Radiology: Artificial Intelligence

DATA RESOURCES

The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.4 Million Screening and Diagnostic Mammographic Images

Jiwoong J. Jeong, MS • Brianna L. Vey, MD* • Ananth Bhimireddy, MS • Thomas Kim, BS • Thiago Santos, MS • Ramon Correa, MS • Raman Dutt, BS • Marina Mosunjac, MD • Gabriela Oprea-Ilies, MD • Geoffrey Smith, MD • Minjae Woo, PhD • Christopher R. McAdams, MD • Mary S. Newell, MD • Imon Banerjee, PhD • Judy Gichoya, MD, MS • Hari Trivedi, MD*

From the School of Computing and Augmented Intelligence, Arizona State University, 699 S Mill Ave, Tempe, AZ 85281 (J.J.J., R.C., I.B.); Departments of Radiology (B.L.V., A.B., R.D., C.R.M., M.S.N., J.G., H.T.), Computer Science (T.S.), and Pathology (M.M., G.O., G.S.), Emory University, Atlanta, Ga; College of Computing, Georgia Institute of Technology, Atlanta, Ga (T.K.); and School of Data Science and Analytics, Kennesaw State University, Kennesaw, Ga (M.W.). Received March 7, 2022; revision requested April 19; revision received November 4; accepted December 16. **Address correspondence to** J.J.J. (email: *jjeong35@asu.edu*).

* J.J.J. and B.L.V. contributed equally to this work.

Supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health (award no. UL1TR002378).

Conflicts of interest are listed at the end of this article.

Supplemental material is available for this article.

Radiology: Artificial Intelligence 2023; 5(1):e220047 • https://doi.org/10.1148/ryai.220047 • Content codes: AI BR

Breast cancer detection remains one of the most frequent commercial and research applications for deep learning (DL) in radiology (1–3). Development of DL models to improve breast cancer screening requires robustly curated and demographically diverse datasets to ensure generalizability. However, most publicly released breast cancer datasets are ethnically and racially homogeneous (4–7), are relatively small, and lack image annotations and/or pathologic data. Specifically, African American and other minority patients are severely underrepresented in breast imaging and other health care–related datasets despite having the worst breast cancer prognosis (8,9). Although a recent large study from Google demonstrated an overall improvement in rates of breast cancer screening recall when their model was used alongside radiologists, very few African American patients were included (10). Given that 45% of the patients at our institution are African American, we were able to curate a diverse dataset to represent African American women in breast imaging research.

Limited datasets lead to weak artificial intelligence models (11–13) that underperform with regard to patients not included in training data, leading to inadvertent systemic racial bias and health care disparities (14–17). For example, many models trained for skin cancer detection and genomics use data consisting of up to 96% White or European patient groups (18,19). Current breast imaging datasets are also lacking in size and granularity. For example, a large dataset created for the DREAM (Dialogue on Reverse Engineering Assessment and Methods) challenge contains 640 000 screening mammograms labeled as benign or malignant from 86 000 patients, but less than 0.2% of cases (4) were positive and there were no regions of interest (ROIs). The CBIS-DDSM (Curated Breast Imaging Subset of the Digital Database for Screening Mammography) dataset (6) contains 2620 scanned film mammograms with lesion annotations; however, scanned filmed mammograms differ from full-field digital mammograms (FFDMs) and therefore cannot be used in isolation to train artificial intelligence models for FFDM. The only large (>10 000 cases), publicly available dataset that contains lesion-level ROIs and stratified pathologic diagnoses is the Optimam Mammography Image Database (OMI-DB) (20), which lacks semantic imaging descriptors. The EMory BrEast imaging Dataset (EMBED) contains lesion-level annotations, pathologic outcomes, and demographic information for 116 000 patients from racially diverse backgrounds and will help bridge the existing gaps in granularity, diversity, and scale in breast imaging datasets.

Materials and Methods

With the approval of the Emory University's institutional review board, this retrospective dataset of curated mammograms was developed using largely automated and semiautomated curation techniques that are detailed below. Dataset development was facilitated by the high level of data homogeneity within and across institutions, which must adhere to the Breast Imaging Reporting and Data System (BI-RADS) (21) and the Mammography Quality Standards Act guidelines (22). The need for written informed consent from patients was waived because of the use of deidentified data.

Data Collection

We identified patients with a screening or diagnostic mammogram at our institution from January 2013 through December 2020. Data were collected from four institutional hospitals (two community hospitals, one large inner-city hospital, and one private academic hospital). Women aged 18 years or older with at least one available mammogram in our picture archiving and communication system were included. Exclusion criteria were any patient younger than aged 18 years. An overview of the full dataset is provided in Figure 1.

Data Extraction

Images

Abbreviations

BI-RADS = Breast Imaging Reporting and Data System, DBT = digital breast tomosynthesis, DICOM = Digital Imaging and Communications in Medicine, DL = deep learning, EMBED = EMory BrEast imaging Dataset, FFDM = full-field digital mammogram, ROI = region of interest, 2D = two-dimensional

Summary

The EMory BrEast imaging Dataset (EMBED) contains two-dimensional and digital breast tomosynthesis screening and diagnostic mammograms with lesion-level annotations and pathologic information in racially diverse patients.

Key Points

- The dataset includes 3 383 659 two-dimensional and digital breast tomosynthesis screening and diagnostic mammograms from 116 000 women, with an equal representation of African American and White patients.
- The dataset also contains 40 000 annotated lesions linked to structured imaging descriptors and 56 ground truth pathologic outcomes grouped into seven severity classes.
- Twenty percent of the dataset is being made freely available for research through the Amazon Web Services Open Data Program.

Keywords

Mammography, Breast, Machine Learning

All mammographic examinations were extracted from the institutional picture archiving and communication system in Digital Imaging and Communications in Medicine (DICOM) format using Niffler (23), an open-source pipeline developed in-house for retrospective image extraction that leverages Pydicom (24). Of 281 509 screening and 83 387 diagnostic examinations, 148 320 (52.7%) screening and 65 265 (78.3%) diagnostic examinations were two dimensional (2D) only, and 133 189 (47.3%) screening and 18 122 (21.7%) diagnostic examinations were 2D plus digital breast tomosynthesis (DBT) plus synthetic 2D. Through use of an open-source in-house Python library (25), all images were de-identified as follows: *(a)* in DICOM format with all Health Insurance Portability and Accountability Act metadata elements removed or de-identified and *(b)* as 16-bit PNG files with de-identified DICOM metadata stored separately. PNG files are valuable in DL pipelines because they can be loaded more quickly for model development. Dates were shifted using fixed patient-level offset to maintain temporality between imaging examinations and clinical data. A master key was retained for regularly adding new examinations to the dataset.

Imaging Findings

Imaging findings were recorded at the time of interpretation in MagView software, version 8.0.2130, and output into a structured database that includes information such as examination type (screening or diagnostic), reason for visit, BI-RADS score, and BI-RADS imaging descriptors (26). Imaging descriptors include information on appearance of masses, calcifications, distribution and location of findings, presence of implants, and additional nonlocalized imaging findings (such as global asymmetries) (Table 1). Findings were noted on a per-finding and per-breast basis, resulting in zero to several findings for each breast on an examination.

Pathologic Outcomes

Ground truth pathologic outcomes were clinically correlated and manually recorded in MagView on a per-finding basis by administrative staff. These outcomes include results of fine-needle aspiration, core and excisional biopsies, lumpectomy, and mastectomy for breast tissue and/or axillary lymph nodes. Each imaging finding may be associated with up to 10 pathologic results but rarely contains more than four. As a secondary check for troubleshooting discrepancies in pathologic outcomes, we separately extracted all free-text pathologic records from an institutional database (CoPath). Pathologic results outside the breast, chest wall, or axilla were excluded, even for primary breast malignancy.

Demographic Characteristics, Family, and Treatment History

Demographic characteristics, including age, race, ethnicity, and insurance status, were collected for each patient through electronic health records. Family, procedure, and treatment histories, including history of hormone replacement therapy, were also available through self-reported intake forms in MagView for many patients but are subject to accurate reporting. Traditional risk factors and Gail and Tyrer-Cuzick risk scores were collected when available.

Data Curation

Imaging

There were four challenges associated with image data curation: *(a)* differentiation of 2D, DBT, and synthetic 2D images; *(b)* differentiation between standard mammographic views and special views (spot, magnification, or procedural views, such as those obtained during biopsies and wire localizations) in diagnostic examinations; *(c)* extraction of burned-in ROIs saved directly into pixel data on a copy of the original mammogram (screen save image); and *(d)* extraction of breast tissue from the inside of spot compression or magnification paddles on diagnostic examinations. We designed a semiautomated supervised machine learning pipeline to address these challenges that combines traditional computer vision and DL techniques, summarized in Figure 2.

Differentiation of image type.— Approximately 42% of our examinations are combined 2D and DBT and therefore can contain up to four image types: 2D, DBT, synthetic 2D, or screen save images containing ROIs. We applied a rules-based approach using the SeriesDescription, ViewPosition, and ImageLaterality tags in the DICOM metadata to identify and label each image type. Results were manually verified on a random test set of 5000 images and were 100% accurate.

Differentiation between standard and special views.— To differentiate between standard and special views for 2D images, supervised image classification and metadata filtration methods were both attempted. To classify on the basis of image appearance alone, a VGG11 DL model (27) with batch normalization was

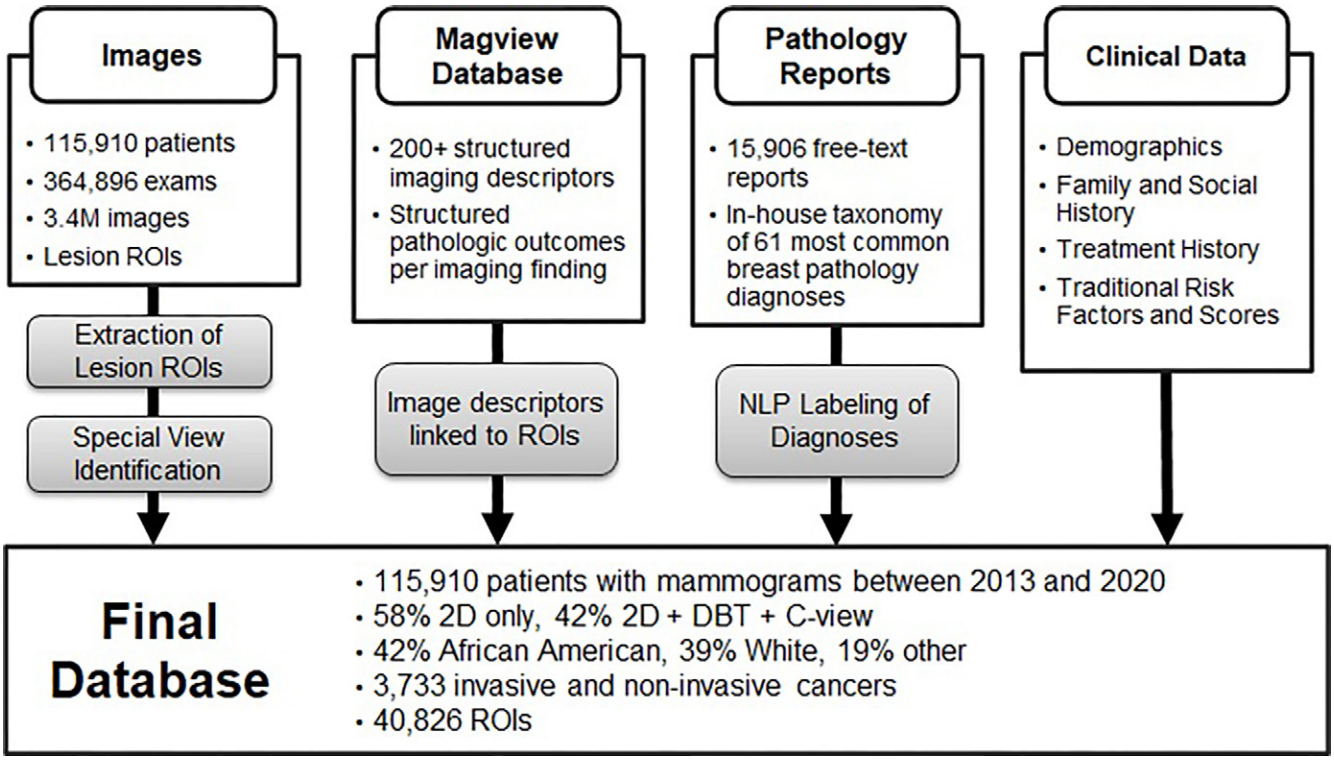

Figure 1: Overview of extraction and curation of the EMory BrEast imaging Dataset (EMBED) with four core components: images and regions of interest (ROIs), structured imaging descriptors and pathologic outcomes from MagView, free-text pathology reports, and additional clinical data. "Other" racial category includes Asian, not reported, and mixed. DBT = digital breast tomosynthesis, NLP = natural language processing, 2D = two-dimensional.

Table 1: Samples of Imaging Descriptor Categories and Commonly Encoded Values Available from MagView

Encoded Descriptors	Possible Values
Mass	
Shape	Generic (G)*, Round (R), Oval (O), Irregular (X), Questioned architectural distortion (Q), Architectural distortion (A), Asymmetric tubular structure/solitary dilated duct (T), Intramammary lymph nodes (N), Global asymmetry (B), Focal asymmetry (F), Developing asymmetry (V), Lymph node (Y)
Margin	Circumscribed (D), Obscured (U), Microlobulated (M), Indistinct (I), Spiculated (S)
Density	High density (+), Isodense (=), Low density (-), Fat containing (0)
Calcification	
Finding	Amorphous (A), Benign (9), Coarse heterogenous (H), Coarse popcornlike (C), Dystrophic (D), Rim (E), Fine-linear (F), Fine linear-branching (B), Generic (G), Fine pleomorphic (I), Large rodlike (L), Milk of calcium (M), Oil cyst (J), Pleomorphic (K), Punctate (P), Round (R), Skin (S), Lucent centered (O), Suture (U), Vascular (V), Coarse (Q)
Distribution	Grouped (G), Segmental (S), Regional (R), Diffuse/scattered (D), Linear (L), Clustered (C)
Other	
Size and position	Side (L or R), Size (in millimeters), Location (quadrant, subareolar, or axillary tail), Depth (anterior, middle, posterior), Distance (in centimeters)†
Related findings to primary finding	Postlumpectomy change (U), Postlumpectomy and radiation change (1), Postsurgical change (P), Biopsy clip (W), Postreduction change (C), Focal asymmetry (Q), Asymmetry (4), Prominent lymph node (2), Mastectomy and flap reconstruction (!)

Note.—Each finding described by the radiologist at the time of interpretation is encoded into a structured format. Most examinations have no associated descriptors because of negative screening results.
* Generic (G) may be coded for a mass in a screening examination in which the radiologist does not wish to further describe mass characteristics.
† Distance refers to the distance from nipple for both mammography and US.

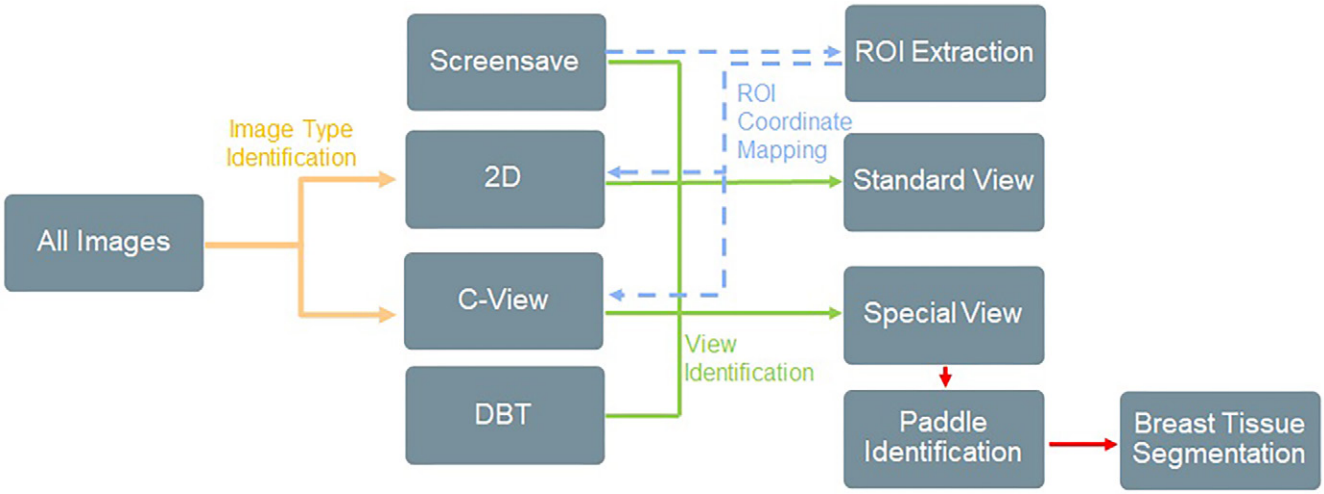

Figure 2: Overview of image filtration for classifying image types and extracting relevant regions of interest (ROIs) and tissue patches. This was achieved using a combination of computer vision techniques, Digital Imaging and Communications in Medicine (DICOM) metadata, and rules-based heuristics. ROI extraction was achieved by identifying ROIs within screen save images, extracting the ROI coordinates, identifying the matching source mammogram, and then mapping the ROI coordinates back to the original image. Examples of ROI extraction and special view tissue segmentation are provided in Figures 3 and 4. DBT = digital breast tomosynthesis, 2D = two-dimensional.

trained, tested, and validated on a manually curated 5000-image dataset containing FFDM and special views, achieving an overall accuracy of 98.46%. Because of the imperfect results, we also examined the DICOM metadata of FFDM and special views and found that a private tag, 0_ViewCodeSequence_0_ViewModifierCodeSequence_CodeMeaning, could be used for differentiation. We manually verified the results on a separate test set of 5000 randomly selected images and confirmed the results were 100% accurate. This metadata tag might be valid only at our institution, so the image-based classification model is available for public use (28) because it may be more generalizable. Using this technique, we identified 208 254 images containing special views.

ROI extraction and mapping to original mammogram.— ROIs were annotated by the interpreting radiologist in a single-reader setting to localize abnormal findings. These are saved as a white circle directly onto a copy of the original mammogram, generating a screen save image. Because most ROIs were created on screening examinations, there are relatively few annotations on diagnostic images. To automate a method of ROI detection and localization, 450 screen save images were randomly selected, and the location of the circular ROI was annotated by a trained student (J.J.J.) using bounding boxes on the MD.ai platform. Annotations were used to train an object detection DL model using the EfficientDet-b0 architecture (29) to localize the ROIs. Detection accuracy on the test set was 99.99% with an intersection over union of 0.95, including images with multiple ROIs. We then ran inference on all remaining screen save images to detect ROIs and manually verified localization accuracy on a test set of 5000 cases. Finally, to map the ROI location from screen save back to the original mammogram, we first identified the original mammogram using an image similarity function from the Simple ITK Python library (30). The ROI coordinates from the screen save were then scaled and translated back to the original mammogram and saved into the database (Fig 3). This process was manually verified on 2000 images to ensure proper matching to the source mammogram and ROI translation, with only three matching errors discovered (<0.2%). Approximately 90% of ROIs were from screening examinations assigned a BI-RADS 0 assessment, and approximately 70% and 30% were from 2D and synthetic 2D images, respectively. No ROIs are available for DBT examinations, although we anticipate some may become available through stored DICOM objects. The description of the increasing the number of available ROIs and ROI mapping to MagView is described in Appendix S1.

Extraction of tissue inside paddles for special diagnostic views.— Twelve distinct paddle types were identified from images with special diagnostic views, such as spot compression and magnification. Paddles were categorized by shape, which may be rectangular or circular (Fig 4). Because paddles are metallic, their pixel intensities at mammography were substantially higher than those of the surrounding tissue. Intensity histogram analysis and postprocessing were used to identify this intensity difference to determine the paddle location and subsequently extract breast tissue within the paddles. For rectangular paddles, a simple thresholding method was used to convert the gray-scale images into binary images to identify paddle edges based on a row and column sum of pixels (Fig 4). For the circular paddles, a feature engineering technique known as Hough circle transformation was used to detect any metallic device in a circular shape that was then used to maximize the area of the tissue within the paddle. Tissue location inside the paddle was saved separately as a binary mask for each special diagnostic view mammogram.

Pathologic Data

Pathologic data were collected using in-house taxonomy for breast abnormalities (Fig 5; full list in Table S1), which was created in consultation with breast pathologists and oncologists to

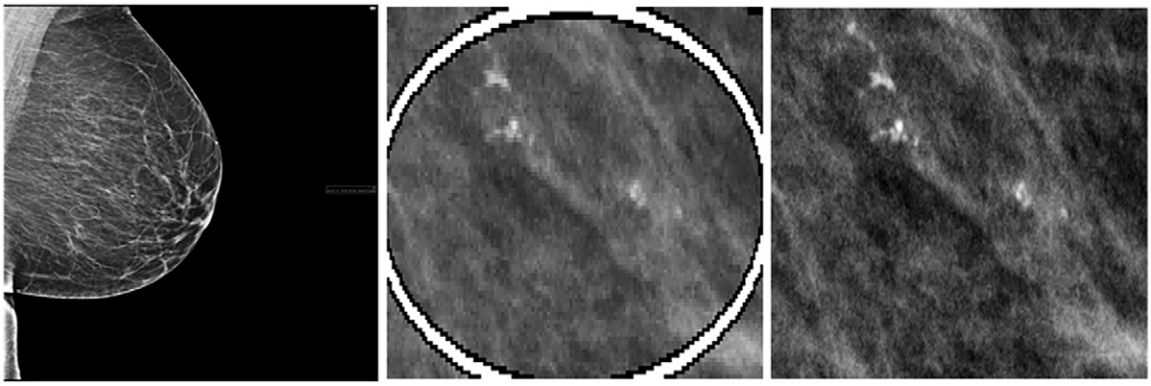

Figure 3: Sample region of interest (ROI) extraction from a 43-year-old African American woman with right craniocaudal screening mammogram, Breast Imaging Reporting and Data System (BI-RADS) 0, followed by US-guided biopsy resulting in a diagnosis of ductal carcinoma in situ. Left: Low-resolution screen save copy of mammogram containing a burned-in ROI annotated by the interpreting radiologist. Middle: Deep learning–based ROI detection and coordinate extraction from the screen save image. Right: Matching original mammogram is found using image comparison from the Python Simple ITK library, and the coordinates of the ROI are mapped to the original image.

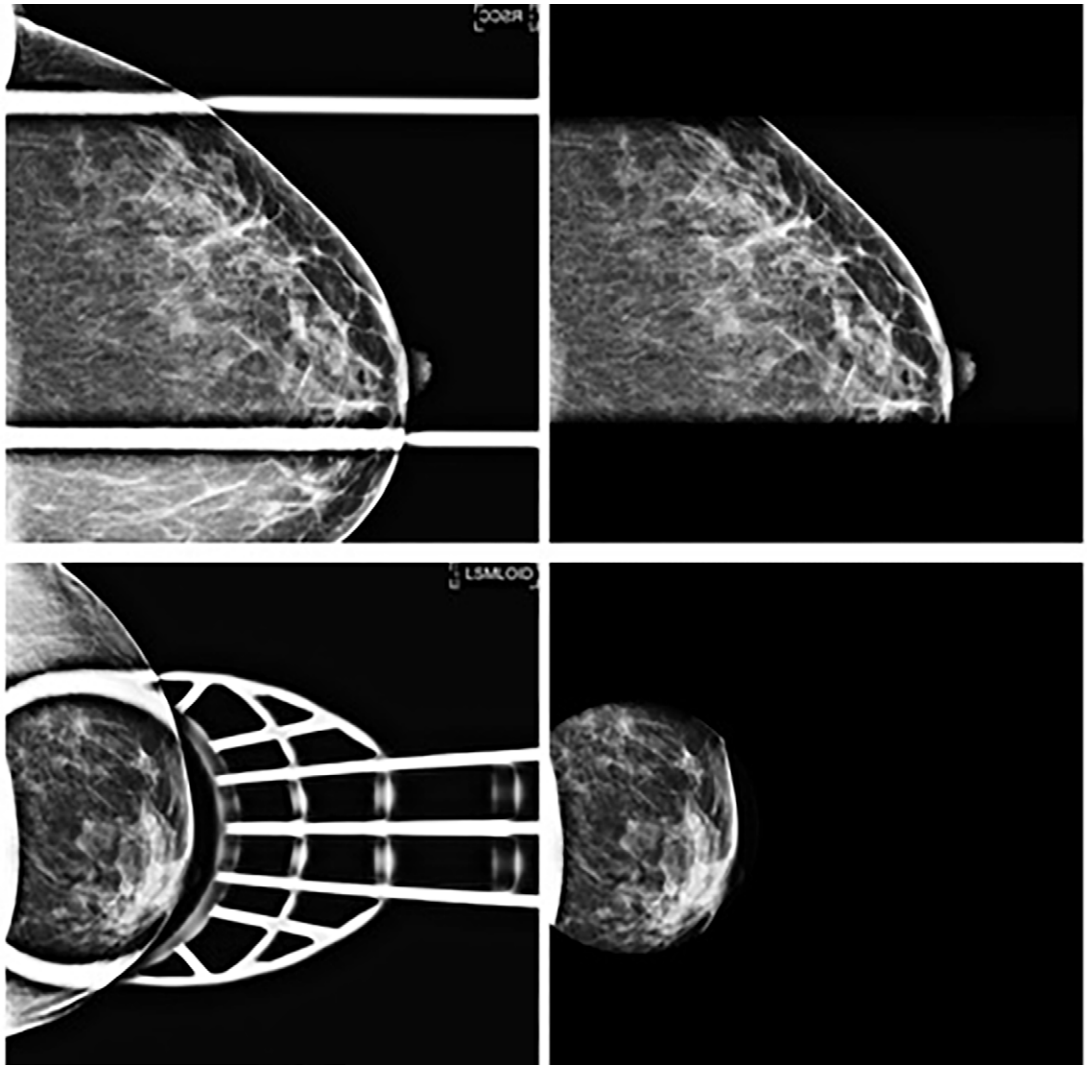

Figure 4: Examples include (top) a 49-year-old White woman with right mediolateral oblique diagnostic mammogram with large spot compression paddle (Breast Imaging Reporting and Data System [BI-RADS] 1) and (bottom) a 40-year-old White woman with left craniocaudal diagnostic mammogram with small spot compression paddle (BI-RADS 4), followed by US-guided biopsy with benign results. This includes special magnification and spot compression views (left) with resultant extracted images of tissue inside the paddle (right) that were achieved using histogram analysis. Extracted tissue was saved as a pixel mask corresponding to the original mammogram.

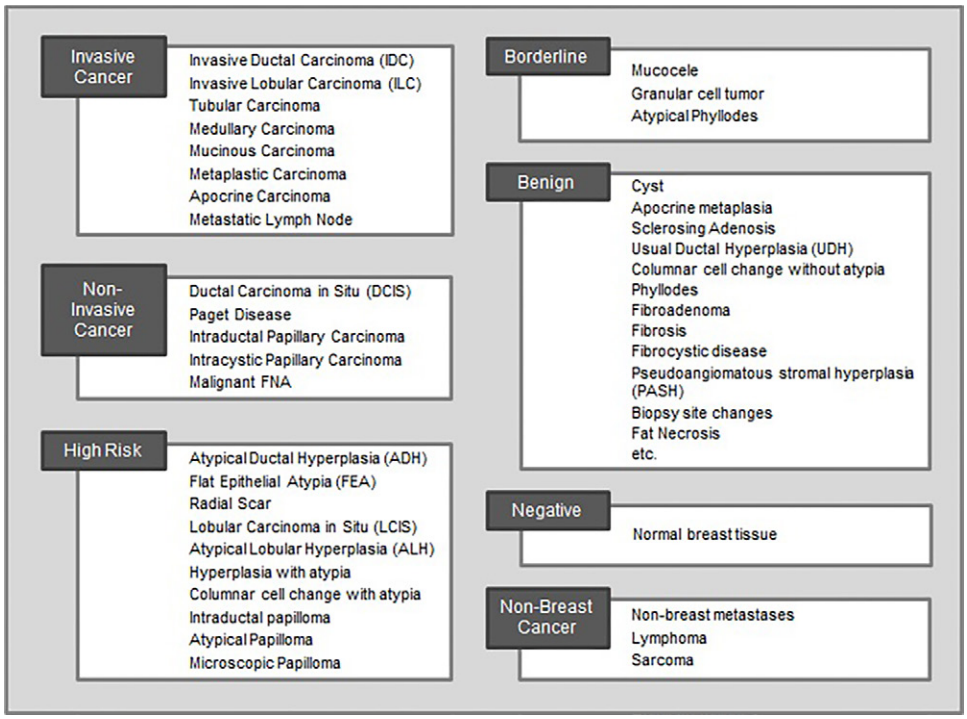

Figure 5: Taxonomy of the 56 most common breast pathologic findings at our institution grouped into seven categories by severity. Each pathology report was tagged with one or more diagnoses from this list. The benign category contains 23 diagnoses but is abbreviated in this figure. A full list is included in Table S1. FNA = fine-needle aspiration.

identify the 56 most common findings in breast pathology reports. Abnormalities were divided into the following seven severity groups: invasive breast cancer, in situ cancer, high-risk lesion, borderline lesion, nonbreast cancer, benign, or negative.

In MagView, pathologic information as well as the source of the sample, such as biopsy or surgery, for each finding was converted to one of the seven severity groups using a lookup table. For example, if the pathologic outcome for a finding was [ductal carcinoma in situ, flat epithelial atypia], these were encoded as [in situ cancer, high-risk lesion]. A second field recorded the worst pathologic diagnosis for that finding, which is in situ cancer.

We performed a secondary check for pathologic outcomes, which were encoded into MagView manually by administrative staff, by training a hierarchical hybrid natural language processing system using 8000 expert-annotated labels to extract pathologic diagnoses from free-text pathology reports. In the first level of the hierarchy, a transformer-based character level BERT (bidirectional encoder representations from transformers) (31) was trained to classify one or many of the seven pathology groups, with an overall F1 score of 94.6% for classifying all pathology groups and 96.8% for only the worst pathology group. Subsequently, at the second level, six independent discriminators were trained to classify individual pathologic diagnoses with an overall F1 score of 90.7%. The negative group contained no individual diagnoses, so secondary discriminators were not required for this class. Model output can be compared directly against the structured pathologic outcomes in MagView, and mismatches can be flagged for human review; however, the efficacy of this strategy is yet to be determined. The model is being validated on external datasets and will be published and released separately to allow other institutions to automatically extract diagnoses from breast pathology reports.

Resulting Dataset

Dataset Characteristics

Patient information is summarized in Table 2. A total of 115 910 unique patients with 364 896 screening and diagnostic mammograms and 3.65 million images were available. The mean age of patients overall and at first mammogram was 59 years ± 12 (SD) and 55 years ± 12, respectively. The self-reported racial distribution was 48 246 (42%) African American, 7552 (7%) Asian, 1130 (1%) Native Hawaiian/Pacific Islander, 13 050 (11%) unknown, and 45 114 (39%) White. The ethnic distribution was 88 025 (76%) non-Hispanic, 6486 (6%) Hispanic, and 21 399 (19%) unknown. The overall distribution of follow-up is shown in Figure 6; 37 939 patients and 24 933 patients had 3- and 5-year follow-ups, respectively. There were 3733 (3%) total patients with cancer, with an annual cancer incidence of 1.16% ± 0.15 at screening mammography.

The distribution of imaging findings separated by BI-RADS score and examination type is shown in Table 3. The number of ROIs linked to each pathologic severity group in screening and diagnostic examinations is shown in Table 4. Lesions that were never biopsied (those that were BI-RADS 1–3 at diagnostic imaging) have no pathologic information and are denoted as such.

Database Size and Structure

The total dataset size is 16.0 terabytes. The MagView data, image metadata, ROI information, and clinical data are stored in a

Table 2: Descriptive Statistics for Full EMory BrEast imaging Dataset (EMBED)

Data	Value
No. of patients	116 902
Mean age (y)*	58.5 ± 12.1
Mean age at first visit (y)*	55.3 ± 12.2
No. of screening mammograms	281 509
No. of diagnostic mammograms	83 387
Mean annual recall rate (%)*	10.6 ± 1.6
Race	
African American	48 246 (41.6)
White	45 114 (38.9)
Asian	7552 (6.5)
Native Hawaiian/Pacific Islander	1130 (1.0)
Multiple	510 (0.4)
American Indian or Alaskan Native	308 (0.3)
Unknown	13 050 (11.3)
Ethnicity	
Hispanic	6486 (5.6)
Non-Hispanic	88 025 (75.9)
Unknown	21 399 (18.5)
Cancer rates	
Total cancers	3733 (3.2)
Annual cancer incidence (%)	1.16 ± 0.15
Regions of interest	
Total	40 826
Directly linked to findings	32 448

Note.—Unless otherwise indicated, data are numbers or numbers with percentages in parentheses. The dataset contains approximately even numbers of African American and White patients. Regions of interest were annotated by interpreting radiologists and could be linked directly to a single finding in approximately 80% of cases. The remaining 20% of regions of interest were from cases with multiple findings and required manual linkage.
* Data are means ± SDs.

MongoDB (32) database. Each data cohort is a collection in the database, and each document in the collection will represent a datapoint. The documents store data in a key-value pair (JSON) format. An anonymized accession number and cohort ID are combined to form the primary key. The dynamic schema of MongoDB allows us to store different data attributes in the same collection. Images are stored both as de-identified DICOM files and 16-bit PNG files in a hierarchical folder structure by cohort, patient, examination, and then image. Filenames were hashed so each filename is unique and can be linked directly to its DICOM metadata.

Discussion

We describe the curation of a dataset of 3 383 659 2D and DBT screening and diagnostic mammograms for 116 000 patients with equal representation for African American and White women. Uniquely, the dataset contains 40 000 lesion-specific ROIs with imaging characteristic and ground truth pathologic outcomes. The dataset specifically addresses the limitations of current datasets, namely the lack of ethnic diversity, limitations in database size, image annotations, and pathologic information. Throughout the process of collecting, organizing, de-identifying, and consolidating the dataset, there were several challenges that may suggest areas for future innovation to increase adoption at other institutions.

One limitation to our study was that about 20% of the lesions were classified as ambiguous because the ROIs of patients with multiple imaging findings could not be automatically linked. Although the specific ROI–lesion classification linkages may be approximated, no solution seemed acceptable at the time of writing.

In addition, DICOM metadata extraction was designed such that each row in the resultant dataframe represents a single image and each column represents a metadata element and its value. However, DICOM metadata varies across manufacturer and model and was sometimes corrupted, resulting in nested values that generated more than 2000 metadata tags for a single file. The resulting dataframe was too large to store in memory, so we decided to retain only metadata present in at least 10% of files. Private metadata fields were dropped. This threshold is a customizable parameter during metadata extraction using Niffler.

We also encountered an obstacle during DICOM-to-PNG conversion because some images appeared low-contrast or washed out as a result of a difference in window-level mapping between GE and Hologic scanners. To address this, the manufacturer DICOM tag was read during PNG conversion. Hologic was converted using min-max windowing and GE was converted by applying the values of interest lookup table function provided in each image's metadata. If this process is replicated at another institution, care should be taken that images from different manufacturers are normalized appropriately by either using built-in normalization functions of Pydicom (24), DCMTK (33), or NumPy (34).

Ground truth for pathologic diagnosis currently relies on the pathologic specimen being obtained at Emory, which may not occur in all cases. Therefore, there may be some cases for which pathologic diagnoses, including cancers, are missed. We are exploring how to best extract these outcomes from state cancer registries to augment the dataset.

Finally, an ongoing challenge remains linkage of ROIs back to imaging and pathologic findings in the MagView database. Although ROIs can be automatically mapped for examinations with a single described finding per breast, this is not possible for examinations with multiple findings per breast. This warrants the development of a new heuristic, which may include automatic selection of the ROI based on coded information in MagView for the breast quadrant and depth, or by manual review.

In summary, this dataset will aid in the development and validation of DL models for breast cancer screening that perform equally across patient demographic characteristics and reduce

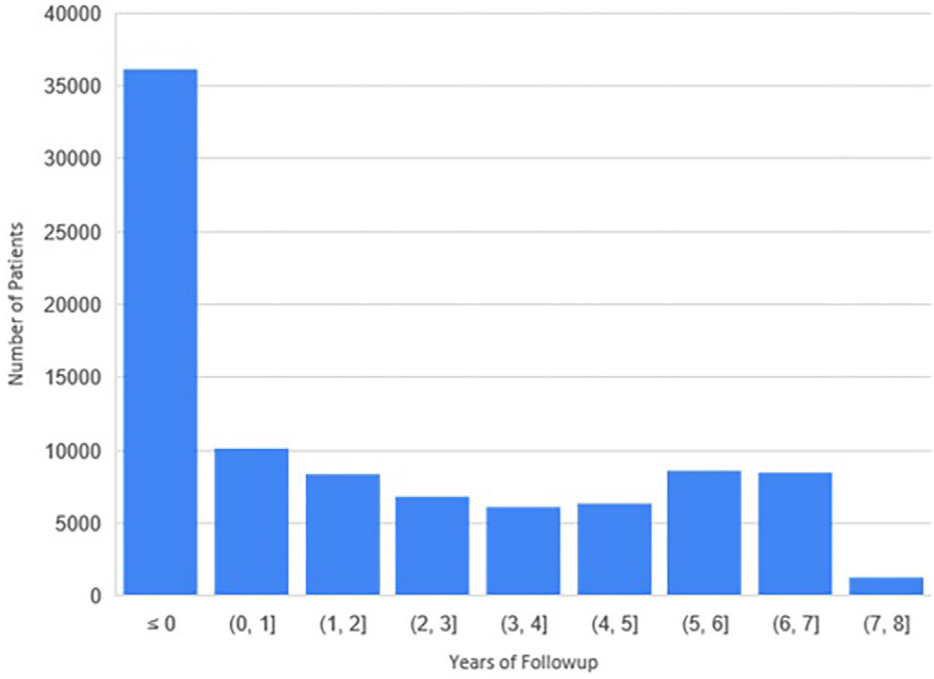

Figure 6: Distribution of the follow-up period available per patient. A total of 37 939 patients had at least 3 years of follow-up, and 24 933 had at least 5 years of follow-up.

Table 3: Sample of Imaging Findings for Training and Validation Datasets

BI-RADS Category	Screening	Diagnostic	Mass	Calcification	Asymmetry	Architectural Distortion
0	34 943	...	6523	7728	23 147	2312
1	167 174	13 776	5	0	132	16
2	23 289	25 960	10 562	7666	8243	434
3	...	17 053	2915	6106	5431	275
4	...	6860	2742	4261	2184	690
5	...	649	911	342	123	60
6	...	1088	659	249	208	52

Note.—Data are numbers of sample imaging findings. Findings categorized broadly by masses, calcifications, asymmetries, and architectural distortions. More detailed information is available, as shown in Table 1. Information regarding findings distributions in the test set is withheld. BI-RADS = Breast Imaging Reporting and Data System.

disparities in health care. To date, EMBED has been used in two research validation studies for breast cancer risk prediction (35,36) and several commercial model validations. Permission for external collaboration by research and industry partners are reviewed on a case-by-case basis by the institutional review board. Following institutional review board approval, we have released 20% of the dataset on the Amazon Web Services Open Data Program *(https://registry.opendata.aws/emory-breast-imaging-dataset-embed/)*, allowing researchers to review the structure and content of EMBED before deciding whether to carry out an analysis on the full dataset.

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request pending permission from the corresponding institutional review board.

Author contributions: Guarantors of integrity of entire study, **J.J.J., A.B., M.M., G.O., H.T.**; 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, **B.L.V., T.S., G.O., G.S., M.W., C.R.M., I.B., J.G.**; clinical studies, **J.J.J., G.O., G.S., C.R.M., M.S.N.**; experimental studies, **J.J.J., R.C., R.D., C.R.M., I.B.**; statistical analysis, **J.J.J., A.B., T.S., I.B., J.G., H.T.**; and manuscript editing, **B.L.V., A.B., T.S., R.C., G.O., M.W., I.B., J.G., H.T.**

Disclosures of conflicts of interest: J.J.J. No relevant relationships. **B.L.V.** No relevant relationships. **A.B.** No relevant relationships. **T.K.** No relevant relationships. **T.S.** No relevant relationships. **R.C.** No relevant relationships. **R.D.** No relevant relationships. **M.M.** No relevant relationships. **G.O.** No relevant relationships. **G.S.** No relevant relationships. **M.W.** No relevant relationships. **C.R.M.** No relevant relationships. **M.S.N.** No relevant relationships. **I.B.** No relevant relationships. **J.G.** NSF and NIH grants; Nightingale Foundation grant; ACR AI Advisory group; SIIM director at large; HL7 board member; trainee editorial board member lead and associate editor for *Radiology: Artificial Intelligence*. **H.T.** Kheiron Medical Technologies, academic-industry collaboration; consultant for Arterys, Sirona Medical, and BioData Consortium; owner of Lightbox AI.

Table 4: Region of Interest Counts by Pathologic Outcome for Training Datasets

Pathologic Outcome	Patients by Category	Total Screening ROIs	Total Diagnostic ROIs	ROIs Linked Directly to Finding
All ROIs	32 514	29 968	2538	25 873
Invasive breast cancer	1765	1383	382	1130
In situ cancer	845	687	158	602
High-risk lesion	1146	849	297	778
Borderline lesion	24	24	0	16
Benign	3281	2625	656	2289
Nonbreast cancer	46	19	27	12

Note.—Data are counts. Approximately 80% of regions of interest (ROIs) could be directly linked to imaging findings and pathologic outcomes. The two most common ROIs are for benign findings, followed by invasive cancers. Information regarding pathologic findings and ROIs in the test set is withheld.

References

1. Tadavarthi Y, Vey B, Krupinski E, et al. The state of radiology AI: considerations for purchase decisions and current market offerings. Radiol Artif Intell 2020;2(6):e200004.
2. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. Expert Rev Med Devices 2019;16(5):351–362.
3. Batchu S, Liu F, Amireh A, Waller J, Umair M. A review of applications of machine learning in mammography and future challenges. Oncology 2021;99(8):483–490.
4. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020;3(3):e200265.
5. Halling-Brown MD, Warren LM, Ward D, et al. OPTIMAM mammography image database: a large scale resource of mammography images and clinical data. arXiv:2004.04742 [preprint] https://arxiv.org/abs/2004.04742. Posted April 9, 2020. Accessed November 2021.
6. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data 2017;4:170177.
7. Buda M, Saha A, Walsh R, et al. Breast Cancer Screening – Digital Breast Tomosynthesis (BCS-DBT) [Data set]. The Cancer Imaging Archive. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=64685580. Published 2020. Accessed November 2021.
8. Yedjou CG, Sims JN, Miele L, et al. Health and racial disparity in breast cancer. Adv Exp Med Biol 2019;1152:31–49.
9. Newman LA, Kaljee LM. Health disparities and triple-negative breast cancer in African American women: a review. JAMA Surg 2017;152(5):485–493.
10. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94. [Published correction appears in Nature 2020;586(7829):E19.]
11. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15(11):e1002683.
12. Pan I, Agarwal S, Merck D. Generalizable inter-institutional classification of abnormal chest radiographs using efficient convolutional neural networks. J Digit Imaging 2019;32(5):888–896.
13. Thrall JH, Li X, Li Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol 2018;15(3 Pt B):504–508.
14. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447–453.
15. Noor P. Can we trust AI not to further embed racial bias and prejudice? BMJ 2020;368:m363.
16. Aizer AA, Wilhite TJ, Chen M-H, et al. Lack of reduction in racial disparities in cancer-specific mortality over a 20-year period. Cancer 2014;120(10):1532–1539.
17. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178(11):1544–1547.
18. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–118. [Published correction appears in Nature 2017;546(7660):686.]
19. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature 2011;475(7355):163–165.
20. Halling-Brown MD, Looney PT, Patel MN, Warren LM, Mackenzie A, Young KC. The oncology medical image database (OMI-DB). In: Law MY, Cook TS, eds. Medical Imaging 2014: PACS and Imaging Informatics: Next Generation and Innovations. Vol 9039. SPIE Proceedings. SPIE; 2014; 903906.
21. Sickles EA, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS Mammography. In: ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. 5th ed. American College of Radiology, 2013; 134–136.
22. Mammography Quality Standards Act and Program. U.S. Food and Drug Administration. https://www.fda.gov/radiation-emitting-products/mammography-quality-standards-act-and-program. Accessed August 13, 2021.
23. Kathiravelu P, Sharma P, Sharma A, et al. A DICOM framework for machine learning and processing pipelines against real-time radiology images. J Digit Imaging 2021;34(4):1005–1013.
24. Mason D. SU-E-T-33: Pydicom: An open source DICOM library. Med Phys 2011;38(6Part10):3493.
25. HITI-anon-internal. TestPyPI. https://test.pypi.org/project/HITI-anon-internal/. Accessed January 7, 2022.
26. Magny SJ, Shikhman R, Keppke AL. Breast Imaging, Reporting and Data System (BI RADS). StatPearls. January 2018. https://www.ncbi.nlm.nih.gov/books/NBK459169/. Accessed November 2021.
27. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [preprint] https://arxiv.org/abs/1409.1556. Posted September 4, 2014. Accessed November 2021.
28. GitHub. EMBED Codebase. https://github.com/Emory-HITI/Mammo. Accessed February 6, 2022.
29. Tan M, Pang R, Le QV. Efficientdet: scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020; 10778–10787.
30. Beare R, Lowekamp B, Yaniv Z. Image segmentation, registration and characterization in R with SimpleITK. J Stat Softw 2018;86:8.
31. El Boukkouri H, Ferret O, Lavergne T, Noji H, Zweigenbaum P, Tsujii J. CharacterBERT: reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, 2020; 6903–6915.
32. Plugge E, Membrey P, Hawkins T. Introduction to MongoDB. In: The Definitive Guide to Mongodb. Apress, 2010; 3–17.
33. dicom.offis.de - DICOM Software made by OFFIS - DCMTK - DICOM Toolkit. https://dicom.offis.de/dcmtk.php.en. Accessed January 10, 2022.
34. Oliphant TE. A guide to NumPy, Vol 1. Trelgol Publishing, 2006.
35. Yala A, Mikhael PG, Strand F, et al. Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. J Clin Oncol 2022;40(16):1732–1740.
36. Yala A, Mikhael PG, Lehman C, et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. Nat Med 2022;28(1):136–143.