Science Translational Medicine
AAAS

# Supplementary Materials for

## Improving breast cancer diagnostics with deep learning for MRI

Jan Witowski *et al.*

Corresponding author: Jan Witowski, jan.witowski@nyulangone.org; Krzysztof J. Geras, k.j.geras@nyu.edu

*Sci. Transl. Med.* **14**, eabo4802 (2022)
DOI: 10.1126/scitranslmed.abo4802

**The PDF file includes:**

Materials and Methods
Tables S1 to S7
Figs. S1 to S19
References (*62–64*)

**Other Supplementary Material for this manuscript includes the following:**

MDAR Reproducibility Checklist

# Supplementary Materials

## Materials and methods

### Processing TCGA-BRCA dataset

In its original form, TCGA-BRCA data set is not suitable for AI evaluation or training purposes. Specifically, it:

- contains studies where series for left and right breasts are separated,
- contains studies where one or more series are multi-volume,
- contains studies where only one breast is imaged,
- does not provide information on which series are pre- and post-contrast,
- does not provide breast-level labels.

To solve this problem, we established a pipeline for processing the TCGA-BRCA data set for AI purposes. This means that the script we share in our manuscript repository (DOI: 10.5281/zenodo.6989320) is able to take a downloaded data set in its current form and return NIfTI files for pre- and two post-contrast series. For series where two breasts are saved separately, the script merges them into a single volume. For series where both breasts are imaged, but multiple acquisitions are saved in a single volume (multi-volume), the script splits the multi-volume into separate series. Additionally, the script excludes studies that are unilateral. Along with the script, we provide a YAML file which defines a list of all TCGA-BRCA studies for inclusion/exclusion, type of laterality and, potential problems (multi-volumes etc.) as well as series numbers corresponding to pre- and post-contrast T1 fat-sat series. Labels for the TCGA-BRCA data set have been generated using one of the supporting files (clinical_patient_brca.txt), specifically anatomic_neoplasm_subdivision column.

**Table S1: MRI manufacturer and model breakdown for all data sets.** MRI scanners are sorted by number of total cases in the data set, descending. If a cell is empty, that means that the specific data set does not contain any cases acquired on the machine. Device names were acquired by extracting information from DICOM tags Manufacturer and ManufacturerModelName. For Duke University Data set, this information was collected from a spreadsheet provided by data set authors and available at The Cancer Imaging Archive (file named "Clinical and Other Features"). †In 28 examinations, manufacturer information was not provided.

| | | **Number of cases** | | | | |
|---|---|---|---|---|---|---|
| **Manufacturer / Model** | **Magnet** | **NYU Langone** | **Jagiellonian University** | **Duke University** | **TCGA-BRCA** | **Total** |
| Siemens Symphony | 1.5T | 9,638 | | | 2 | 9,640 |
| Siemens Trio Tim | 3T | 8,142 | | 58 | | 8,200 |
| Siemens Skyra | 3T | 1,940 | 1 | 57 | | 1,998 |
| Siemens Espree | 1.5T | 668 | | | 1 | 669 |
| Philips Achieva | 1.5T | 477 | | | 16 | 493 |
| Siemens MAGNETOM Sola | 1.5T | | 392 | | | 392 |
| Siemens Avanto | 1.5T | 132 | 1 | 179 | 3 | 315 |
| GE SIGNA HDx | 1.5T | | | 272 | 8 | 280 |
| GE SIGNA HDxt | 1.5T | | | 248 | 6 | 254 |
| Siemens Verio | 3T | 175 | | | | 175 |
| GE SIGNA HDe | 1.5T | 112 | | | | 112 |
| Siemens Aera | 1.5T | 72 | | | | 72 |
| Siemens Verio Dot | 3T | 65 | | | | 65 |
| Siemens MAGNETOM Vida | 3T | 64 | | | | 64 |
| Hitachi ECHELON | 1.5T | 24 | | | | 24 |
| GE Optima MR450w | 1.5T | | | 98 | | 98 |
| GE SIGNA EXCITE | 1.5T | | | 10 | 85 | 95 |
| Siemens Sonata | 3T | | | | 9 | 9 |
| GE DISCOVERY MR750 | 3T | | | | 1 | 1 |
| Unknown† | - | 28 | | | | 28 |
| | | **21,537** | **394** | **922** | **131** | **22,984** |

## 1. Reader study results

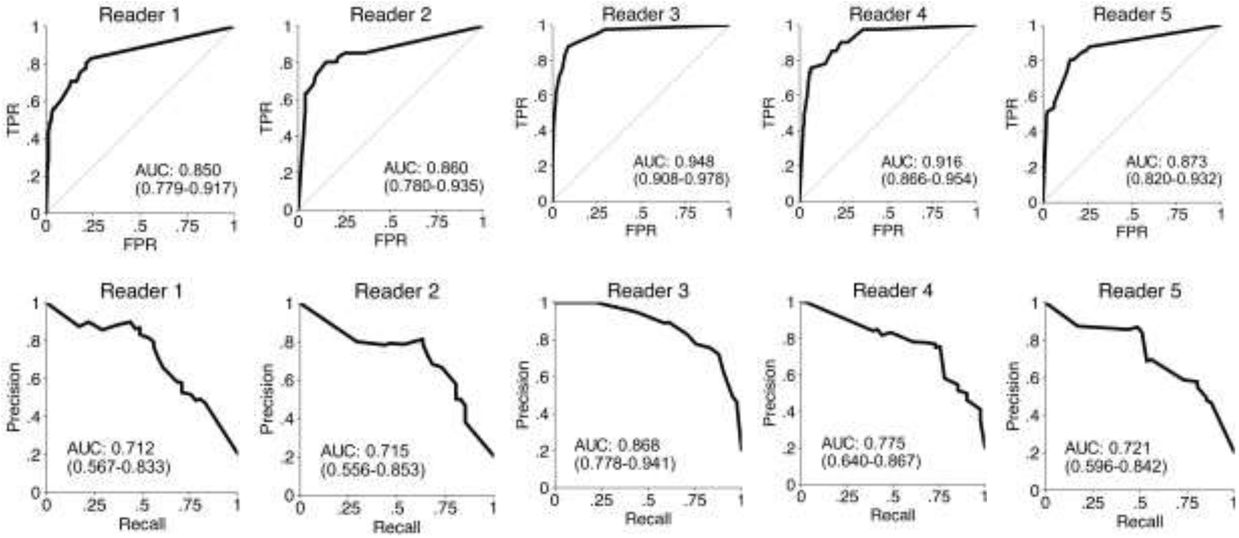


**Figure S1: All receiver operating characteristic (ROC) and precision-recall (PR) curves from the reader study on the NYU Langone subset. Top:** ROC curves for each of the 5 readers. **Bottom:** PR curves for these readers. All ROC and PR curves are non-parametric (empirical) and were generated from predictions of probabilities of malignancy provided by radiologists. All curves are displayed with 95% confidence intervals estimated with bootstrap (N=2,000 replicates). *TPR*, true positive rate; *FPR*, false positive rate.

**Table S2: Results of the reader study on the NYU Langone subset**, reported with 95% confidence intervals estimated with bootstrap (N=2,000). We also report an average performance across all 5 readers ("Avg Reader"). Average reader performance was calculated as a simple mean of metrics for all readers. To calculate sensitivity and specificity for readers, we used BI-RADS 4 as a binarization threshold. That is, studies classified by radiologists as BI-RADS 4 or 5 were considered as positive and BI-RADS 1, 2, 3 as negative. For AI predictions, a decision threshold was selected such that the AI system's sensitivity closely matches average reader sensitivity.

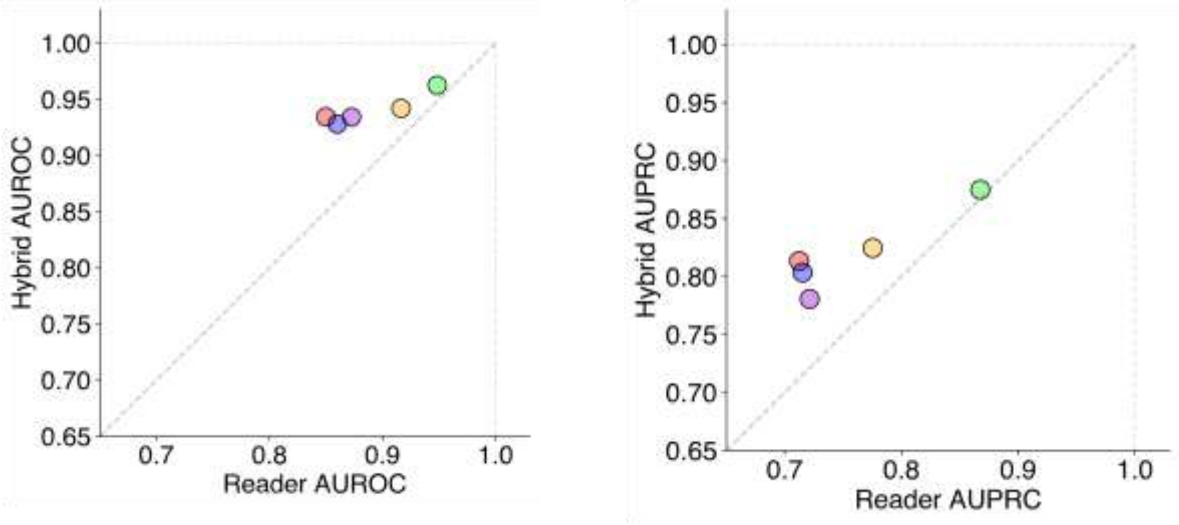
| Reader | AUROC | AUPRC | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|
| Reader 1 | 0.850<br>(0.779-0.917) | 0.712<br>(0.567-0.833) | 0.780<br>(0.659-0.895) | 0.786<br>(0.721-0.849) | 0.485<br>(0.381-0.593) | 0.933<br>(0.883-0.972) |
| Reader 2 | 0.860<br>(0.780-0.935) | 0.715<br>(0.556-0.853) | 0.854<br>(0.737-0.969) | 0.660<br>(0.582-0.745) | 0.393<br>(0.294-0.490) | 0.946<br>(0.904-0.991) |
| Reader 3 | 0.948<br>(0.908-0.978) | 0.868<br>(0.778-0.941) | 0.976<br>(0.913-1.000) | 0.704<br>(0.634-0.764) | 0.460<br>(0.366-0.556) | 0.991<br>(0.971-1.000) |
| Reader 4 | 0.916<br>(0.866-0.954) | 0.775<br>(0.640-0.867) | 0.976<br>(0.917-1.000) | 0.610<br>(0.536-0.679) | 0.392<br>(0.291-0.487) | 0.990<br>(0.965-1.000) |
| Reader 5 | 0.873<br>(0.820-0.932) | 0.721<br>(0.596-0.842) | 0.854<br>(0.750-0.949) | 0.761<br>(0.700-0.822) | 0.479<br>(0.370-0.582) | 0.953<br>(0.915-0.985) |
| Avg Reader | 0.890 | 0.758 | 0.888 | 0.704 | 0.442 | 0.962 |
| AI System | 0.924<br>(0.880-0.962) | 0.784 (0.656-0.887) | 0.897<br>(0.786-0.976) | 0.796<br>(0.728-0.856) | 0.517<br>(0.388-0.629) | 0.969<br>(0.937-0.993) |

**Table S3. Results of the secondary reader study on the Jagiellonian University subset**, reported with 95% confidence intervals estimated with bootstrap (N=2,000). Average reader performance was calculated as a simple mean of metrics for all readers. To calculate sensitivity and specificity for readers, we used BI-RADS 4 as a binarization threshold. That is, studies classified by radiologists as BI-RADS 4 or 5 were considered as positive and BI-RADS 1, 2, 3 as negative. For AI predictions, a decision threshold was selected such that the AI system's sensitivity closely matches average reader sensitivity.

| Reader | AUROC | AUPRC | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|
| Reader 1 | 0.787<br>(0.708-0.859) | 0.667<br>(0.515-0.786) | 0.690<br>(0.548-0.816) | 0.816<br>(0.747-0.873) | 0.508<br>(0.380-0.635) | 0.905<br>(0.853-0.949) |
| Reader 2 | 0.849<br>(0.776-0.916) | 0.699<br>(0.559-0.827) | 0.834<br>(0.719-0.938) | 0.717<br>(0.665-0.791) | 0.449<br>(0.338-0.562) | 0.939<br>(0.894-0.982) |
| Avg Reader | 0.818 | 0.683 | 0.762 | 0.767 | 0.479 | 0.922 |
| AI System | 0.802<br>(0.712-0.881) | 0.558<br>(0.407-0.731) | 0.762<br>(0.622-0.886) | 0.762<br>(0.695-0.828) | 0.469<br>(0.353-0.589) | 0.921<br>(0.870-0.966) |

## 2. Hybrid predictions

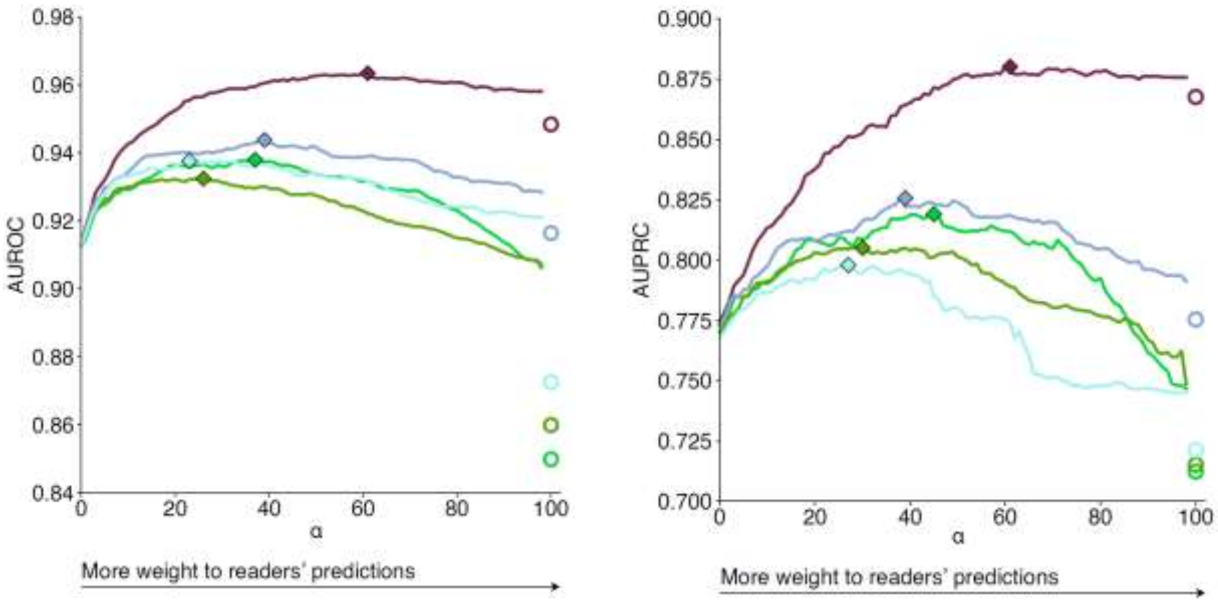
### 2.1. Diagnostic performance of hybrids



(a)

(b)

**Figure S2: Hybrid predictions are stronger than readers' predictions alone.** We demonstrate that an equally weighted average of radiologists and AI model predictions (a "hybrid") on the NYU Langone reader study subset consistently yields a stronger performance in terms of both AUROC **(a)** and AUPRC **(b)**. Each colored circle represents the results of each radiologist and their hybrid. If the circle is above the diagonal, it means that the hybrid had better results than the reader. Performance increase is more marked in radiologists who performed slightly worse. However, even for the strongest reader's predictions the results were higher when averaged with AI model.



(a) (b)

**Figure S3: Performance of a hybrid model, as a function of *α* ∈ (0, 99]%**. Plots show how (**a)**, AUROC and **(b)**, AUPRC change when the α multiplier changes. At *α* = 0%, the hybrid performance is equal to the model only performance. At *α* = 100%, the hybrid performance is equal to the reader only performance (here plotted as an empty circle on the far right of the figures). Results demonstrate that utilizing AI predictions even at low weights (high *α*) substantially improves performance. Each line represents performance for a different reader. Diamond-shaped points represent maximum performance for each metric and reader.

## 2.2. Inter-reader variability analysis

We analyzed whether hybrids improve inter-reader diagnostic variability. First, we calculated Fleiss' κ.

**Table S4: Inter-reader variability analysis results.** To calculate κ, we binarized the predictions (made by both AI and readers on the NYU Langone reader study cases) into positive and negative classes. We set the binarization threshold at 2% [probability of malignancy]. We selected this value, because 2% is a cut-off for BI-RADS 4 category, according to the clinical guidelines. Readers were asked to follow these guidelines, and they had to assign a probability of >2% to any BI-RADS 4 or 5 prediction.

| Category | Hybrid type / Description | Fleiss' κ (95% CI) | Avg AUROC | AUROC s$^2$ |
|---|---|---|---|---|
| **No hybrid** | **Interreader variability between readers** | **0.5567 (0.50-0.63)** | **0.890** | **1.7 × 10$^{-3}$** |
| **AI hybrid** | **Unweighted average of reader prediction and AI prediction** | **0.77 (0.72-0.82)** | **0.939** | **2.1 × 10$^{-4}$** |
| Baseline 1 | Unweighted average of reader prediction and fixed median AI prediction equal to 3.71% | 0.5608 (0.48-0.63) | 0.890 | 1.7 × 10$^{-3}$ |
| Baseline 2 | Unweighted average of reader prediction and fixed mean AI prediction equal to 16.82% | 1.0 (1.0-1.0) | 0.890 | 1.7 × 10$^{-3}$ |

As shown in table S3, we used two trivial baselines to validate the results. For Baseline 1, the change is small. For Baseline 2, we reached a perfect agreement. However, although averaging averaging radiologists' predictions with a fixed scalar does not impact the AUC, averaging with AI's predictions improves the AUC noticeably. As seen in this experiment, the change in Fleiss' κ depends on the value that we use to average readers' predictions. This is visualized in figure S4. The changes in agreement were small until we reached a point when all binarized predictions are positive and the agreement trivially reaches 1.0.

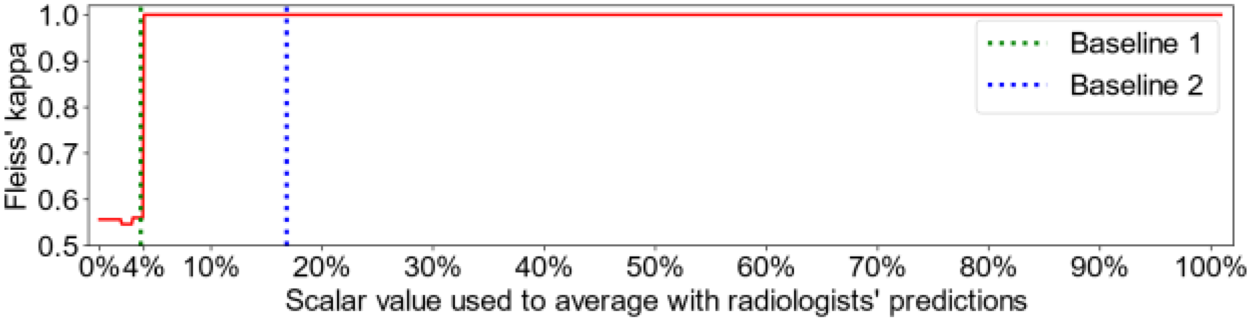
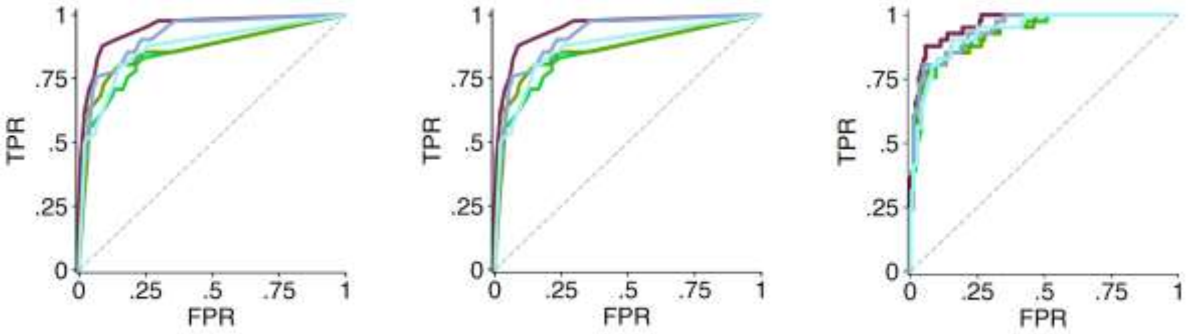


**Figure S4: Changes in Fleiss’ kappa for hybrids with a fixed scalar, depending on the scalar value.**

Next, as an alternative method of analysis, we visualize ROC curves for all approaches and compute variance ($s^2$) in AUROCs within different scenarios. Visualizing ROC curves as a method of evaluating inter-reader variability has been used previously in literature, for example in Winkel et al. *(62)* or Obuchowski et al. *(63, 64)*.



(a) Radiologists alone

(b) Hybrids with a fixed value (identical for Baselines 1 and 2)

(c) AI hybrids

**Figure S5: ROC curves for all readers and hybrids.**

As averaging with a constant does not change ROC curves, variance of AUROCs within the hybrids do not change for Baselines 1 and 2 ($s^2 = 1.7 \times 10^{-3}$). Within AI hybrids, variance is an order of magnitude lower ($s^2 = 2.1 \times 10^{-4}$) than within radiologists on their own.

## 3. Subgroup performance

### 3.1. Full numerical results for all subgroups

**Table S5: Subgroup performance**. Reported values are *n* (95% confidence intervals). Confidence intervals were calculated with a bootstrap (2,000 replicates). *PPV*, positive predictive value; *NPV*, negative predictive value. As there were no malignant examples in BI-RADS 1 and 2 categories in our test set, AUROC would not be defined for those groups. BI-RADS 1 and 2 were combined with BI-RADS 3 to generate the results. For AI predictions, a decision threshold was selected such that the AI system's sensitivity closely matches average reader sensitivity.

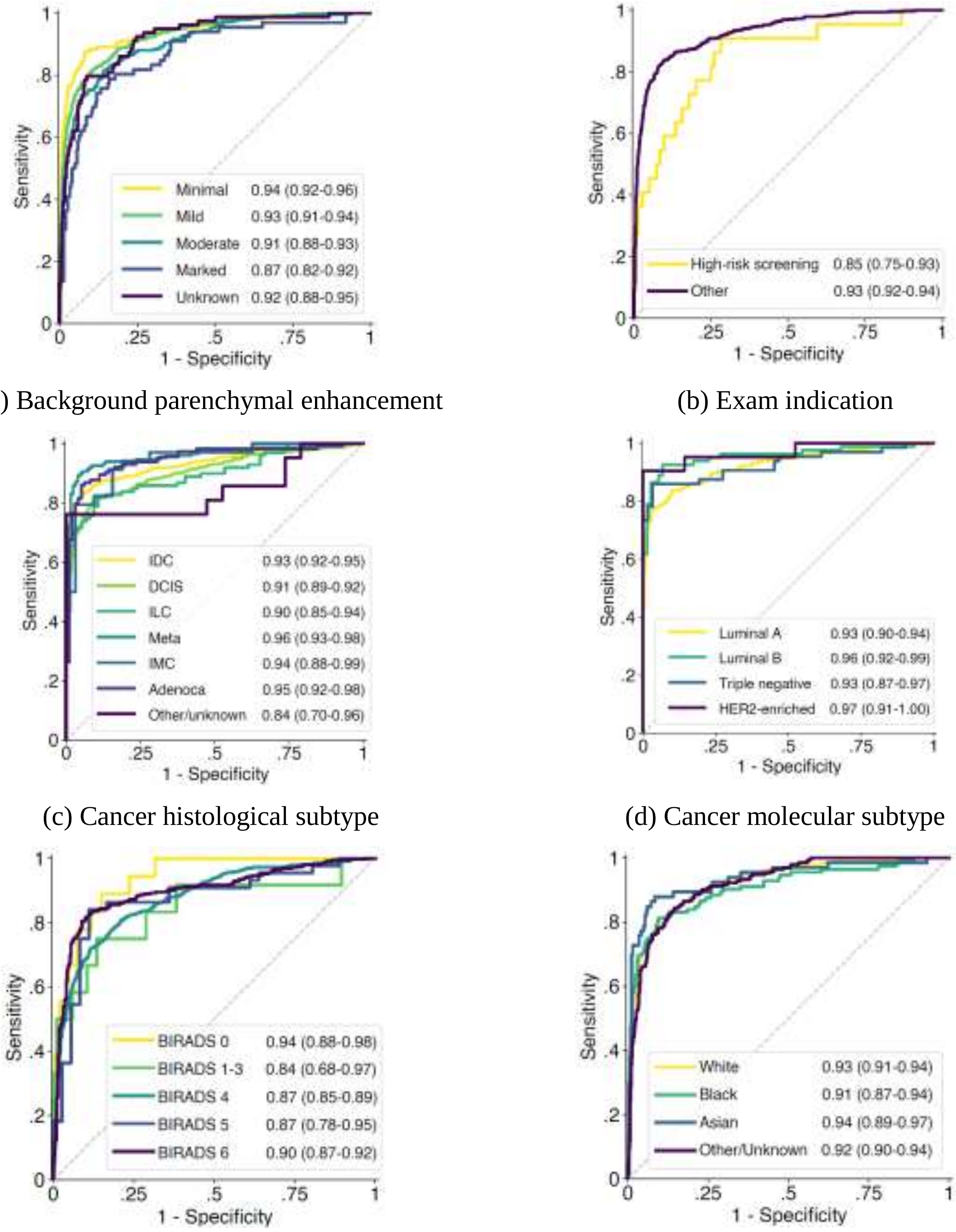
| Group | *n* | AUROC | AUPRC | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|---|
| ***BI-RADS risk assessment category*** | | | | | | | |
| **BIRADS 1/2/3** | 2,307 | 0.84 (0.68-0.97) | 0.09 (0.01-0.27) | 0.75 (0.46-1.00) | 0.83 (0.82-0.84) | 0.01 (0.00-0.02) | 1.00 (1.00-1.00) |
| **BIRADS 4** | 956 | 0.87 (0.85-0.89) | 0.72 (0.67-0.76) | 0.88 (0.85-0.91) | 0.61 (0.58-0.63) | 0.42 (0.39-0.45) | 0.94 (0.93-0.96) |
| **BIRADS 5** | 40 | 0.87 (0.78-0.95) | 0.89 (0.78-0.97) | 0.91 (0.82-0.98) | 0.47 (0.31-0.64) | 0.68 (0.56-0.79) | 0.81 (0.64-0.95) |
| **BIRADS 6** | 385 | 0.90 (0.87-0.92) | 0.88 (0.85-0.92) | 0.90 (0.86-0.93) | 0.67 (0.62-0.72) | 0.70 (0.65-0.74) | 0.88 (0.85-0.92) |
| **BIRADS 0** | 102 | 0.94 (0.88-0.98) | 0.61 (0.35-0.84) | 0.94 (0.82-1.00) | 0.75 (0.68-0.81) | 0.27 (0.16-0.38) | 0.99 (0.98-1.00) |
| **unknown** | 146 | 0.92 (0.87-0.96) | 0.75 (0.61-0.85) | 0.89 (0.81-0.98) | 0.79 (0.74-0.84) | 0.45 (0.35-0.55) | 0.97 (0.95-0.99) |
| ***Patient age at the time of examination*** | | | | | | | |
| **Age <40** | 399 | 0.91 (0.87-0.94) | 0.65 (0.53-0.76) | 0.89 (0.81-0.95) | 0.74 (0.70-0.77) | 0.27 (0.22-0.33) | 0.98 (0.97-0.99) |
| **Age <50** | 1,294 | 0.91 (0.89-0.93) | 0.67 (0.61-0.73) | 0.89 (0.85-0.92) | 0.73 (0.71-0.75) | 0.30 (0.27-0.33) | 0.98 (0.97-0.99) |
| **Age ≥50** | 2,642 | 0.93 (0.92-0.94) | 0.74 (0.71-0.78) | 0.89 (0.86-0.91) | 0.79 (0.78-0.80) | 0.37 (0.35-0.40) | 0.98 (0.98-0.99) |
| ***Breast cancer histological subtype*** | | | | | | | |
| **DCIS** | 570 | 0.91 (0.89-0.92) | 0.93 (0.92-0.95) | 0.89 (0.86-0.91) | 0.68 (0.64-0.72) | 0.76 (0.72-0.79) | 0.84 (0.80-0.87) |
| **IDC** | 523 | 0.93 (0.92-0.95) | 0.95 (0.94-0.96) | 0.92 (0.90-0.94) | 0.64 (0.60-0.68) | 0.74 (0.71-0.78) | 0.88 (0.84-0.91) |
| **Meta** | 138 | 0.96 (0.93-0.98) | 0.97 (0.94-0.99) | 0.96 (0.93-0.99) | 0.58 (0.49-0.66) | 0.72 (0.66-0.79) | 0.93 (0.86-0.98) |
| **Adenoca** | 106 | 0.95 (0.92-0.98) | 0.96 (0.93-0.98) | 0.98 (0.95-1.00) | 0.55 (0.45-0.65) | 0.72 (0.65-0.79) | 0.96 (0.91-1.00) |
| **ILC** | 87 | 0.90 (0.85-0.94) | 0.94 (0.90-0.96) | 0.86 (0.79-0.92) | 0.63 (0.52-0.74) | 0.75 (0.67-0.83) | 0.77 (0.67-0.87) |
| **IMC** | 33 | 0.94 (0.88-0.99) | 0.95 (0.89-0.99) | 0.94 (0.85-1.00) | 0.75 (0.59-0.90) | 0.80 (0.67-0.92) | 0.92 (0.81-1.00) |
| **Other/unknown** | 20 | 0.84 (0.70-0.96) | 0.91 (0.79-0.98) | 0.81 (0.63-0.95) | 0.53 (0.29-0.75) | 0.65 (0.46-0.83) | 0.71 (0.45-0.93) |
| ***Breast cancer molecular subtype*** | | | | | | | |
| **Luminal A** | 326 | 0.93 (0.90-0.94) | 0.95 (0.93-0.96) | 0.91 (0.88-0.94) | 0.68 (0.63-0.73) | 0.77 (0.73-0.81) | 0.86 (0.82-0.90) |
| **Luminal B** | 78 | 0.96 (0.92-0.99) | 0.97 (0.94-0.99) | 0.96 (0.91-1.00) | 0.67 (0.56-0.78) | 0.76 (0.67-0.84) | 0.94 (0.87-1.00) |
| **Triple negative** | 63 | 0.93 (0.87-0.97) | 0.95 (0.91-0.98) | 0.91 (0.82-0.97) | 0.71 (0.59-0.83) | 0.76 (0.66-0.86) | 0.88 (0.78-0.96) |
| **HER2-enriched** | 21 | 0.97 (0.91-1.00) | 0.98 (0.93-1.00) | 0.95 (0.85-1.00) | 0.67 (0.45-0.85) | 0.74 (0.57-0.89) | 0.93 (0.79-1.00) |
| ***Background parenchymal enhancement*** | | | | | | | |
| **Minimal** | 884 | 0.94 (0.92-0.96) | 0.78 (0.71-0.84) | 0.89 (0.84-0.94) | 0.85 (0.83-0.86) | 0.35 (0.30-0.39) | 0.99 (0.98-0.99) |
| **Mild** | 1,614 | 0.93 (0.91-0.94) | 0.72 (0.68-0.77) | 0.89 (0.86-0.92) | 0.79 (0.77-0.80) | 0.37 (0.34-0.40) | 0.98 (0.97-0.99) |
| **Moderate** | 884 | 0.91 (0.88-0.93) | 0.71 (0.65-0.77) | 0.88 (0.84-0.92) | 0.68 (0.66-0.71) | 0.32 (0.28-0.35) | 0.97 (0.96-0.98) |
| **Marked** | 184 | 0.87 (0.82-0.92) | 0.66 (0.54-0.77) | 0.92 (0.85-0.98) | 0.60 (0.54-0.65) | 0.33 (0.27-0.40) | 0.97 (0.95-0.99) |
| **Unknown** | 370 | 0.92 (0.88-0.95) | 0.67 (0.56-0.77) | 0.86 (0.78-0.94) | 0.79 (0.76-0.82) | 0.33 (0.27-0.39) | 0.98 (0.97-0.99) |
| ***Patient's race*** | | | | | | | |
| **White** | 2,738 | 0.93 (0.91-0.94) | 0.72 (0.68-0.75) | 0.88 (0.85-0.90) | 0.78 (0.77-0.80) | 0.32 (0.30-0.34) | 0.98 (0.98-0.99) |
| **Black** | 244 | 0.91 (0.87-0.94) | 0.82 (0.75-0.89) | 0.88 (0.82-0.94) | 0.73 (0.69-0.78) | 0.49 (0.43-0.57) | 0.95 (0.93-0.98) |
| **Asian** | 164 | 0.94 (0.89-0.97) | 0.83 (0.71-0.93) | 0.92 (0.86-0.98) | 0.71 (0.65-0.76) | 0.44 (0.36-0.52) | 0.97 (0.95-0.99) |
| **Other/Unknown** | 790 | 0.92 (0.90-0.94) | 0.65 (0.57-0.73) | 0.91 (0.86-0.94) | 0.75 (0.73-0.77) | 0.33 (0.29-0.37) | 0.98 (0.98-0.99) |
| ***MRI scanner magnet strength*** | | | | | | | |
| **1.5T** | 2,102 | 0.93 (0.91-0.94) | 0.67 (0.62-0.72) | 0.84 (0.80-0.87) | 0.86 (0.85-0.87) | 0.36 (0.32-0.39) | 0.98 (0.98-0.99) |
| **3T** | 1,834 | 0.92 (0.91-0.93) | 0.75 (0.72-0.79) | 0.92 (0.90-0.94) | 0.66 (0.65-0.68) | 0.34 (0.32-0.36) | 0.98 (0.97-0.98) |

### 3.2. Model performance per exam indication

**Table S6: AI system's performance on the NYU Langone test set per exam indication.** Indications were automatically extracted from radiology reports using regular expressions. The list of all regular expressions is available in the report explaining creation and composition of the dataset at https://cs.nyu.edu/~kgeras/reports/MRI_datav1.0.pdf. The extraction was possible for 1,883 (47.8%) examinations. In the remaining 2,053 exams, the script was not able to accurately extract the exam indication due to the unstructured and narrative reporting. Results for all metrics are presented with 95% confidence intervals (estimated with bootstrap; N=2,000 replicates). To calculate sensitivity and specificity from AI predictions, a decision threshold was selected such that the overall AI system's sensitivity closely matches average reader sensitivity.

| Exam indication | n | AUC ROC | AUC PR | Sensitivity | Specificity |
|---|---|---|---|---|---|
| High-risk screening | 903 | 0.85 (0.75-0.93) | 0.24 (0.07-0.43) | 0.73 (0.54-0.91) | 0.81 (0.79-0.83) |
| Other indication | 980 | 0.93 (0.92-0.94) | 0.84 (0.81-0.88) | 0.91 (0.88-0.93) | 0.73 (0.71-0.75) |
| Extent of disease | 481 | 0.91 (0.89-0.93) | 0.90 (0.86-0.93) | 0.91 (0.88-0.94) | 0.63 (0.58-0.67) |
| Follow-up/surveillance | 286 | 0.91 (0.84-0.97) | 0.02 (0.01-0.07) | 1.00 (1.00-1.00) | 0.78 (0.74-0.81) |
| Further evaluation | 110 | 0.87 (0.76-0.98) | 0.49 (0.05-0.87) | 0.86 (0.50-1.00) | 0.77 (0.71-0.82) |
| Other | 103 | 0.93 (0.84-0.99) | 0.75 (0.49-0.94) | 0.90 (0.73-1.00) | 0.81 (0.75-0.87) |
| Unknown | 2,053 | 0.91 (0.90-0.93) | 0.63 (0.57-0.68) | 0.87 (0.84-0.90) | 0.77 (0.76-0.78) |
| Total | 3,936 | 0.92 (0.92-0.93) | 0.72 (0.69-0.75) | 0.90 (0.79-0.98) | 0.80 (0.73-0.86) |

### 3.3. ROC curves for key subgroups



(a) Background parenchymal enhancement

(b) Exam indication

(c) Cancer histological subtype

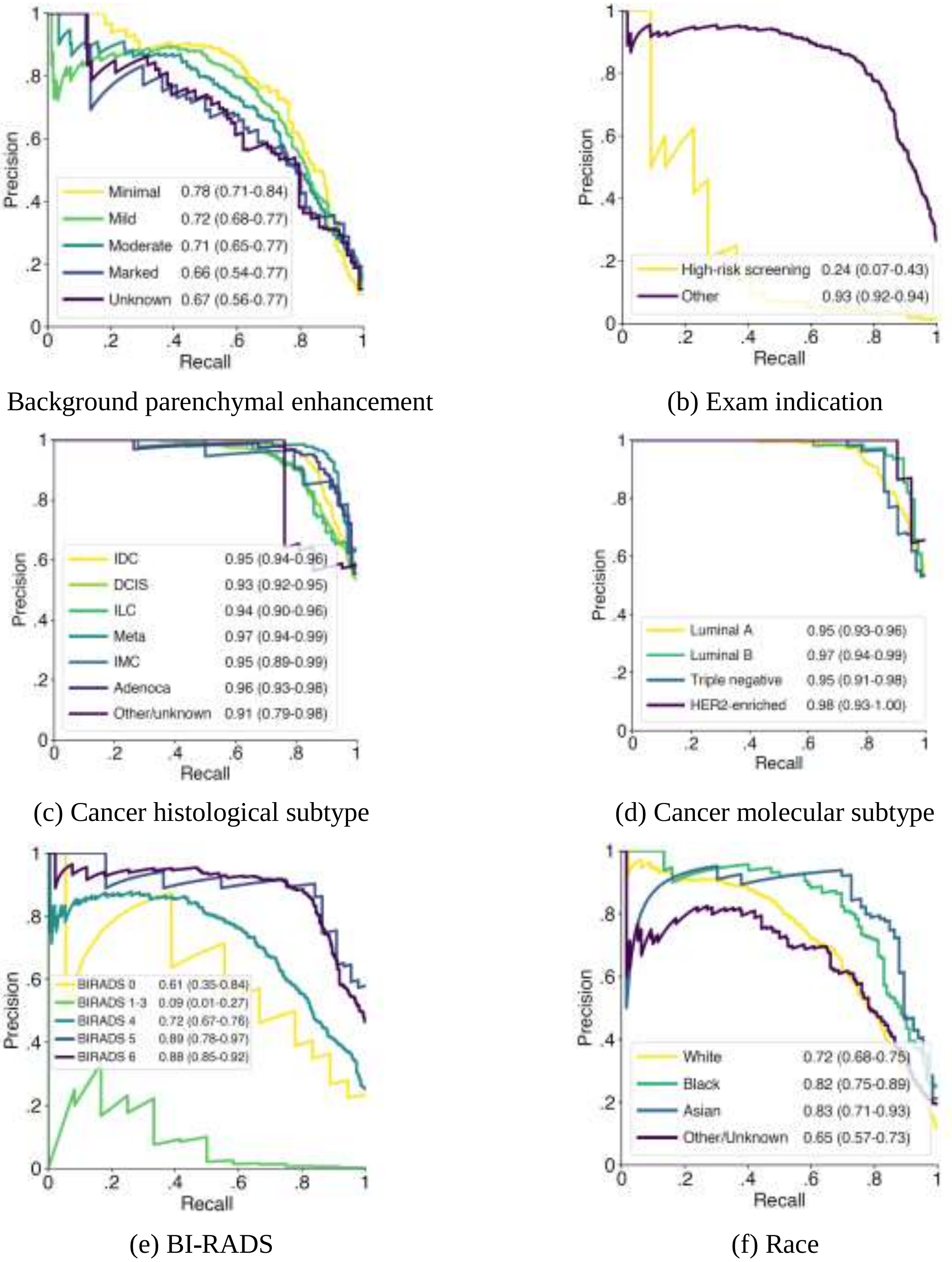
(d) Cancer molecular subtype

(e) BI-RADS

(f) Race

**Figure S6: Empirical ROC curves for subgroups** per background parenchymal enhancement category **(a)**, exam indication **(b)**, cancer histological subtype **(c)**, cancer molecular subtype **(d)**, BI-RADS category **(e)**, race **(f)**. As there were no malignant examples in BI-RADS 1 and 2 categories in our test set, AUC ROC would not be defined for those groups. BI-RADS 1 and 2 were combined with BI-RADS 3 to generate curves and calculate AUC ROC.

### 3.4. Precision-recall curves for key subgroups



(a) Background parenchymal enhancement

(b) Exam indication

(c) Cancer histological subtype

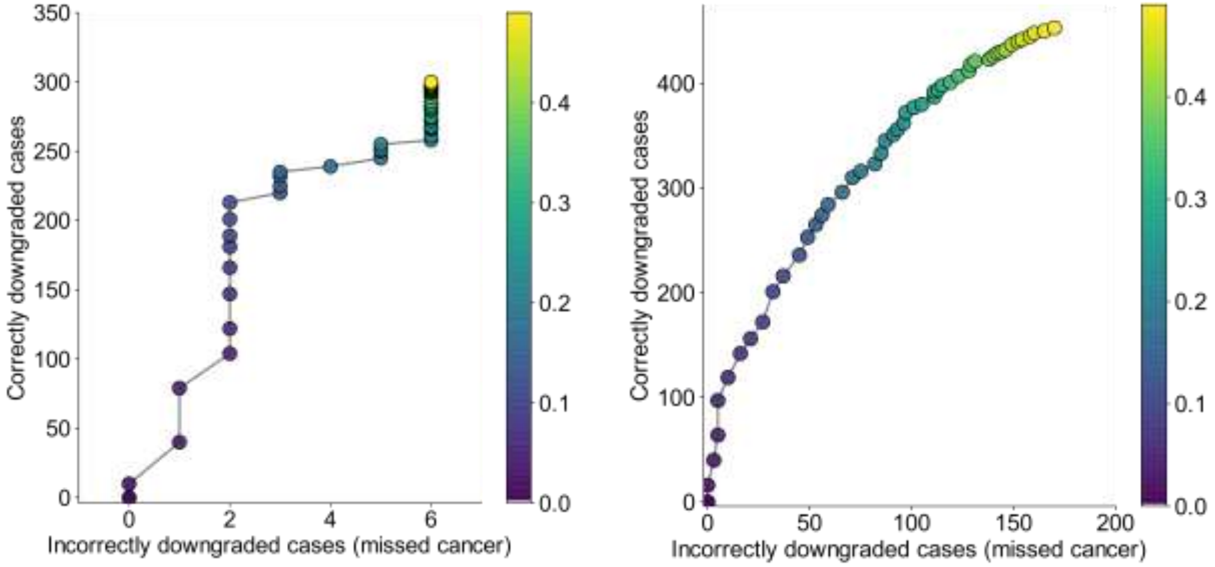
(d) Cancer molecular subtype

(e) BI-RADS

(f) Race

**Figure S7: Empirical precision-recall curves for subgroups** per background parenchymal enhancement category **(a)**, exam indication **(b)**, cancer histological subtype **(c)**, cancer molecular subtype **(d)**, BI-RADS category **(e)**, race **(f)**. As there were no malignant examples in BI-RADS 1 and 2 categories in our test set, AUC PR would not be defined for those groups. BI-RADS 1 and 2 were combined with BI-RADS 3 to generate AUC PR and curves.

## 4. BI-RADS downgrading



(a) BI-RADS 3

(b) BI-RADS 4

**Figure S8: Trade-off in missed cancers versus correctly avoided biopsies when using only AI system to decide on management. (a)** shows the trade-off when using only AI system's predictions to decide whether the patient should return for a 6-month follow-up or not in cases with BI-RADS 3 findings. "Correctly downgraded" patients would return to a regular screening, while "missed cancers" prevent the opportunity to detect cancer if it would be imaged again in 6 months. **(b)** shows the trade-off in BI-RADS 4 cases. "Correctly downgraded" cases from BI-RADS 4 to BI-RADS 3 represent patients who would avoid an unnecessary biopsy (by downgrading BI-RADS 4 lesion to BI-RADS 3), while "missed cancers" are situations where patients do have breast cancer but would not be biopsied because of the AI system's predictions. Both **a, b** do not take into consideration patient's and physician's preferences and do not weigh the trade-off items (e.g. one missed cancer case is more important than one avoided biopsy). They also ignore the potential effect of physician ultimately making a decision based on their own knowledge supported by the AI system. **a, b** show the trade-off at different operating points. Operating points are color-coded by increasing binarization thresholds (warmer colors are higher thresholds).

## 7. Breast-level labels

**Table S7: Breast-level breakdown of labels in the NYU Langone data set**. Malignant and benign labels are not mutually exclusive. A patient might have both a malignant and a benign change in the same breast.

| | **Training set** | **Validation set** | **Test set** | **Total** |
|---|---|---|---|---|
| **Left benign** | 2,117 | 518 | 715 | 3,350 |
| **Right benign** | 2,111 | 477 | 705 | 3,293 |
| **Left malignant** | 1,278 | 326 | 478 | 2,082 |
| **Right malignant** | 1,211 | 293 | 427 | 1,931 |
| **Left negative** | 11,539 | 2,747 | 2,992 | 17,278 |
| **Right negative** | 11,617 | 2,798 | 3,060 | 17,475 |

## 8. Error analysis

Below are several examinations selected from the NYU Langone reader study subset that show situations where our AI system is compared with radiologists' predictions. We present probabilities of malignancy (POMs) for all readers and the AI system with a short case description.

### 8.1. Correctly identified cancers

*Case 1.* In the following imaging exam, all five radiologists gave it a very high probability of malignancy in the right breast (one BI-RADS 4C, four BI-RADS 5). The AI system also correctly identified the malignancy and gave the examination a 97% probability of cancer in the right breast. Oone radiologist found a suspicious lesion in the left breast. Based on the patient's history, that lesion was also identified by the radiologist originally interpreting the exam. Upon biopsy, the lesion was found to be benign.

| | Left breast POM | Right breast POM |
|---|---|---|
| **Reader 1** | 0 | 100 |
| **Reader 2** | 0 | 99 |
| **Reader 3** | 0 | 98 |
| **Reader 4** | 10 | 99 |
| **Reader 5** | 0 | 90 |
| **AI system** | **2** | **97** |

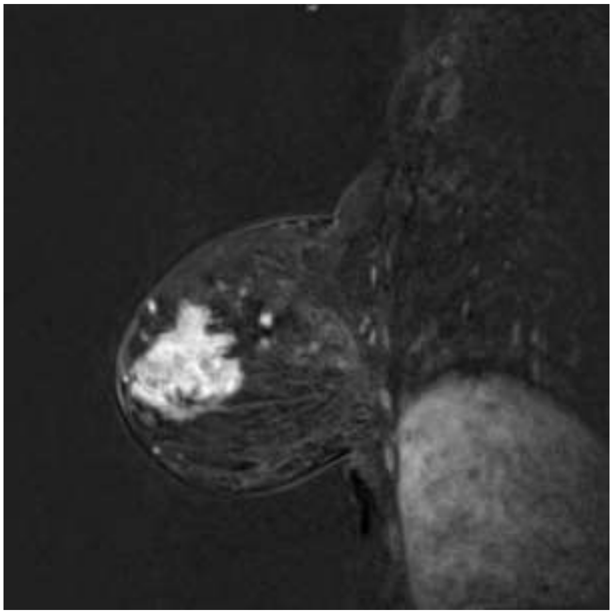


**Figure S9:** Sagittal view of the adenocarcinoma in the right breast. There are multiple irregular heterogeneously enhancing masses suspicious for satellite lesions.

*Case 2.* Here, only three out of five readers found any lesions in the examination. Out of the three who did, only one gave it a high probability of malignancy (reader 5, 30%). The suspicious lesion was later confirmed to be malignant. Our AI model correctly predicted the malignancy, giving a 39% probability in the left breast, and 0% POM in the right breast.

| | **Left breast POM** | **Right breast POM** |
|---|---|---|
| **Reader 1** | 2 | 0 |
| **Reader 2** | 0 | 0 |
| **Reader 3** | 5 | 0 |
| **Reader 4** | 0 | 0 |
| **Reader 5** | 30 | 2 |
| **AI system** | **39** | **0** |

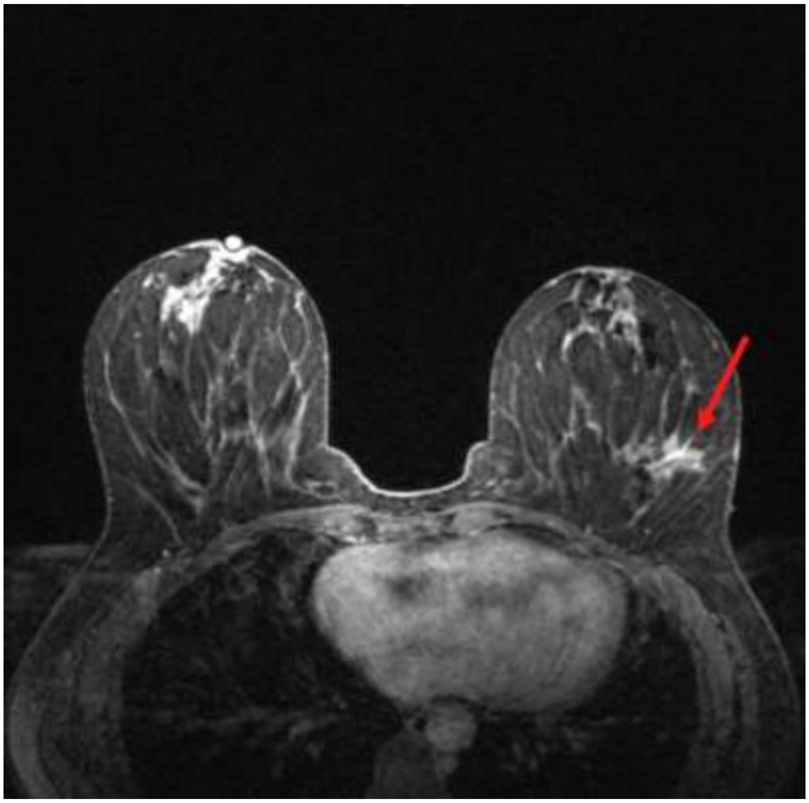


**Figure S10:** From the radiology report: "A 2 cm biopsy tract [red arrow] is present in the left outer breast at 3:00 posterior depth, associated with mild inflammatory changes and a biopsy clip in its medial aspect, concordant with the site of biopsy-proven malignancy."

*Case 3.* This examination was performed in the diagnostic process of evaluating bloody left nipple discharge which demonstrated atypical cells. Although there were no suspicious findings in the left breast, all radiologists agreed that the enhancement in the right breast was highly suspicious. This prediction was matched by AI output. The lesion was found to be malignant.

| | **Left breast POM** | **Right breast POM** |
|---|---|---|
| **Reader 1** | 0 | 85 |
| **Reader 2** | 0 | 50 |
| **Reader 3** | 0 | 40 |
| **Reader 4** | 0 | 85 |
| **Reader 5** | 0 | 95 |
| **AI system** | **0** | **26** |

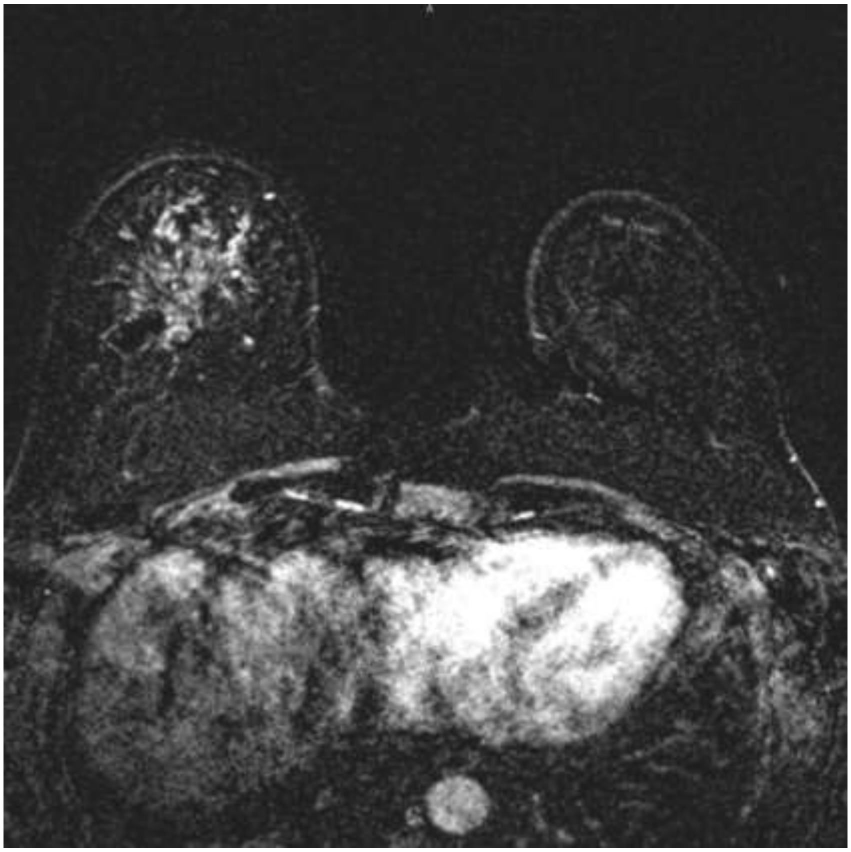
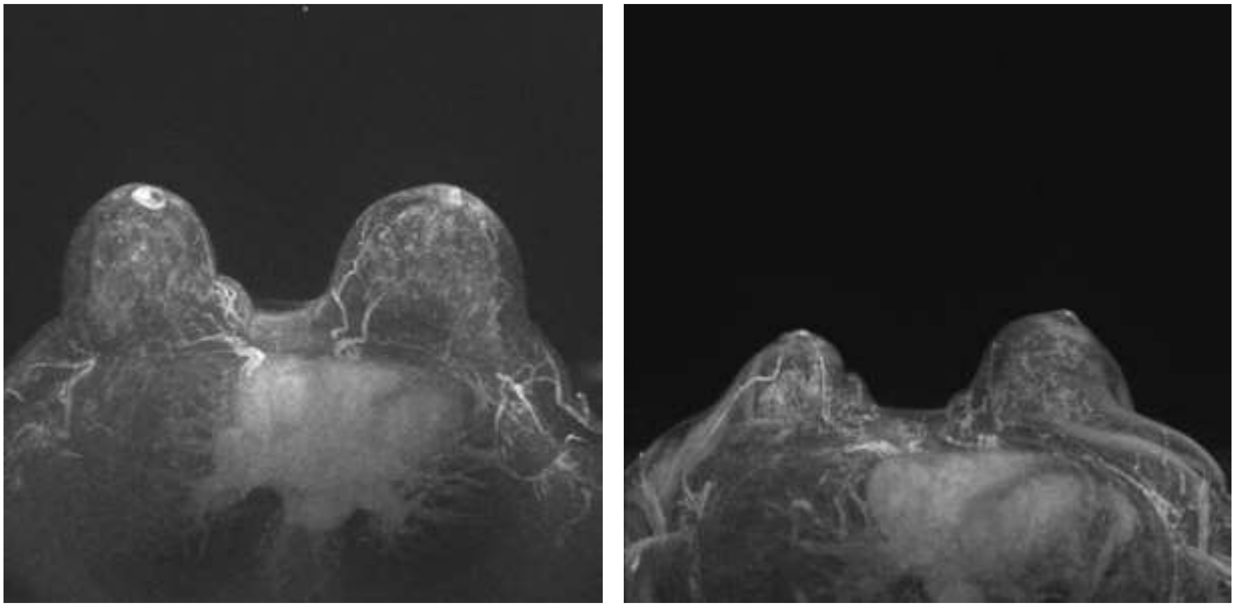


**Figure S11:** A slide from T1-weighted subtraction series with visible suspicious lesion in the right breast. From the radiology report: “Extensive nonmass enhancement in the inferior right breast with questionable mild architectural distortion”.

### 9.2. Correctly identified negative examinations

*Case 4.* and *Case 5.* below are two sample imaging exams where all radiologists agreed that there are no suspicious lesions in the exam, and our AI system gave very low probabilities of malignancy as well. Predictions in the table below were appropriate for both Case 4 and Case 5.

| | **Left breast POM** | **Right breast POM** |
|---|---|---|
| **Reader 1** | 0 | 0 |
| **Reader 2** | 0 | 0 |
| **Reader 3** | 0 | 0 |
| **Reader 4** | 0 | 0 |
| **Reader 5** | 0 | 0 |
| **AI system** | **1** | **1** |



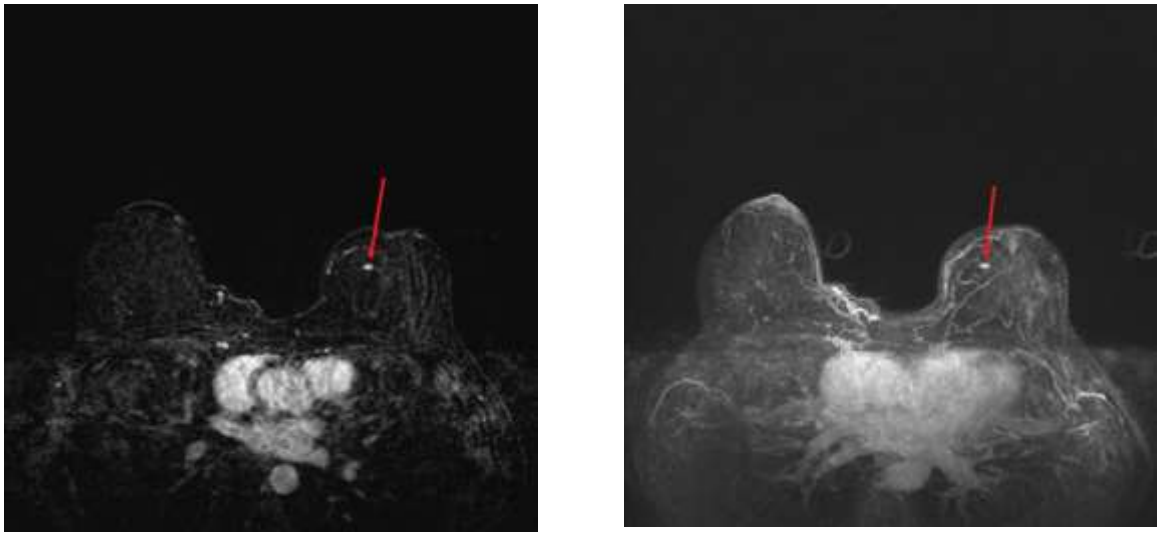
(a) (b)

**Figure S12:** Maximum intensity projection images for Case 4 **(a)** and Case 5 **(b)**.

### 9.3. Opportunity to avoid biopsies

*Case 6.* shows an examination where all radiologists would biopsy the lesion in the left breast. One reader classified this exam as BI-RADS 4A, three readers as BI-RADS 4B and one as BI-RADS 4C. Looking into patient history, the suspicious lesion in the left breast was indeed biopsied and yielded a benign result. Our AI system correctly outputted a low POM. This raises questions whether radiologists would be more likely to revisit their first diagnosis when provided with AI output.

| | **Left breast POM** | **Right breast POM** |
|---|---|---|
| **Reader 1** | 15 | 0 |
| **Reader 2** | 10 | 0 |
| **Reader 3** | 50 | 0 |
| **Reader 4** | 2 | 0 |
| **Reader 5** | 10 | 0 |
| **AI system** | **1** | **0** |



(a)

(b)

**Figure S13:** Axial view of a subtraction image **(a)** and maximum intensity projection **(b)** of the Case 6 with visible lesion [red arrow] that was interpreted as suspicious by radiologists, but turned out to be benign after a core biopsy. Diagnosis from the pathology report said: “benign breast tissue with dense stroma, focal sclerosing adenosis, benign adipose tissue”.

## 9.4. Missed cancers

In this section we will investigate a few situations where patients were diagnosed with breast cancer, but our AI system output suggested low or very low probability of malignancy. We identified two examinations where our system dramatically underestimated the POM (Case 7. and Case 8.). We also present two more imaging exams where the POM was higher, but still lower than preferable.

*Case 7.* Here, all radiologists agreed that right breast had a high POM with BI-RADS 4C/5. This was a situation where our model failed completely, yielding only 1% POM for the right breast. This study was performed to evaluate the extent of disease.

| | **Left breast POM** | **Right breast POM** |
|---|---|---|
| **Reader 1** | 0 | 75 |
| **Reader 2** | 0 | 99 |
| **Reader 3** | 0 | 80 |
| **Reader 4** | 0 | 95 |
| **Reader 5** | 0 | 70 |
| **AI system** | **2** | **1** |

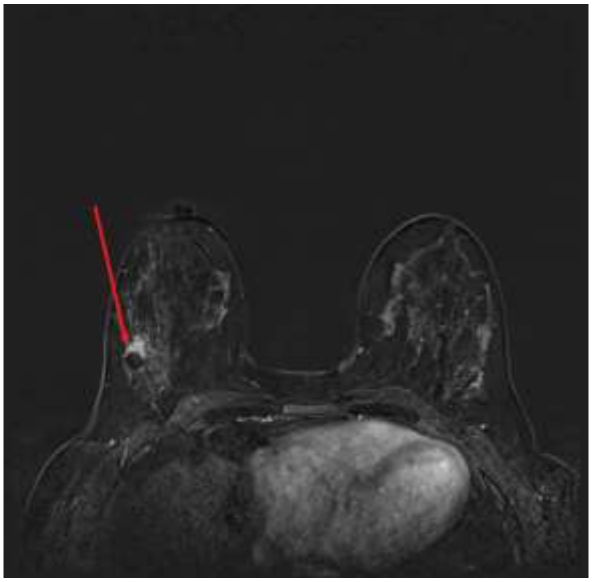


**Figure S14:** From radiology report: “2.5 x 1.6 x 2.2 cm enhancing mass containing susceptibility artifact from biopsy marker clip in the right breast 9:00 axis, 8 cm from the nipple, biopsy proven malignancy”.

*Case 8*. Similarly to Case 7, all radiologists agreed that the left breast has a relatively high POM. This exam was performed for extent of disease.

| | **Left breast POM** | **Right breast POM** |
|---|---|---|
| **Reader 1** | 20 | 0 |
| **Reader 2** | 50 | 0 |
| **Reader 3** | 20 | 0 |
| **Reader 4** | 40 | 0 |
| **Reader 5** | 75 | 0 |
| **AI system** | **2** | **1** |

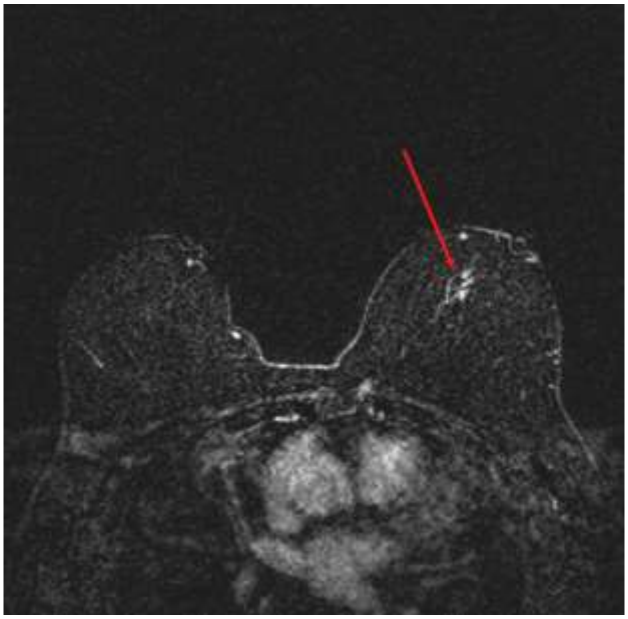
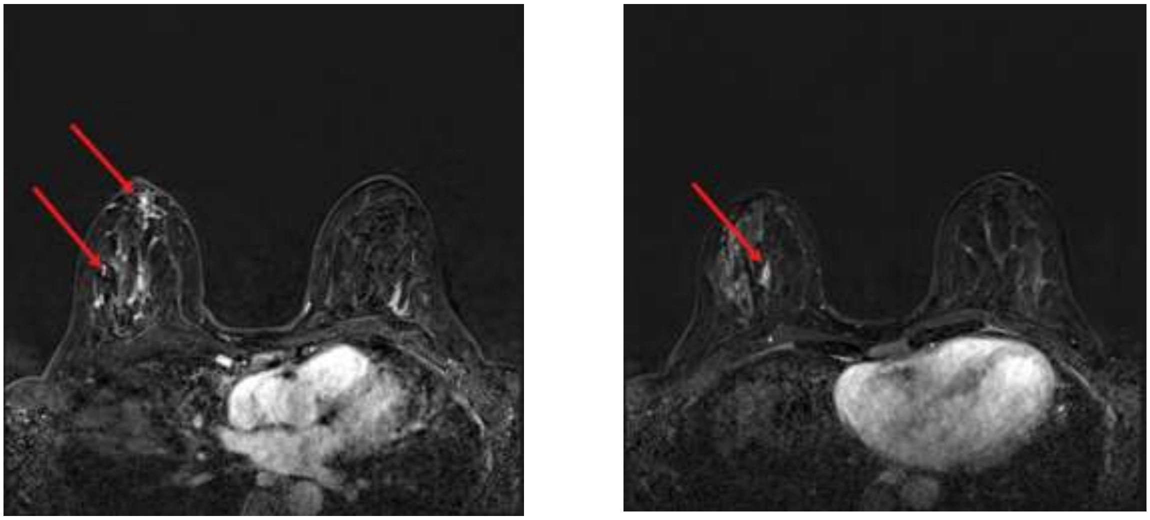


**Figure S15:** From radiology report: “Susceptibility artifact from a metallic clip is seen in the left breast mid depth with surrounding non mass enhancement collectively measuring 1.5 x 2.6 cm consistent with biopsy proven malignancy”. Suspicious area marked with the red arrow.

*Case 9.* In this case, there were multiple suspicious findings in the right breast. Both radiologists and our AI system identified higher-than-average POM. However, the AI's POM was lower than expected from a highly accurate system. On the other hand, this POM was on par with some radiologists' predictions. Reader 1 would not recommend a biopsy, and Reader 4 gave a 10% POM for the right breast, the same value that the AI system did.

| | **Left breast POM** | **Right breast POM** |
|---|---|---|
| **Reader 1** | 0 | 1 |
| **Reader 2** | 0 | 30 |
| **Reader 3** | 0 | 50 |
| **Reader 4** | 0 | 10 |
| **Reader 5** | 0 | 85 |
| **AI system** | **2** | **10** |



(a)

(b)

**Figure S16:** Axial subtraction images of multiple suspicious findings identified by radiologists, marked with red arrows. Two specimens were obtained in the core biopsy following the MRI, and they both yielded ductal carcinoma in situ (high nuclear grade, solid and cribriform types, with necrosis and focal microcalcifications).

### 9.5. Overestimated POM on negative/benign cancers

Here we investigate a few situations where the AI system outputted a high probability of malignancy, even though the case turned out to be benign or negative.

*Case 10.* Here, four out of five radiologists interpreted the examination as negative. One radiologist (Reader 1) would biopsy the right breast. Our system gave a relatively high POM for the right breast (68%).

| | **Left breast POM** | **Right breast POM** |
|---|---|---|
| **Reader 1** | 0 | 5 |
| **Reader 2** | 0 | 0 |
| **Reader 3** | 0 | 0 |
| **Reader 4** | 0 | 0 |
| **Reader 5** | 0 | 0 |
| **AI system** | **6** | **68** |

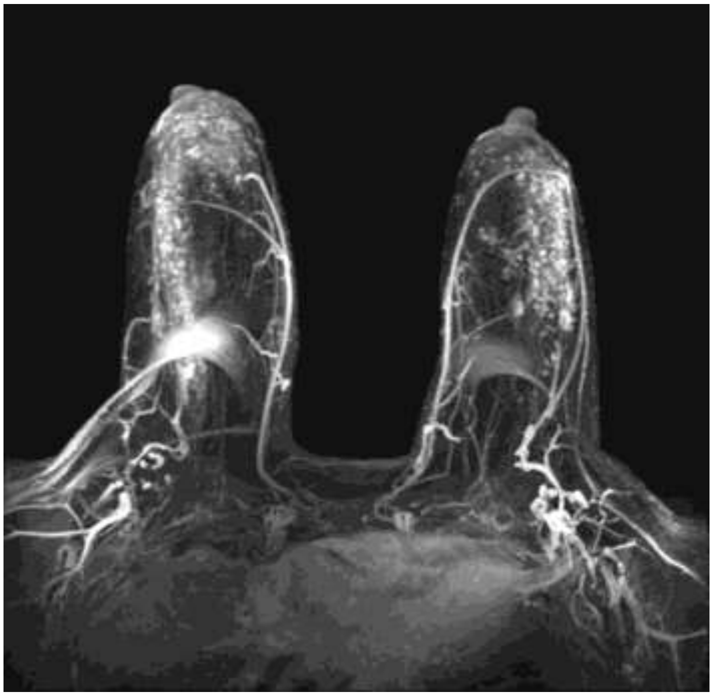


**Figure S17:** Axial maximum intensity projection from Case 10.

*Case 11.* In this examination, four out of five radiologists would biopsy the finding in right breast, and POM given to this imaging exam varied significantly. Ultimately, the finding was biopsied and was found benign. Although our system's POM was very similar to radiologists, we would expect a highly accurate model to give lower POM to benign cases.

| | **Left breast POM** | **Right breast POM** |
|---|---|---|
| **Reader 1** | 0 | 0 |
| **Reader 2** | 0 | 75 |
| **Reader 3** | 0 | 10 |
| **Reader 4** | 0 | 30 |
| **Reader 5** | 0 | 75 |
| **AI system** | **4** | **55** |

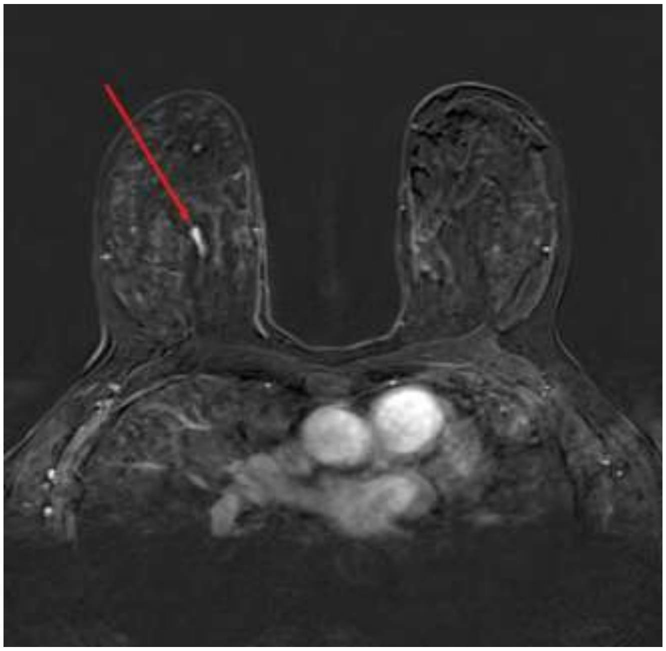
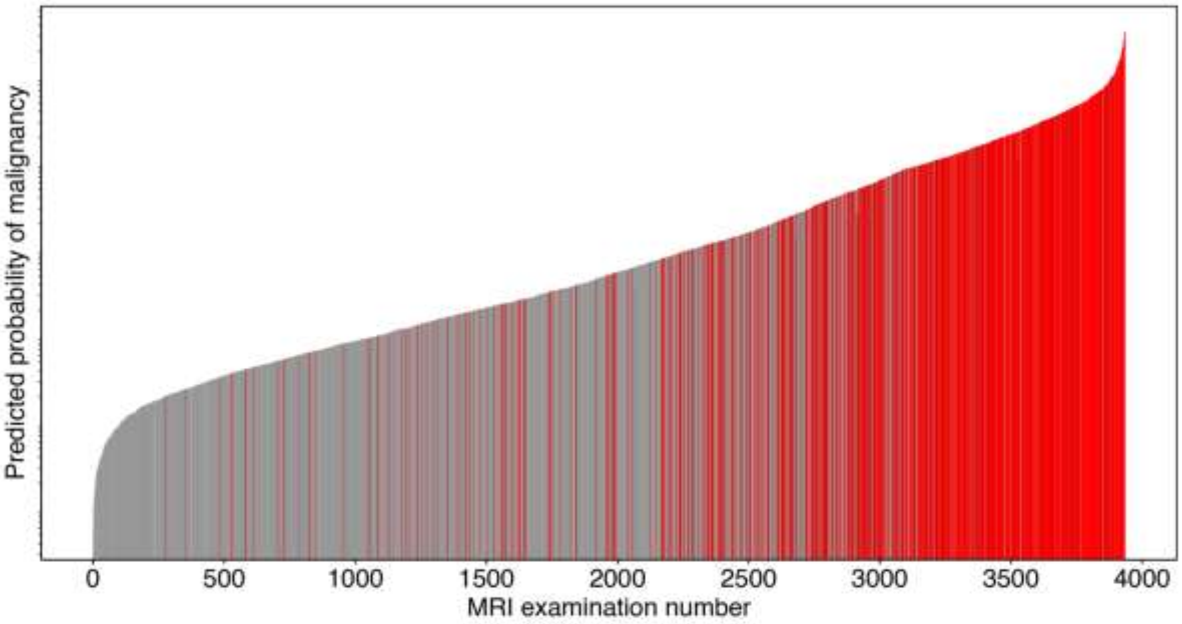


**Figure S18:** Axial subtraction image showing suspicious lesion that was later biopsied. Pathology report showed that the finding was benign, yielding fibrocystic changes, including columnar changes and stromal fibrosis.

## 9. Distribution of predicted probabilities of malignancy



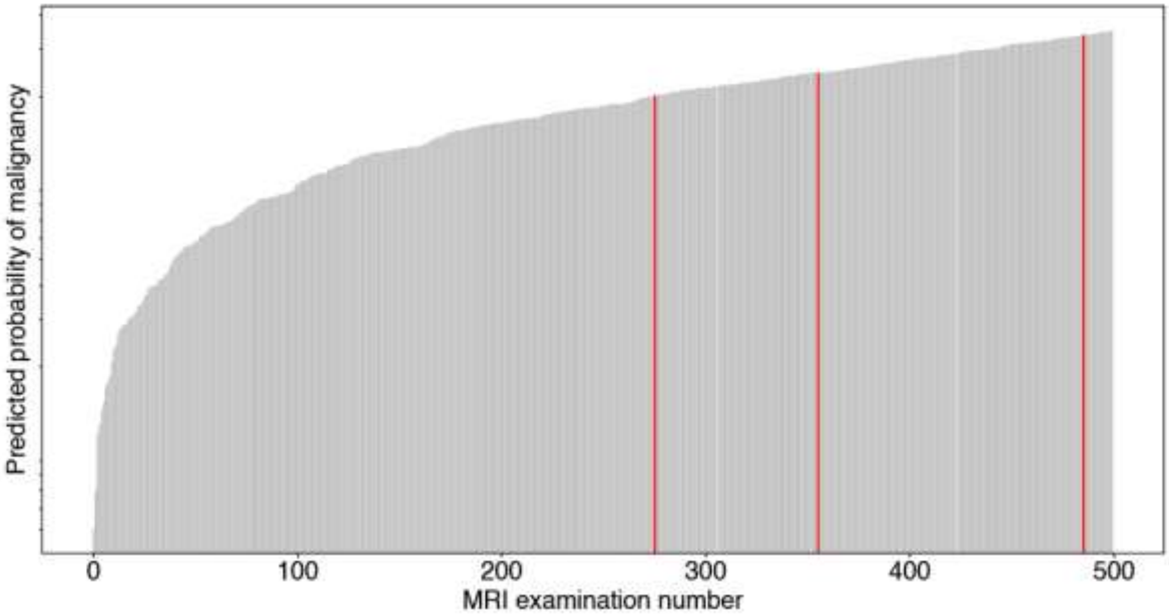


**Figure S19: Distribution of predicted probabilities of malignancy (POM) on the NYU Langone test set.** Each bar represents a POM for a single study (maximum between left and right breast POMs) and all bars are ordered by POM in an increasing manner. Red bars represent malignant cases, whereas black bars are non-malignant. **Top figure** shows all NYU Langone test set cases, meanwhile the **bottom figure** zooms in on the first 500 cases.