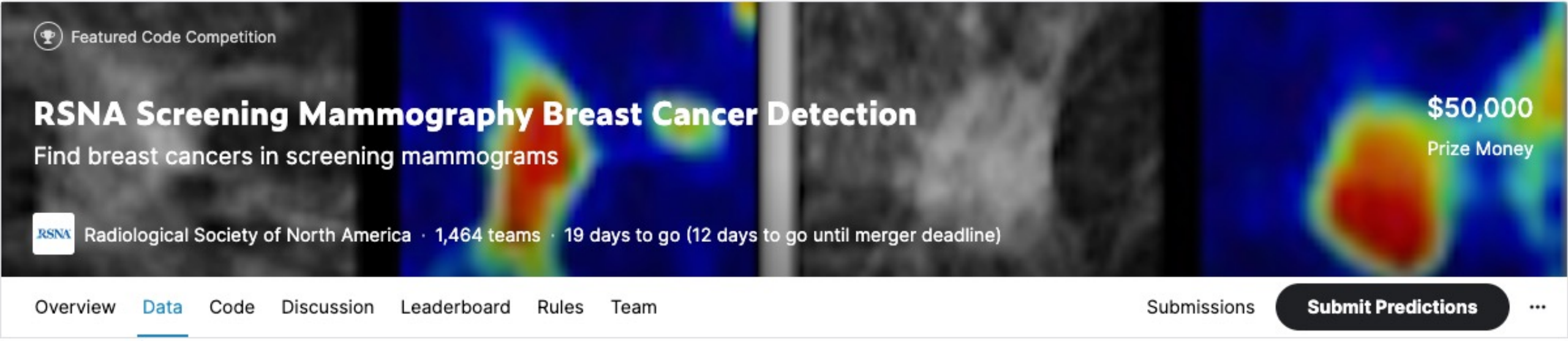


Goal: Identify breast cancer in mammograms!

Mortality in high-income countries dropped 40% when regular mammography screening was implemented.

Early detection and treatment are critical to reducing deaths

We need to streamline the process radiologists use to evaluate screening mammograms – but highly trained personnel is expensive!

Data: train – 11,913 mammograms in dicom format (4 images per patient)
test – 4 available, but can submit test once the notebook is running

Look out for: - Same patient several (up to four) images
- Metadata with files
- Images come from different machines
- implants only for the patient not breasts
- False positives as well as negatives to be taken into account

Data: train.csv

Out[7]:

	site_id	patient_id	image_id	laterality	view	age	cancer	biopsy	invasive	BIRADS	implant	density	machine_id	difficult_negative_case
0	2	10006	462822612	L	CC	61.0	0	0	0	NaN	0	NaN	29	False
1	2	10006	1459541791	L	MLO	61.0	0	0	0	NaN	0	NaN	29	False
2	2	10006	1864590858	R	MLO	61.0	0	0	0	NaN	0	NaN	29	False
3	2	10006	1874946579	R	CC	61.0	0	0	0	NaN	0	NaN	29	False
4	2	10011	220375232	L	CC	55.0	0	0	0	0.0	0	NaN	21	True

In [8]:
```
# ID code for the source hospital.
data.site_id.unique()
```

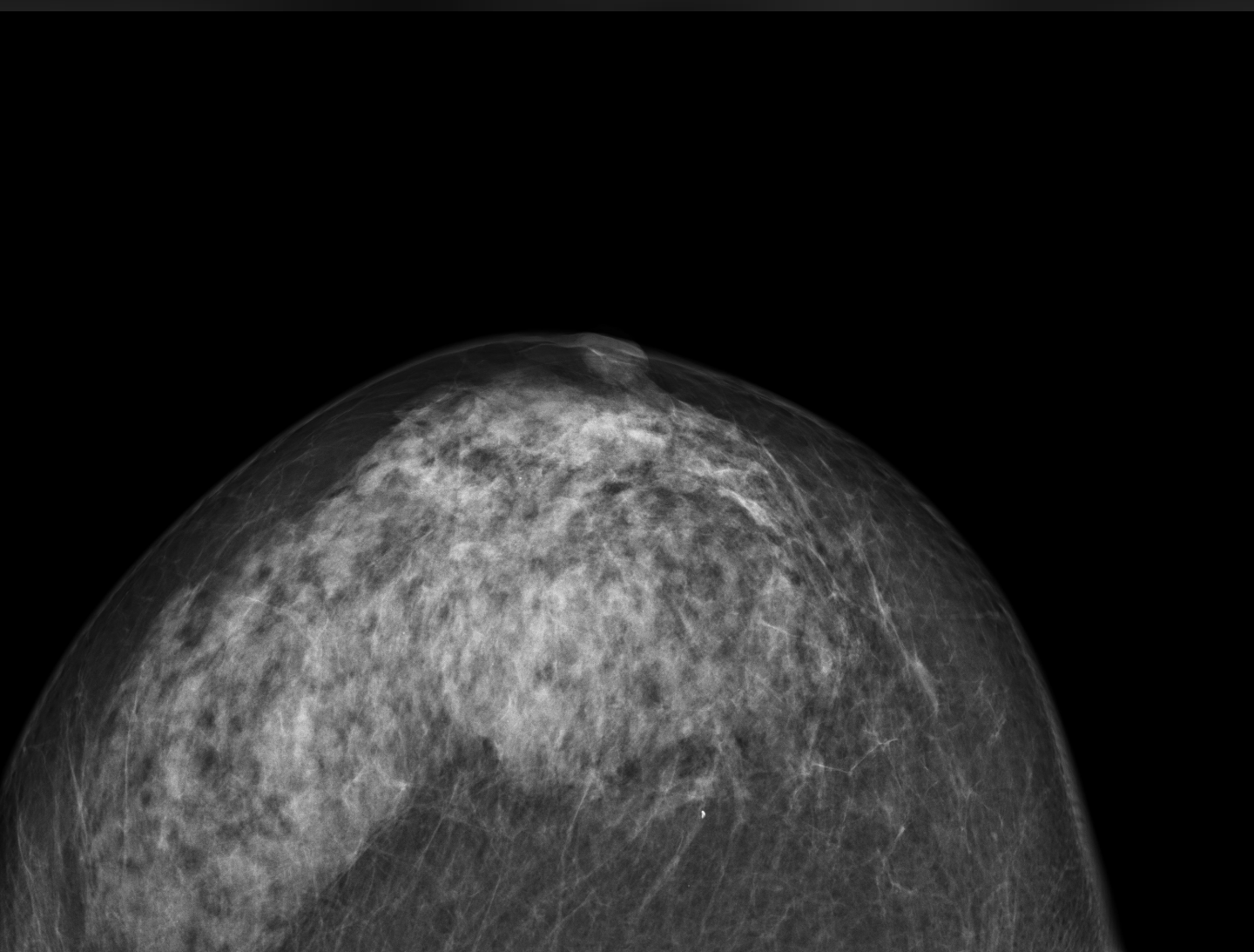
Out[8]:
```
array([2, 1])
```

In [9]:
```
# An ID code for the imaging device.
data.machine_id.unique()
```

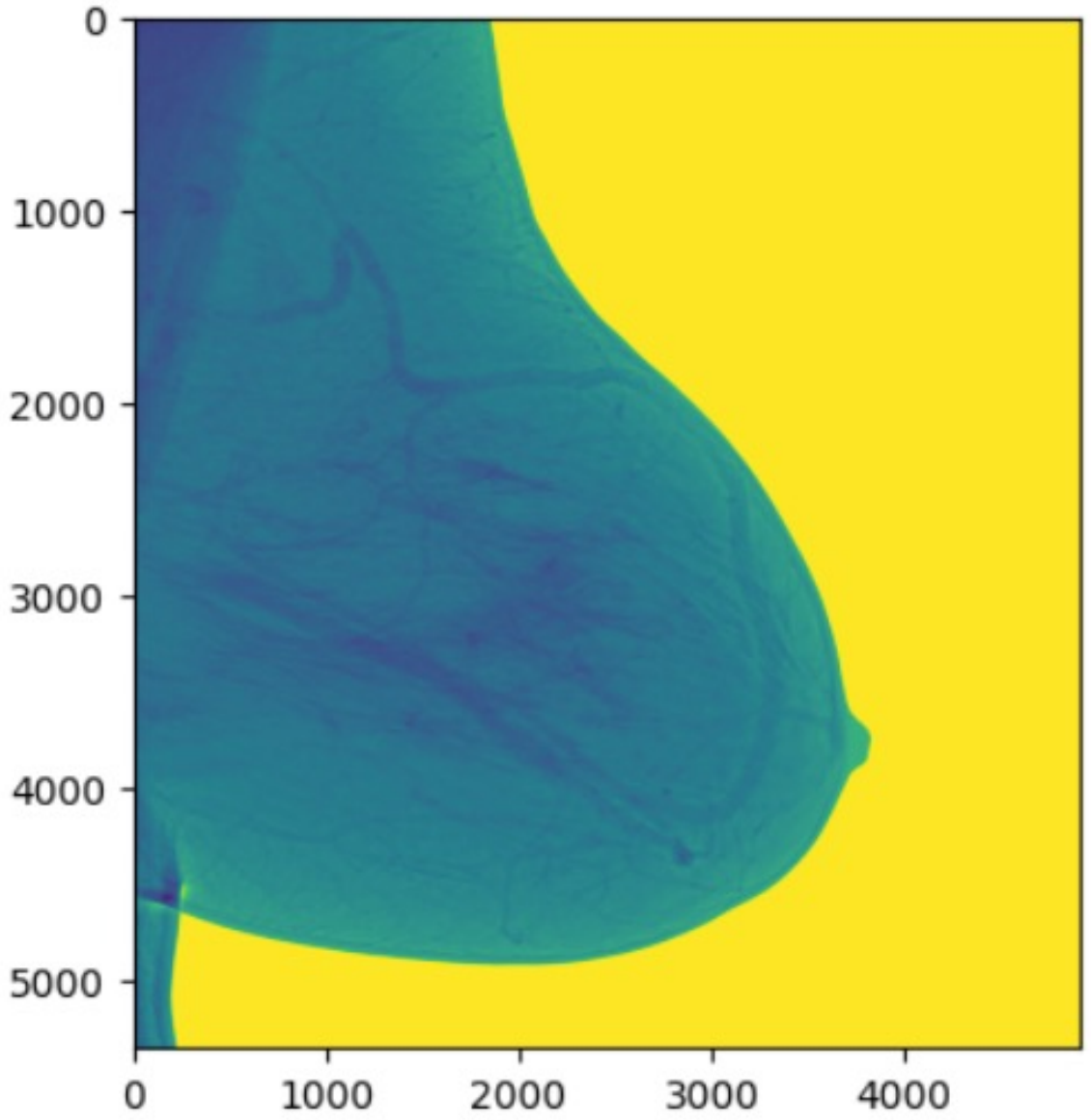
Out[9]:
```
array([ 29,  21, 216,  93,  49,  48, 170, 210, 190, 197])
```

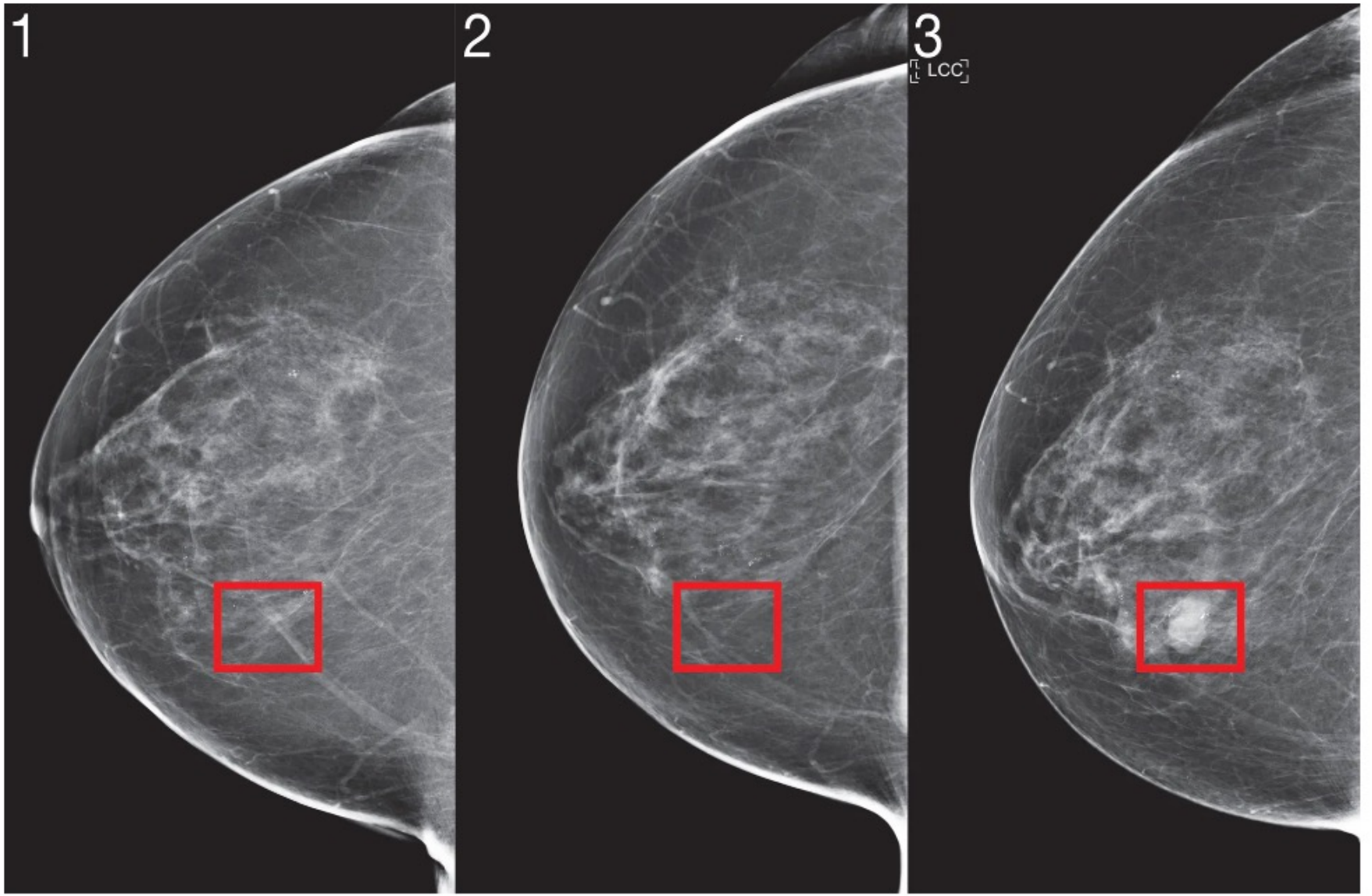
[train/test].csv Metadata for each patient and image. Only the first few rows of the test set are available for download.

- `site_id` - ID code for the source hospital.
- `patient_id` - ID code for the patient.
- `image_id` - ID code for the image.
- `laterality` - Whether the image is of the left or right breast.
- `view` - The orientation of the image. The default for a screening exam is to capture two views per breast.
- `age` - The patient's age in years.
- `implant` - Whether or not the patient had breast implants. Site 1 only provides breast implant information at the patient level, not at the breast level.
- `density` - A rating for how dense the breast tissue is, with A being the least dense and D being the most dense. Extremely dense tissue can make diagnosis more difficult. Only provided for train.
- `machine_id` - An ID code for the imaging device.
- `cancer` - Whether or not the breast was positive for malignant cancer. The target value. Only provided for train.
- `biopsy` - Whether or not a follow-up biopsy was performed on the breast. Only provided for train.
- `invasive` - If the breast is positive for cancer, whether or not the cancer proved to be invasive. Only provided for train.
- `BIRADS` - 0 if the breast required follow-up, 1 if the breast was rated as negative for cancer, and 2 if the breast was rated as normal. Only provided for train.
- `prediction_id` - The ID for the matching submission row. Multiple images will share the same prediction ID. Test only.
- `difficult_negative_case` - True if the case was unusually difficult. Only provided for train.

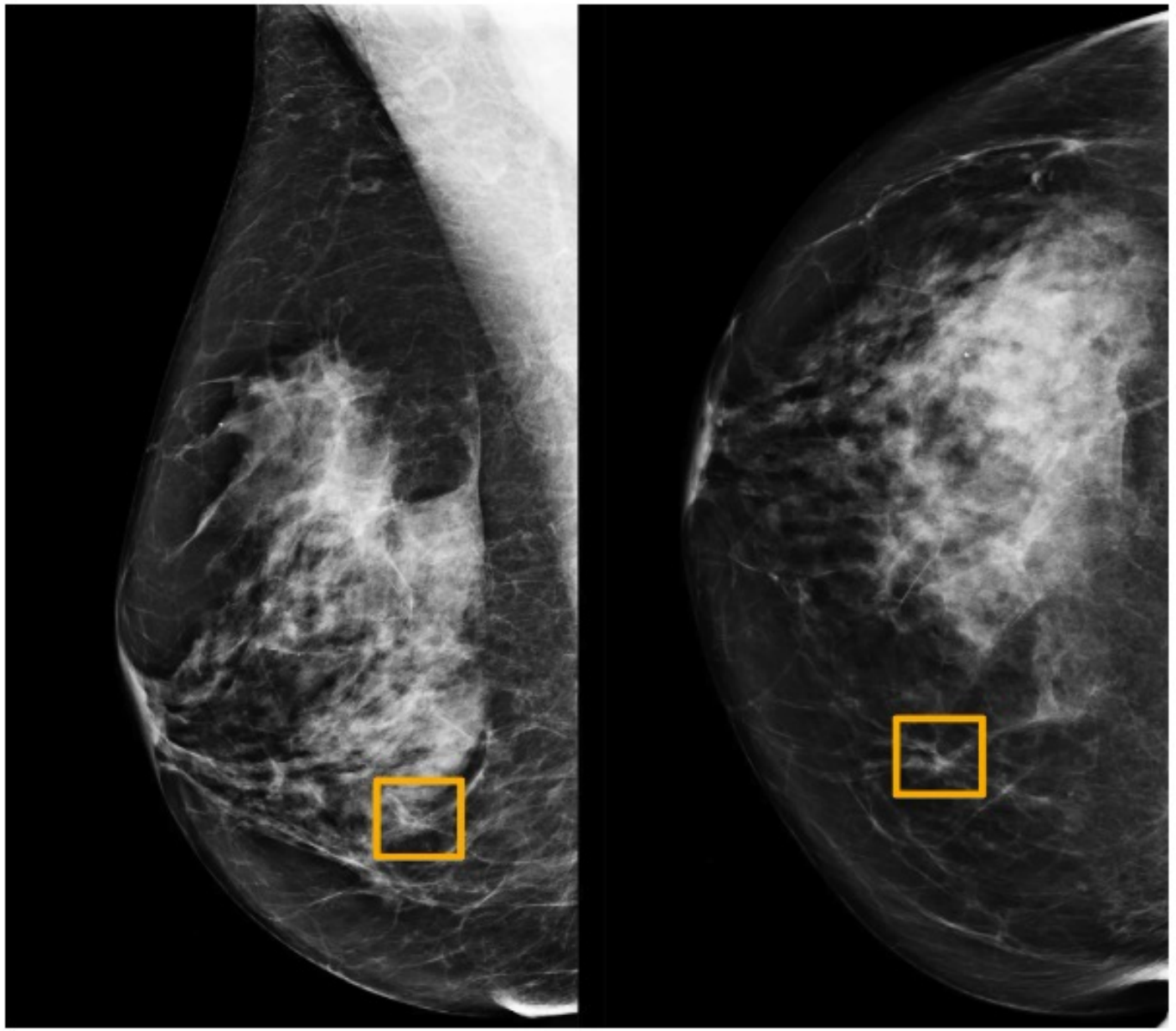


In [15]:
```
imgplot = plt.imshow(pixel_array_numpy)
```





Images from a mammogram of a patient whom the algorithm identified as high risk four years before cancer was diagnosed. COURTESY OF MIT



A yellow box indicates where an A.I. system found cancer hiding inside breast tissue. Six previous radiologists failed to find the cancer in routine mammograms.

Northwestern University

```
In [16]: meta = dicom.read_file(path+"train_images/1028/315268478.dcm")
```

```
In [17]: meta
```

```
Out[17]: Dataset.file_meta -------------------------------
(0002, 0001) File Meta Information Version       OB: b'\x00\x01'
(0002, 0002) Media Storage SOP Class UID         UI: Digital X-Ray Image Storage - For Presentation
(0002, 0003) Media Storage SOP Instance UID      UI: 1.2.840.10009.1.2.3.1028.1.315268478
(0002, 0010) Transfer Syntax UID                 UI: JPEG 2000 Image Compression (Lossless Only)
(0002, 0012) Implementation Class UID            UI: 1.2.840.113654.2.3.1995.2.12.0
(0002, 0013) Implementation Version Name         SH: 'PYDICOM 2.3.0'
-------------------------------------------------
(0008, 0018) SOP Instance UID                    UI: 1.2.840.10009.1.2.3.1028.1.315268478
(0008, 0023) Content Date                        DA: '20221118'
(0008, 0033) Content Time                        TM: '184049.243873'
(0010, 0020) Patient ID                          LO: '1028'
(0020, 000d) Study Instance UID                  UI: 1.2.840.10009.1.2.3.1028
(0020, 000e) Series Instance UID                 UI: 1.2.840.10009.1.2.3.1028.1
(0020, 0013) Instance Number                     IS: '315268478'
(0020, 0062) Image Laterality                    CS: 'L'
(0028, 0002) Samples per Pixel                   US: 1
(0028, 0004) Photometric Interpretation          CS: 'MONOCHROME1'
(0028, 0010) Rows                                US: 5355
(0028, 0011) Columns                             US: 4915
(0028, 0100) Bits Allocated                      US: 16
(0028, 0101) Bits Stored                         US: 16
(0028, 0102) High Bit                            US: 15
(0028, 0103) Pixel Representation                US: 0
(0028, 0120) Pixel Padding Value                 US: 4196
(0028, 1040) Pixel Intensity Relationship        CS: 'LOG'
(0028, 1041) Pixel Intensity Relationship Sign   SS: 1
(0028, 1050) Window Center                       DS: [1802.310000, 1802.310000, 2020.704000, 1583.916000]
(0028, 1051) Window Width                        DS: [1091.970000, 1091.970000, 1091.970000, 1091.970000]
```