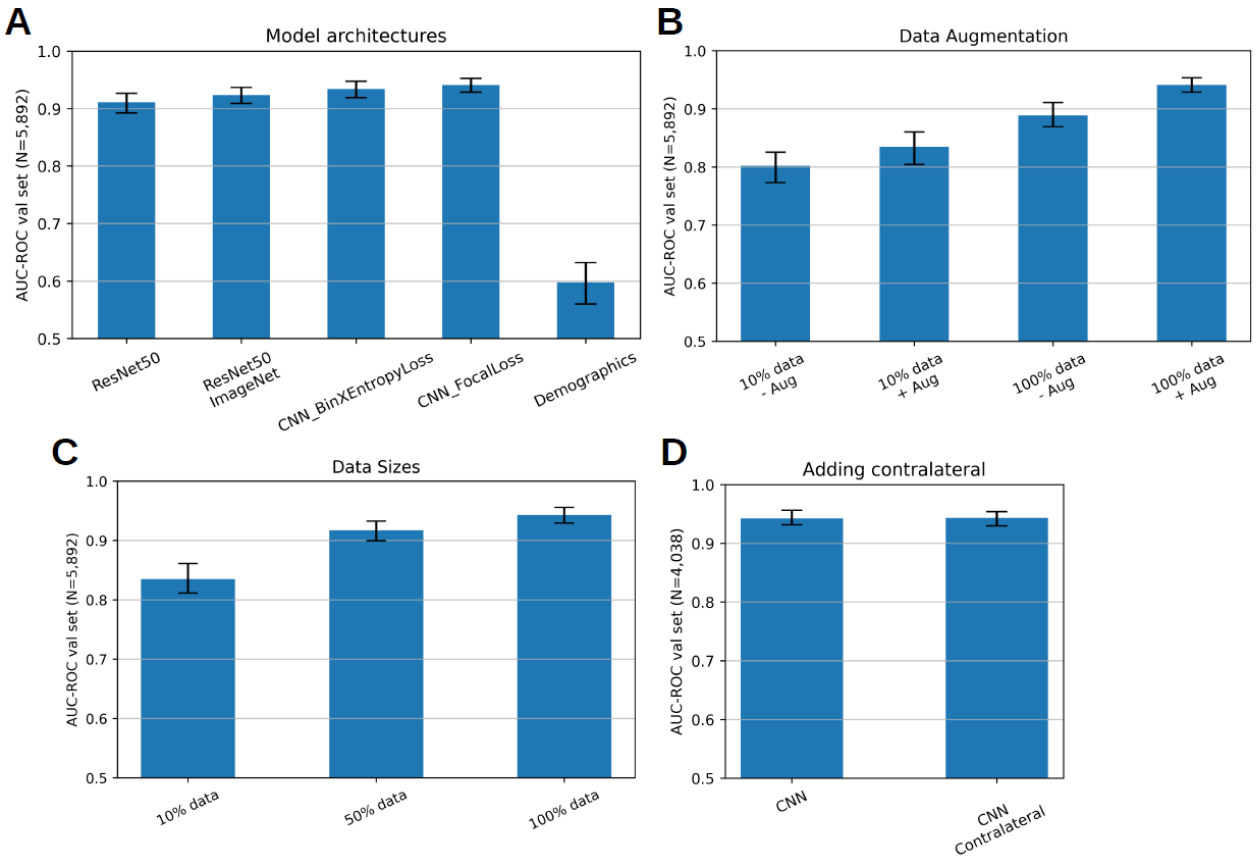


Figure S2: Performance of different models and data sets, evaluated on the validation set. In each barplot the performance (vertical axis) is measured as the area under the receiver operating characteristic curve (AUC-ROC). A) A simple CNN performs just as well as the much larger state-of-art model ResNet50. ResNet50 trained on ImageNet numerically outperformed training it from scratch (DeLong test: p=0.077, z=1.77). However, the simpler CNN model, trained from scratch, outperformed a ResNet50 (DeLong test: p=0.0014, z=3.19). Demographic information adds little information to this high-risk population. B) Data augmentation techniques for 2D images are very beneficial for performance, even more when there is a larger data set available. C) Increasing data sizes improve model performance in an asymptotic way. D) Adding contralateral breast does not significantly improve the AUC-ROC, although it does improve precision metric (See Table S2).

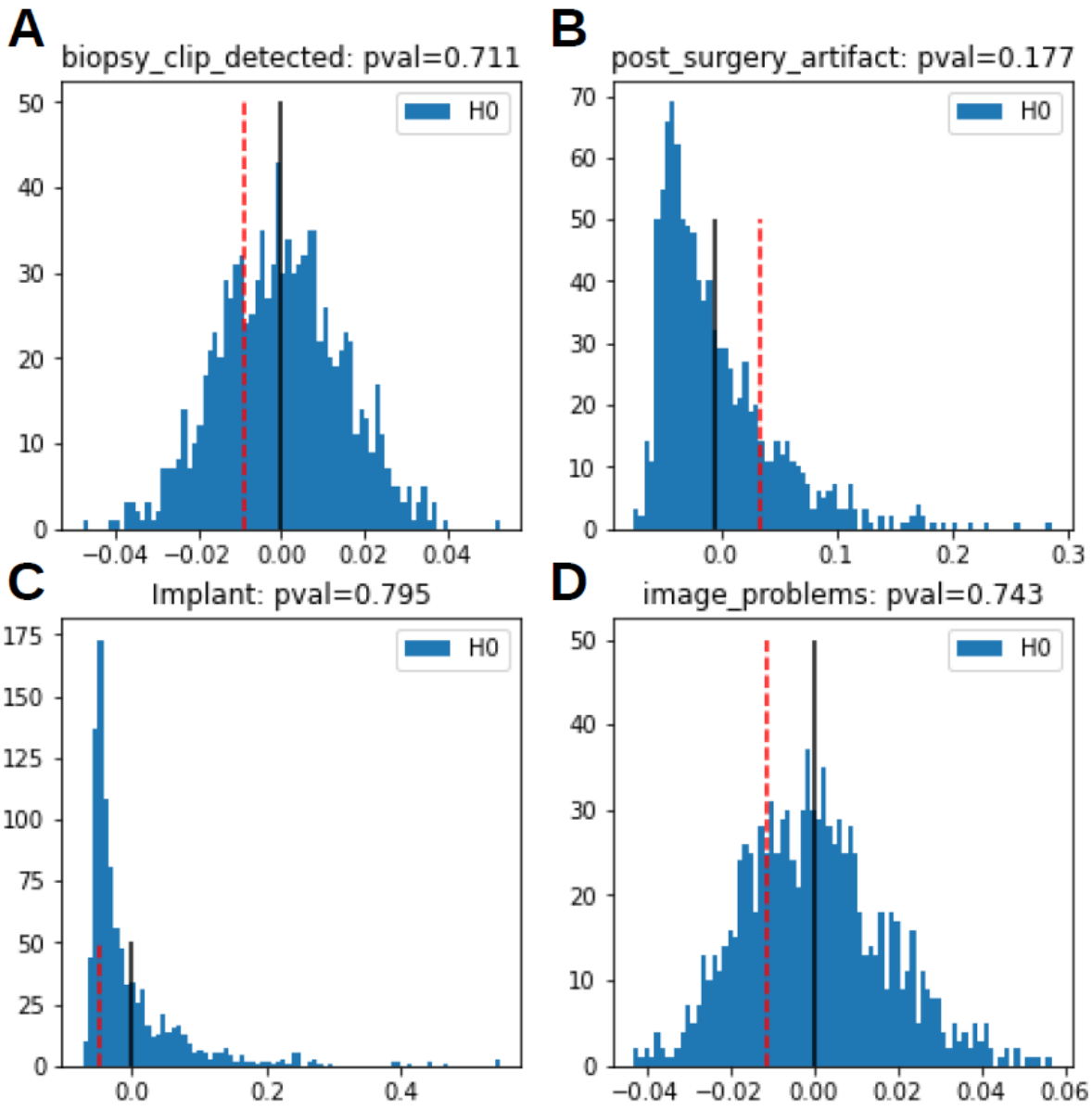


Figure S3: Difference in performance for data quality characteristics (as in Table 1). For each characteristic we measure the difference in performance in terms of AUC-ROC against a null distribution. Here we test that the presence of clinical or imaging abnormalities is not significantly worse than its absence, in a one-sided test. (A): Presence or absence of biopsy clips do not affect model performance significantly. (B): Presence or absence of post surgery changes do not affect model performance significantly. (C): Presence or absence of implants do not affect model performance significantly. (D): Problems related to image quality do not affect model performance significantly. Here we account for: Image artifacts, movement artifacts, fat-saturation artifacts, and any other image quality problems.

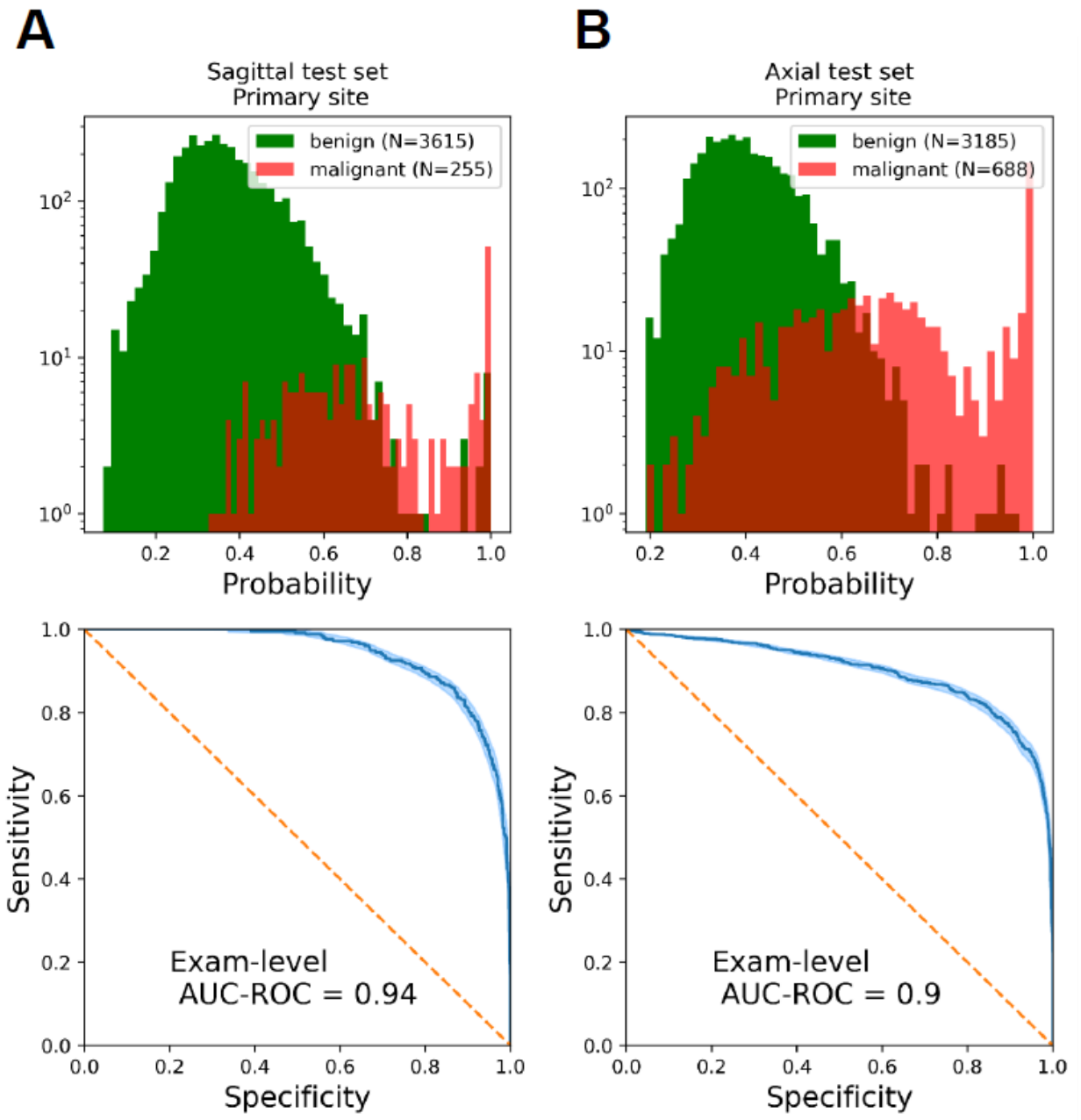


Figure S4: Performance per exam instead of breast. (A) When evaluated at the exam level, performance drops from 0.95 to 0.94 in the in-plane sagittal data from the primary site, and (B) from 0.92 to 0.9 in the out-of-plane axial data from the primary site. This evaluation does not apply to data from the secondary site, as it already consists of one exam per patient.

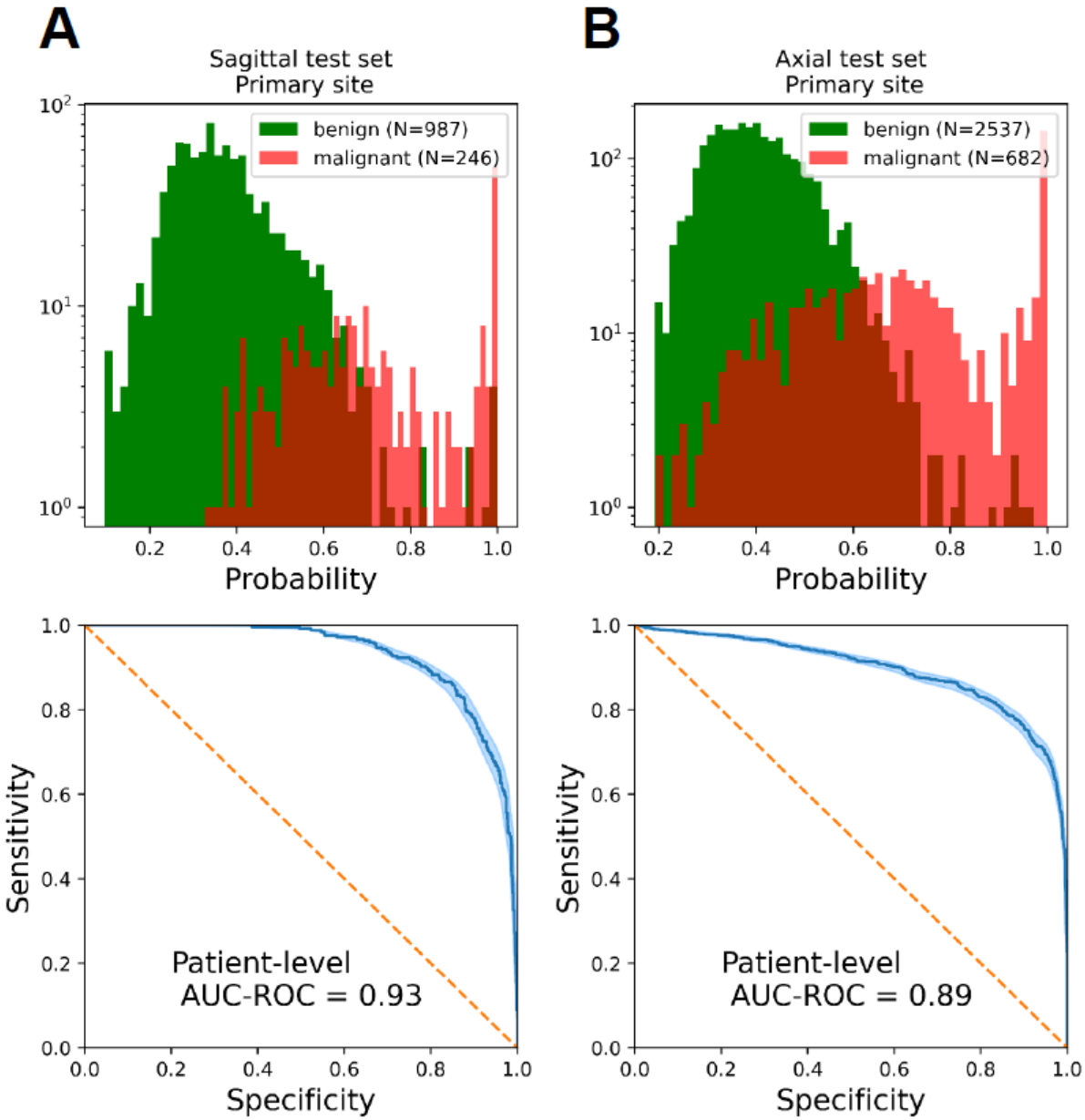


Figure S5: Results at the patient-level. For each patient, we keep only the last exam, except for positive findings, in which case we keep the exam at which cancer was identified. Per exam we take the maximum predicted value across both breasts, thereby removing any multiple entries of breasts or patients. This primarily reduced the many benign follow-up exams in screening patients. (A) Performance for the primary site sagittal data. (B) Performance for the primary site axial data. This evaluation does not apply to data from the secondary site, as it already consists of individual exams from individual patients.

	Breasts	
BI-RADS	**Benign**	**Malignant**
1	867	0
2	3793	0
3	820	3
4	453	46
5	31	48
6	8	164

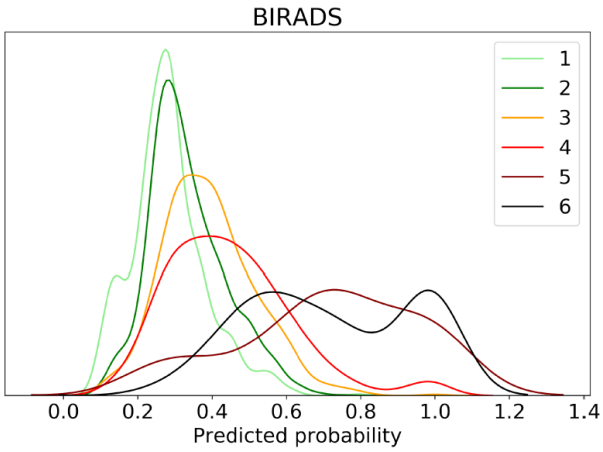


Figure S6: Distribution of cancer predictions by BI-RADS in the sagittal test set from the primary site. (Left) number of breasts broken down by BI-RADS. (Right) Model prediction of probability of cancer per breast.

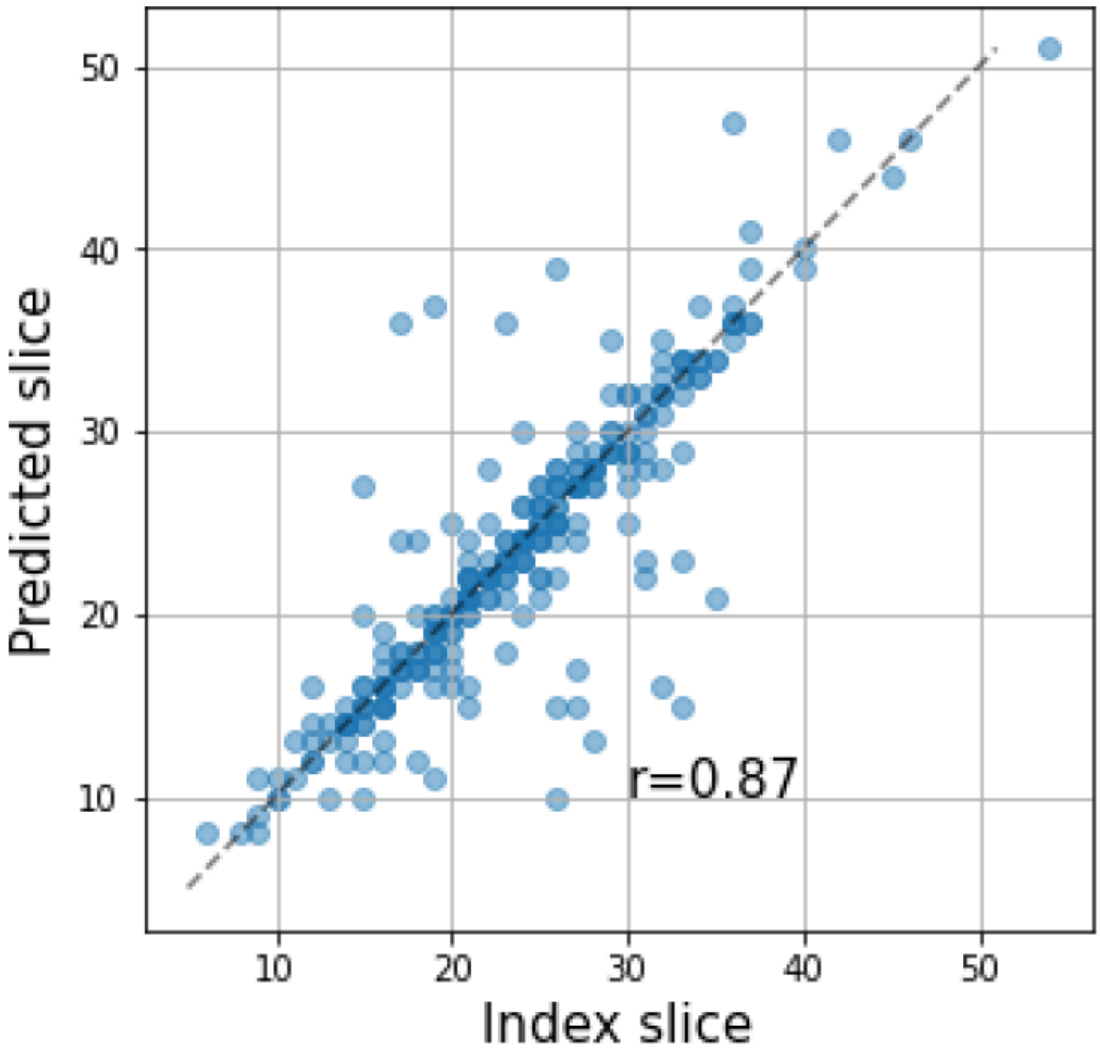


Figure S7: Correlation between AI predicted slice with maximum probability of cancer and the index slice selected by a radiologist.

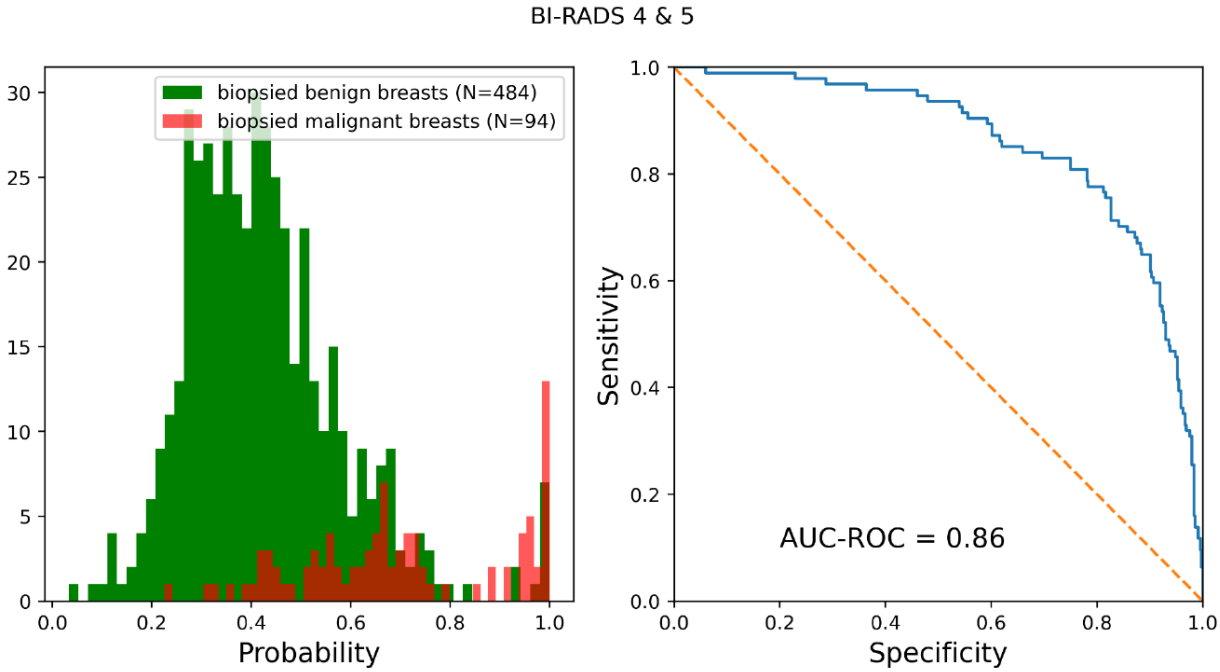


Figure S8: ROC curve for biopsied breasts only (BI-RADS 4 and 5) in the sagittal test set from the primary site.

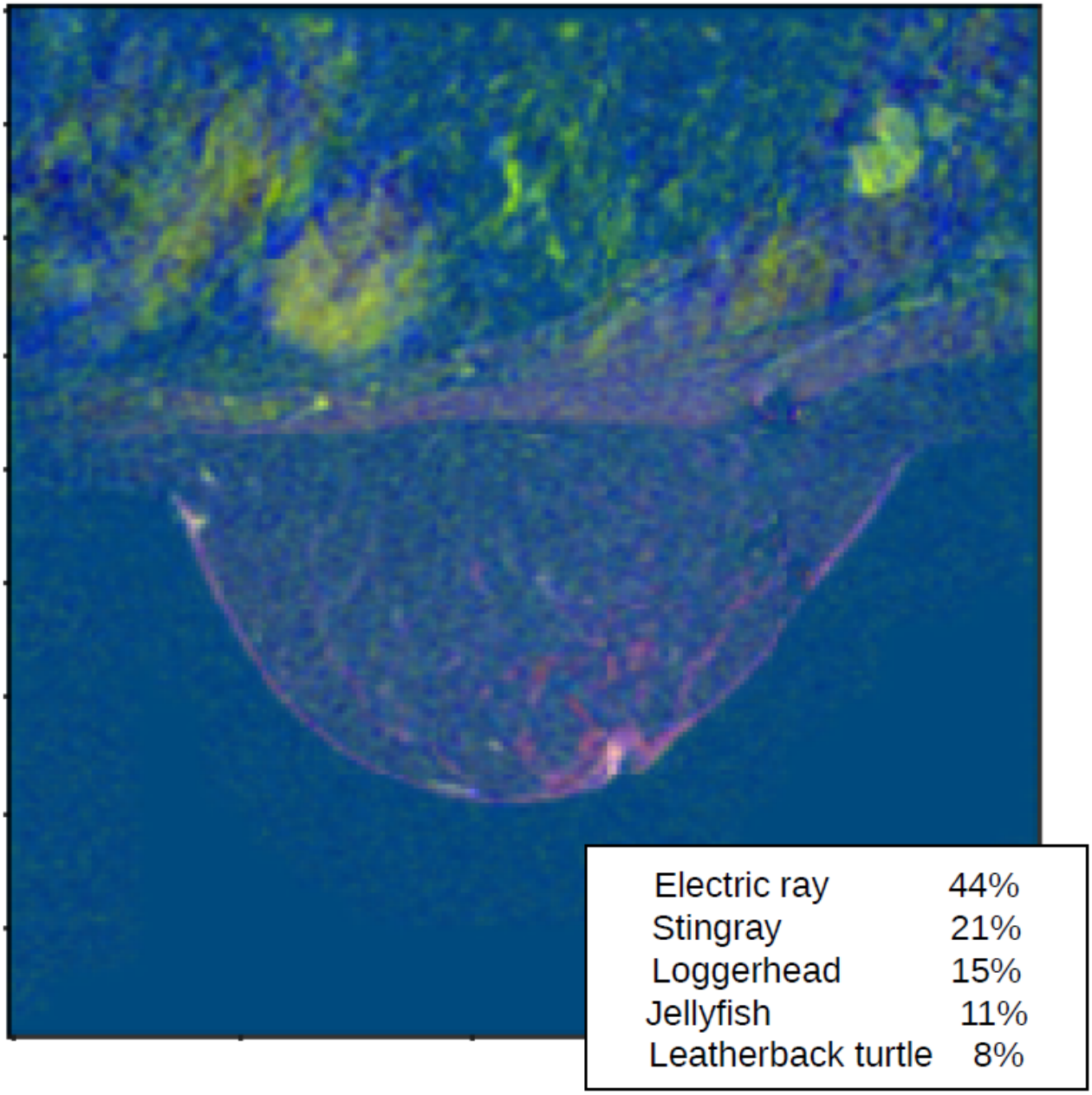


Figure S9: Fine-tuning a model pre-trained on natural images (e.g. ImageNet) will learn high-level features that do not apply well to clinical data. Evaluating the image of a breast MRI through a ResNet50 pre-trained on ImageNet extracts features that prompts the classifier to recognize the image as a marine animal.

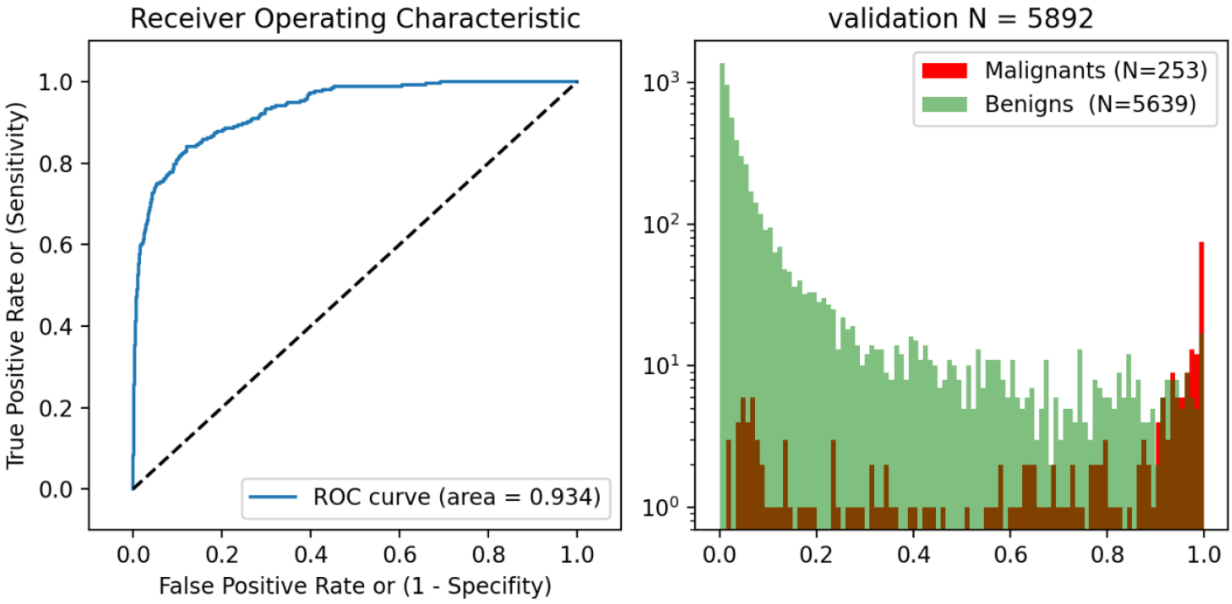


Figure S10: Use of binary cross-entropy as a loss function leads to qualitatively different results: Classification results in the validation set per breast. Left: ROC curve achieving an AUC-ROC of 0.934. Right: Histogram of all samples in the validation set, ordered by predicted risk value by the model and color-coded by true pathology (red = cancer, green = healthy).

Table S1: Summary of performance of different models in the validation set (n=5,892). Performances are measured in area under the receiver operating characteristic curve (AUC-ROC), F1 score, and average precision measured as the area under the precision-recall curve.

Model	**ROC AUC**	**F1 Score**	**Average Precision**
CNN	0.943	0.454	0.661
CNN + Contralateral (3)	0.943	0.58	0.677
CNN + Contralateral (2)	0.943	0.597	0.644
CNN BinaryCrossEntropy	0.934	0.484	0.597
ResNet50 (ImageNet)	0.923	0.541	0.44
CNN 50% Data	0.917	0.373	0.52
CNN + Contralateral (1)	0.915	0.458	0.448
ResNet50	0.91	0.513	0.457

CNN No Data Aug	0.887	0.376	0.413
CNN 10% Data	0.835	0.426	0.375
CNN 10% Data			
No Data Aug	0.802	0.29	0.229
ANN Demographics (no image)	0.602	0	0.076

Table S2: Demographic and clinical information associated with pathology.

Variable	Chi2	p-value
Family History of Breast Cancer	329.98	< 0.001
Ethnicity Hispanic/Latino	0.84	0.36
Ethnicity not Hispanic	8.932	0.002
Ethnicity Unknown	13.528	< 0.001
Race Asian-Far East/Indian Subcontinent	4.121	0.042
Race Black or African American	1.9	0.168
Race Native American-Am Indian/Alaska	0	1
Race Native Hawaiian or Pacific Isl.	0.02	0.886
Race Unknown	13.321	< 0.001
Race White	3.216	0.0729
Ages 0-10	0	1
Ages 10-20	0.751	0.386
Ages 20-30	8.141	0.004
Ages 30-40	1.098	0.294
Ages 40-50	0.061	0.804
Ages 50-60	2.967	0.084
Ages 60-70	0.316	0.574
Ages 70-80	4.606	0.0318
Ages 80-90	0	1
Ages 90-100	0.14	0.7086

Radiology: Artificial Intelligence

High-performance Open-Source Deep Learning Model for Breast Cancer Detection and Localization on MRI

Key Result
An open-source deep learning model developed and trained on the largest breast MRI dataset to date achieved state-of-the-art in performance in breast cancer detection and localization.

Datasets:

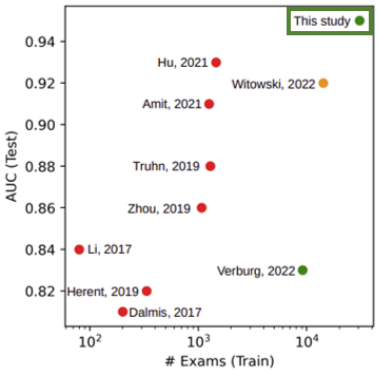
- Validation set: 6,615 sagittal MRI scans
- External dataset: 1,840 axial MRI scans

Methods:

- A 2D convolutional neural network for breast cancer detection and localization was trained on the largest breast MRI dataset to date.
- The model was validated and tested on internal and external data.

Results:

- The model achieved an AUC of 0.95 for detecting breast cancer in the primary site.
- The model generalized well to axial scans and data from a secondary site.



Comparison of related studies to date

Note: Color indicates if studies make code and trained weights openly available (red: not available, orange: upon request, green: no restriction

Hirsch L et al. Published Online: June 25, 2025
https://doi.org/10.1148/ryai.240550

Radiology: Artificial Intelligence