SCIENTIFIC DATA

OPEN

Data Descriptor: A curated mammography data set for use in computer-aided detection and diagnosis research

Received: 5 July 2016

Accepted: 9 October 2017

Published: 19 December 2017

Rebecca Sawyer Lee[1], Francisco Gimenez[1], Assaf Hoogi[2], Kanae Kawai Miyake[3], Mia Gorovoy[3] & Daniel L. Rubin[2]

Published research results are difficult to replicate due to the lack of a standard evaluation data set in the area of decision support systems in mammography; most computer-aided diagnosis (CADx) and detection (CADe) algorithms for breast cancer in mammography are evaluated on private data sets or on unspecified subsets of public databases. This causes an inability to directly compare the performance of methods or to replicate prior results. We seek to resolve this substantial challenge by releasing an updated and standardized version of the Digital Database for Screening Mammography (DDSM) for evaluation of future CADx and CADe systems (sometimes referred to generally as CAD) research in mammography. Our data set, the CBIS-DDSM (Curated Breast Imaging Subset of DDSM), includes decompressed images, data selection and curation by trained mammographers, updated mass segmentation and bounding boxes, and pathologic diagnosis for training data, formatted similarly to modern computer vision data sets. The data set contains 753 calcification cases and 891 mass cases, providing a data-set size capable of analyzing decision support systems in mammography.

Design Type(s)	parallel group design • feature extraction objective • image processing objective
Measurement Type(s)	Mammography
Technology Type(s)	digital curation
Factor Type(s)	diagnosis
Sample Characteristic(s)	Homo sapiens • breast

[1]Biomedical Informatics Training Program, Stanford University, Stanford, CA 94305, USA. [2]Department of Radiology and Medicine (Biomedical Informatics Research), Stanford University, Stanford, CA 94305, USA. [3]Department of Radiology (Breast Imaging), Stanford University, Stanford, CA 94305, USA. Correspondence and requests for materials should be addressed to R.S.L. (email: rebeccalslee15@gmail.com) or to D.L.R. (email: dlrubin@stanford.edu).

Background & Summary

Computer-aided detection (CADe) and diagnosis (CADx) systems are designed to assist radiologists for mammography interpretation. CADe is employed to discover abnormal structures within the mammogram while CADx is used to determine the significance of the discovered abnormality. Despite promising results, current CADe systems are limited by high false-positive rates[1], and CADx systems for mammography are not yet approved for clinical use[2]. Although the technical difficulty of CAD in mammography has been substantial, there is another obstacle that must be addressed to enable this research: decision support system evaluation.

Our review of the CAD literature reveals inconsistent data sources and data-set sizes. In addition, only few of the published results can be reproduced directly, as most evaluation data are not public. Tables 1 and 2 contain a sample of many systems, CADe and CADx, respectively, that have been evaluated using private data sets or undefined portions of public data sets. Without common data sets, it is impossible to rigorously compare methods. This is hindering CAD research in mammography. We seek to address this challenge by providing a standard data set for evaluation.

The non-medical computer vision community has adopted an open research approach. This includes provision of standard data sets for evaluation of algorithms. For example, ImageNet is a database of 14,197,122 images from 27 'high-level' categories including animals, food, and vehicles. Each category has at least 51 sub-categories, allowing for highly specific classifications[3]. Other public databases include Mixed National Institute of Standards and Technology (MNIST) database, a database of hand-written digits[4], and Caltech 256, a database of 265 object categories such as helicopters, planes, motorbikes, and school buses[5]. These data sets and others like them have provided a benchmark for computer vision research. Many researchers point to the existence of these open data sets as the primary drivers for recent successes in image classification technologies, such as deep learning.

Conversely, few well-curated public data sets have been provided for the mammography community. These include the Digital Database for Screening Mammography (DDSM)[6], the Mammographic Imaging Analysis Society (MIAS) database[7], and the Image Retrieval in Medical Applications (IRMA) project[8]. Although these public data sets are useful, they are limited in terms of data set size and accessibility. For example, most researchers using the DDSM do not leverage all its images for a variety of historical reasons. When the database was released in 1997, computational resources to process hundreds or thousands of images were not widely available. Additionally, the DDSM images are saved in non-standard compression files that require the use of decompression code that has not been updated or maintained for modern computers. Finally, the region-of-interest (ROI) annotations for the abnormalities in the DDSM were provided to indicate a general position of lesions, but not a precise segmentation for them. Therefore, many researchers must implement segmentation algorithms for accurate feature extraction.

While there are substantial challenges in using the DDSM for method evaluation, due to its size and other unique characteristics, we believe that it can still be a powerful resource for imaging research. The DDSM is a database of 2,620 scanned film mammography studies. It contains normal, benign, and malignant cases with verified pathology information. The scale of the database along with ground truth validation makes the DDSM a useful tool in the development and testing of decision support systems despite the fact that the images are scanned film instead of full field digital mammograms. This is because there is currently no mammography database of this size publicly available. We report here the development of, and propose to release, the CBIS-DDSM (Curated Breast Imaging Subset of DDSM), an updated version of the DDSM providing easily accessible data and improved ROI segmentation. This resource will contribute to the advancement of decision support system research in mammography, supplying a standardized mammography data.

Performance statistics of selected CADe methods for the detection of abnormalities					
Authors	**Size of Data set (Cases)**	**Public or Private Data**	**Accuracy**	**Sensitivity**	**False Positives Per Image**
Karssemeijer and te Brake[13]	50	Public (MIAS*)	NA	90%	1
Mudigonda *et al.*[14]	56	Public (MIAS*)	NA	81%	2.2
Liu *et al.*[15]	38	Public (MIAS*)	NA	90%	1
Li *et al.*[16]	94	Private	NA	91%	3.21
Baum *et al.*[17]	63	Private	NA	89%	0.61
Kim *et al.*[18]	83	Private	NA	96%	0.2
Yang *et al.*[19]	203	Private	96.1%	95–98%	1.8
The *et al.*[20]	123	Private	NA	94%	2.3 per case
Sadaf *et al.*[21]	127	Private	NA	91%	NA
Chu *et al.*[22]	230	Public (DDSM†)	NA	98.5%	0.84

Table 1. **Sample Set of CADe Systems Reported in the Literature.** *Mammographic Imaging Analysis Society. †Digital Database for Screening Mammography.

Performance statistics of selected CADx methods for the classification of masses				
Authors	**Size of Data set (Cases)**	**Public or Private Data**	**Classification Accuracy**	**Az***
Brzakovic *et al.*[23]	25	Private	85%	NA
Huo *et al.*[24]	65	Private	NA	0.94
Rangayyan *et al.*[25]	54	Public (MIAS†) and Private	91%	NA
Mudigonda *et al.*[26]	39	Public (MIAS†)	82.1%	0.85
Sahiner *et al.*[27]	102	Private	NA	0.91
Timp *et al.*[28]	465	Private	NA	0.77
Ganesan *et al.*[29]	282	Private	88.8%	NA
Görgel *et al.*[30]	78, 65	Private, Public (MIAS†)	91.4%, 90.1%	NA
Qiu *et al.*[31]	560	Private	77.14%	0.81
Choi *et al.*[32]	600	Public (DDSM‡)	NA	0.88

Table 2. **Sample Set of CADx Systems Reported in the Literature.** *Area under the receiver-operator characteristic curve. †Mammographic Imaging Analysis Society. ‡Digital Database for Screening Mammography.

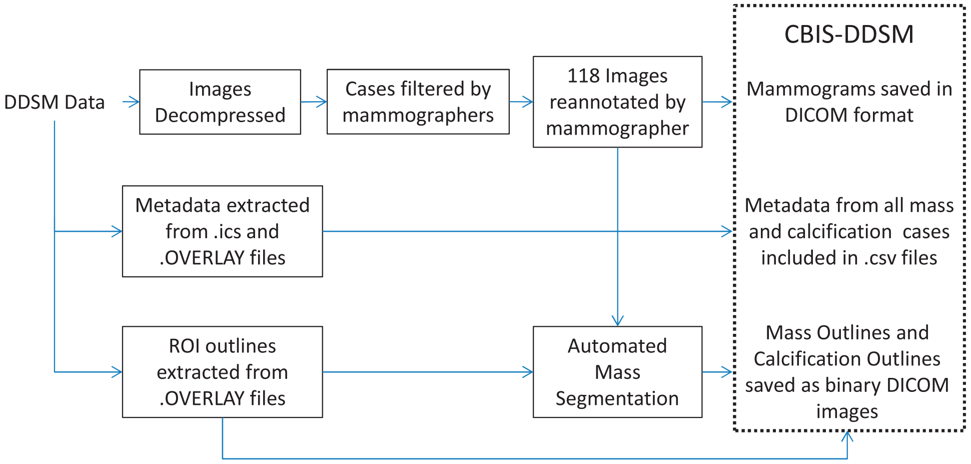
DDSM Data
Images Decompressed
Cases filtered by mammographers
118 Images reannotated by mammographer
CBIS-DDSM
Mammograms saved in DICOM format
Metadata extracted from .ics and .OVERLAY files
Metadata from all mass and calcification cases included in .csv files
ROI outlines extracted from .OVERLAY files
Automated Mass Segmentation
Mass Outlines and Calcification Outlines saved as binary DICOM images

Figure 1. **Flow diagram of preparation of CBIS-DDSM.**

Methods

The DDSM already contains a large amount of information for each of its 2,620 cases. However, some information is limited, specifically the ROI annotations, while other information is difficult to access. We have solved these issues by updating the ROI segmentations and by gathering and reformatting the metadata into a more accessible format. Figure 1 shows a diagram of the processes performed to prepare the data set: image decompression and reannotation and metadata extraction and reformatting.

Description of DDSM

The DDSM is a collection of mammograms from the following sources: Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital, and Washington University of St Louis School of Medicine. The DDSM was developed through a grant from the DOD Breast Cancer Research Program, US Army Research and Material Command, and the necessary patient consents were obtained by the original developers of the DDSM[6]. The cases are annotated with ROIs for calcifications and masses, as well as the following information that may be useful for CADe and CADx algorithms: Breast Imaging Reporting and Data System (BI-RADS) descriptors for mass shape, mass margin, calcification type, calcification distribution, and breast density; overall BI-RADS assessment from 0 to 5; rating of the subtlety of the abnormality from 1 to 5; and patient age.

Parse semantic features

DDSM provides metadata in the form of.ics files. These files include the patient age, the date of the study, as well as the date of digitization, the dense tissue category, the scanner used to digitize, and the resolution of each image. Additionally, those cases with abnormalities have.OVERLAY files that contain information about each abnormality, including type of abnormality (mass or calcification) and the

BI-RADS descriptors mentioned above. These metadata have been extracted and compiled into a single comma-separated values (CSV) file.

Removal of questionable mass cases

It has been noted by other researchers that not all DDSM ROI annotations are accurate[9], and we found that some annotations indicate suspicious lesions that are not seen in the image. Due to this issue, we acquired the assistance of a trained mammographer who reviewed the questionable cases. In this process, we found 339 images in which a mass was not clearly seen. These images were removed from the final data set. Additionally, several cases were removed by TCIA due to personal health information in the images.

Image decompression

DDSM images are distributed as lossless joint photographic experts group (JPEG) files (LJPEG), an obsolete image format. The only library capable of decompressing these images is the Stanford PVRG-JPEG Codec v1.1, which was last updated in 1993. We modified the PVRG-JPEG codec to successfully compile on an OSX 10.10.5 (Yosemite) distribution using Apple GCC clang-602.0.53. The original decompression code outputs data in 8-bit or 16-bit raw binary bitmaps[6]. We wrote python tools to read this raw data and store it as 16-bit gray scale Tag Image File Format (TIFF) files. These files were later converted to Digital Imaging and Communications in Medicine (DICOM) format, which is standard for medical images. This process is entirely lossless and preserved all information from the original DDSM files.

Image processing

The original DDSM files were distributed with a set of bash and C tools for Linux to perform image correction and metadata processing. These tools were very difficult to refactor for use on modern systems. Instead, we re-implemented these tools in Python to be cross-platform and easy to understand for modern users.

All images in the DDSM were derived from several different scanners at different institutions. The DDSM data descriptions provide methods to convert raw pixel data into 64-bit optical density values, which are standardized across all images. Optical density values were then re-mapped to 16-bit gray scale TIFF files and later converted to DICOM format for the data repository.

The DDSM automatically clips optical density values to be between 0.05 and 3.0 for noise reduction. We perform this clipping as well, but provide a flag to remove the clipping and retain the original optical density values.

Image cropping

Several CAD tasks require only analyzing abnormalities (the portion of the image in the ROI) without needing the full mammogram image. We provide a set of convenience images, which are focused crops of abnormalities. Abnormalities were cropped by determining the bounding rectangle of the abnormality with respect to its ROI.

Mass segmentation

Mass margin and shape have long been proven substantial indicators for diagnosis in mammography. Because of this, many methods are based on developing mathematical descriptions of the tumor outline. Due to the dependence of these methods on accurate ROI segmentation and the imprecise nature of many of the DDSM-provided annotations, as seen in Fig. 2, we applied a lesion segmentation algorithm (described below) that is initialized by the general original DDSM contours but is able to supply much more accurate ROIs. Figure 2 contains example ROIs from the DDSM, our mammographer, and the automated segmentation algorithm. As shown, the DDSM outlines provide only a general location and not a precise mass boundary. The segmentation algorithm was designed to provide exact delineation of the mass from the surrounding tissue. This segmentation was done only for masses and not calcifications.

Lesion segmentation was accomplished by applying a modification to the local level set framework as presented in Chan and Vese[10–12]. Level set models follow a non-parametric deformable model, thus can handle topological changes during evolution. Chan-Vese model is a region-based method that estimates spatial statistics of image regions and finds a minimal energy where the model best fits the image, resulting in convergence of the contour towards the desired object. Our modification of the local framework includes automated evaluation of the local region surrounding each contour point. For low contrast lesions, small local region is determined, and excessive curve evolution is thus prevented. On the other hand, for noisy or heterogeneous lesions, a relatively large local region is assigned to the contour point to prevent convergence of the level set contour into local minima. Local frameworks require an initialization of the contour, and thus in our case the original DDSM annotation was used as the level set segmentation initialization.

Standardized train/test splits

Separate sets of cases for training and testing algorithms are important for ensuring that all researchers are using the same cases for these tasks. Specifically, the test set should contain cases of varying difficulty

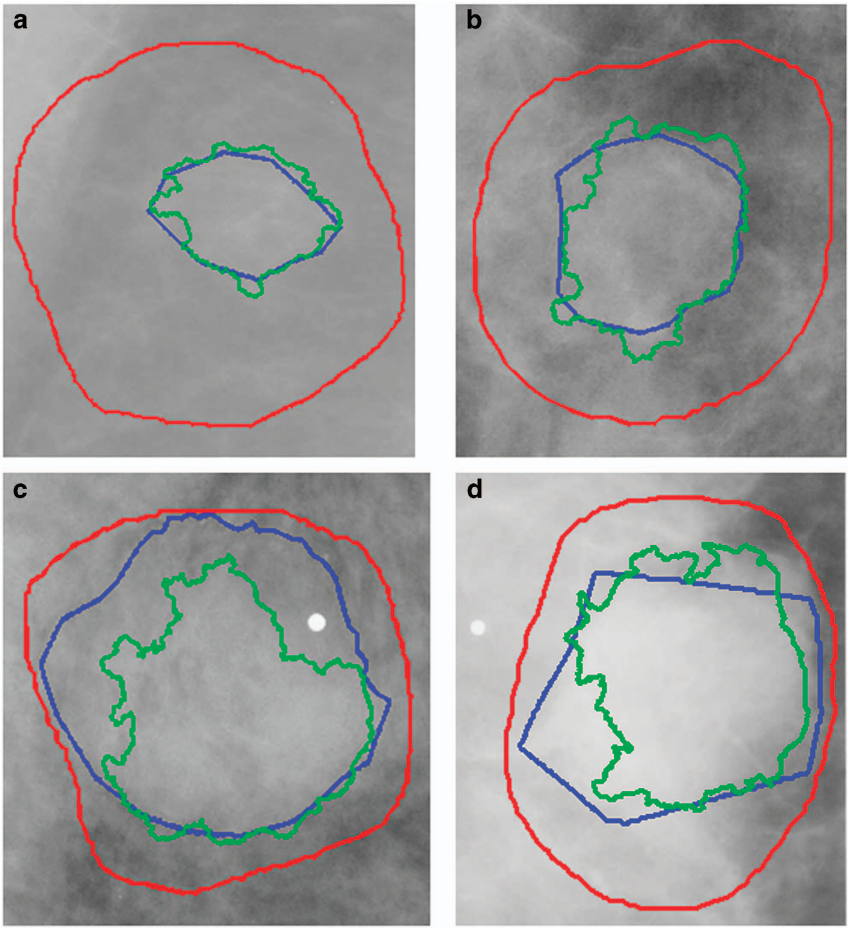

Figure 2. Example ROI outlines from each of the four BI-RADS density categories. The Dice's coefficients are provided for each. (**a**) Density 1 ROI, $D_{H,C}=0.904$, $D_{H,D}=0.237$, (**b**) Density 2 ROI, $D_{H,C}=0.886$, $D_{H,D}=0.423$, (**c**) Density 3 ROI, $D_{H,C}=0.749$, $D_{H,D}=0.797$, (**d**) Density 4 ROI, $D_{H,C}=0.808$, $D_{H,D}=0.682$. Outlines: DDSM (red), hand-drawn (blue), automated (green).

in order to ensure that the method is tested thoroughly. The data were split into a training set and a testing set based on the BI-RADS category. This allows for an appropriate stratification for researchers working on CADe as well as CADx. Note that many of the BI-RADS assessments likely were updated after additional information was gathered by the physician, as it is unconventional to subscribe BI-RADS 4 and 5 to screening images. The split was obtained using 20% of the cases for testing and the rest for training. The data were split for all mass cases and all calcification cases separately. Here 'case' is used to indicate a particular abnormality, seen on the craniocaudal (CC) and/or mediolateral oblique (MLO) views, which are the standard views for screening mammography. Figure 3 displays the histograms of BI-RADS assessment and pathology for the training and test sets for calcification cases and mass cases. As shown, the data split was performed in such a way to provide equal level of difficulty in the training and test sets. Table 3 contains the number of benign and malignant cases for each set.

Code availability

The code used to generate these data sets from a raw dump of the DDSM was written in Python v2.7.9. It is publicly available as a git repository on GitHub at github.com/fjeg/ddsm_tools.

Data Records

The images are distributed at the full mammography and abnormality level as DICOM files. Full mammography images include both MLO and CC views of the mammograms.

Abnormalities are represented as binary mask images of the same size as their associated mammograms. These mask images delineate the ROI of each abnormality. Users can perform an element-wise selection of pixels within an abnormality mask that was created for each mammogram. For convenience of abnormality analysis, we have also distributed images containing just the abnormalities cropped by the bounding box of their ROI. Abnormality files have been separated into Java Network

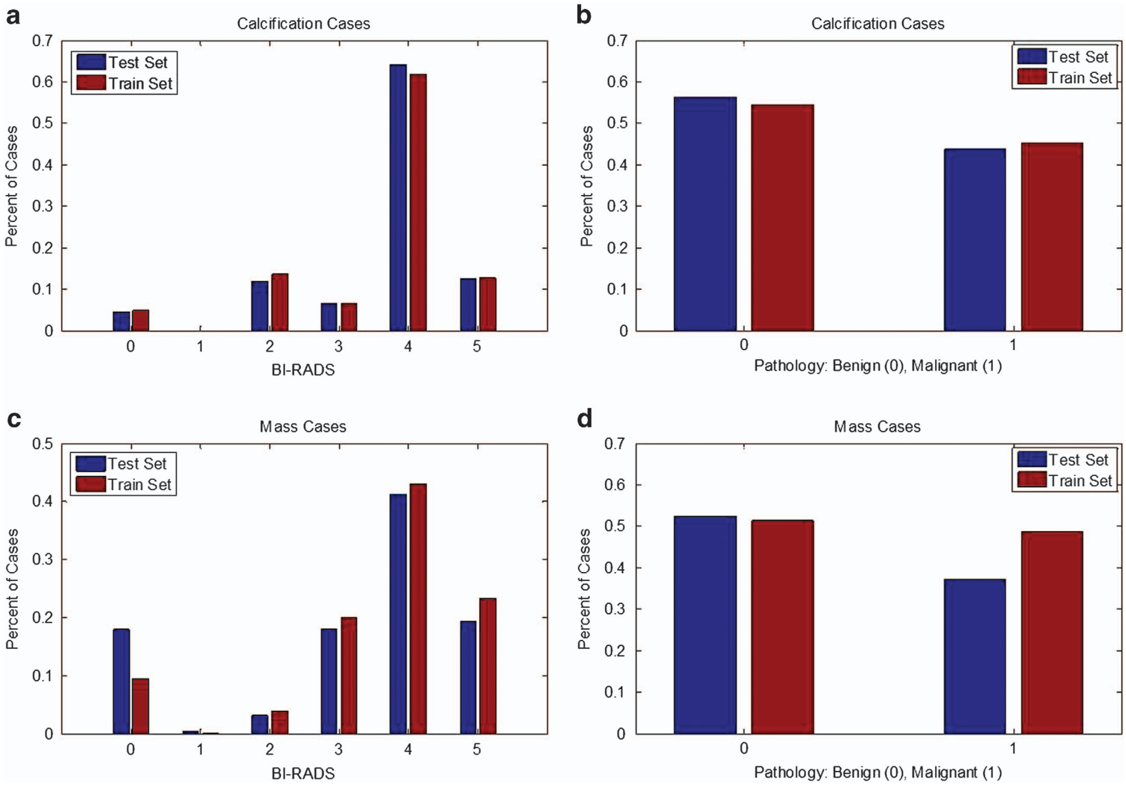

Figure 3. Histograms showing distribution of reading difficulty for training and test sets. Mass and calcification cases were split into training and test sets based on BI-RADS assessment. (**a**) Histogram of BIRADS for each abnormality in training and test sets with calcifications, (**b**) Histogram of benign and malignant cases for training and test sets with calcifications, (**c**) Histogram of BIRADS for each abnormality in training and test sets with masses, (**d**) Histogram of benign and malignant cases for training and test sets with masses.

Launch Protocol (JNLP) files based on abnormality type, training or test set, and image type (full mammogram, ROI mask, or cropped mammogram).

The following files contain the mammograms and ROIs for the cases with calcifications:

- Calc-Training_full_mammogram_images_1-doiJNLP-PrQ05L6k.jnlp
- Calc-Training_ROI-mask_and_crpped_images-doiJNLP-kTGQKqBk.jnlp
- Calc-Test_full_mammogram_images-doiJNLP-SiXj6kpS.jnlp
- Calc-Test_ROI-mask_and_crpped_images-doiJNLP-PsjCfTdf.jnlp

The following files contain the mammograms and ROIs for the cases with calcifications:

- Mass-Training_full_mammogram_images_1-doiJNLP-wv6aeYDn.jnlp
- Mass-Training_ROI-mask_and_crpped_images_1-doiJNLP-07gmVj4b.jnlp
- Mass-Test_full_mammogram_images-doiJNLP-6ccCrb8t.jnlp
- Mass-Test_ROI-mask_and_crpped_images-doiJNLP-SmEOyQFn.jnlp

Note that there is some overlap since some cases contain both calcifications and masses. Metadata for each abnormality is included as an associated CSV file containing the following:

- Patient ID: the first 7 characters of images in the case file
- Density category
- Breast: Left or Right
- View: CC or MLO
- Number of abnormality for the image (This is necessary as there are some cases containing multiple abnormalities.
- Mass shape (when applicable)

	Benign Cases	Malignant Cases
Calcification Training Set	329 cases (552 abnormalities)	273 cases (304 abnormalities)
Calcification Test Set	85 cases (112 abnormalities)	66 cases (77 abnormalities)
Mass Training Set	355 cases (387 abnormalities)	336 cases (361 abnormalities)
Mass Test Set	117 cases (135 abnormalities)	83 cases (87 abnormalities)

Table 3. Number of Cases and Abnormalities in the Training and Test Sets. These numbers are different as some cases contain more than one abnormality.

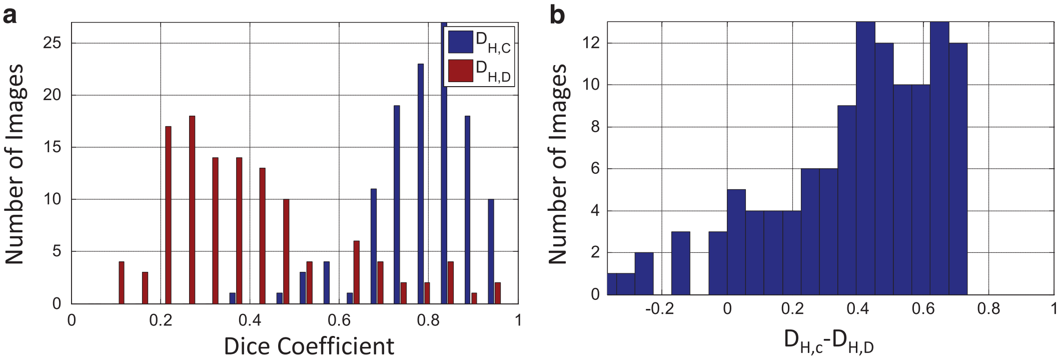

Figure 4. Results from comparison of new hand-drawn outlines with computer-generated and original DDSM outlines. (**a**) Histogram of Dice's coefficients for new hand-drawn and computer-generated segmentation ($D_{H,C}$) and new hand-drawn and DDSM ($D_{H,D}$). The Dice's coefficients were computed for 118 images. The mean $D_{H,C}$ is 0.792 ± 0.108, and the mean $D_{H,D}$ is 0.398 ± 0.195. (**b**) Histogram of $D_{H,C} - D_{H,D}$. The mean difference in Dice's coefficient was 0.395 ± 0.257. The Wilcoxon signed rank test between $D_{H,C}$ and $D_{H,D}$ yielded a *P*-value of 5.54 × 10 − 19.

- Mass margin (when applicable)
- Calcification type (when applicable)
- Calcification distribution (when applicable)
- BI-RADS assessment
- Pathology: Benign, Benign without call-back, or Malignant
- Subtlety rating: Radiologists' rating of difficulty in viewing the abnormality in the image
- Path to image files

There are individual files for mass and calcification training and test sets:

- mass_case_description_train_set.csv
- mass_case_description_test_set.csv
- calc_case_description_train_set.csv
- calc_case_description_test_set.csv

All these files are available via Data Citation 1.

Technical Validation

The details of data collection may be found at the primary DDSM website6. We have improved the quality of this data set by distributing the data in a more accessible format, specifically as decompressed images and updated metadata extraction code, as well as by providing improved ROI segmentation and training and testing splits for evaluation. The methods for the validation of the ROI segmentation are given below.

Segmentation evaluation

Ideally, we would provide hand-drawn segmentations for each mass lesion in CBIS-DDSM. However, this would be a prohibitively large task to accomplish. We thus used an automated segmentation algorithm with the goal of providing better segmentations than those available in the DDSM. Since the segmentations we provide for the mass lesions in CBIS-DDSM were generated by our automated algorithm, we evaluated the segmentations in CBIS-DDSM by comparing ROIs from 118 images in CBIS-DDSM with hand-drawn outlines that were provided by an experienced radiologist. We computed the

Dice coefficients between the computer and hand-drawn ROIs, $D_{H,C}$. This was compared to the Dice coefficients between the original DDSM annotations and the newly hand-drawn annotations, $D_{H,D}$. The Dice coefficient is a common metric for validity of image segmentation. Additionally, we use the Wilcoxon signed rank test to determine the statistical significance of $D_{H,C}$ as compared to $D_{H,D}$. We utilized this test instead of a standard t-test due to the non-normal distributions of Dice coefficients, as seen in Fig. 4. Finally, we examined the histogram and statistics of the difference between $D_{H,C}$ and $D_{H,D}$.

The average $D_{H,C}$ was 0.792 ± 0.108, and the average $D_{H,D}$ was 0.398 ± 0.195. Figure 4 shows the histogram of $D_{H,C}$ and $D_{H,D}$ for each of the 118 new hand-drawn images. The histogram shows that the majority of the images were automatically segmented with high correlation to the hand-drawn ROIs. The Wilcoxon signed rank test between $D_{H,C}$ and $D_{H,D}$ yielded a P-value of 5.54×0^{-19}. The histogram in Fig. 4 shows the difference between $D_{H,C}$ and $D_{H,D}$. The mean difference in Dice's coefficient was 0.395 ± 0.257.

Figure 2 contains example ROIs from each of the BI-RADS density categories represented in the data set: (1) almost entirely fat, (2) containing scattered areas of fibroglandular density, (3) heterogeneously dense, and (4) extremely dense. The red outline indicates the original DDSM ROI, the blue is the new hand-drawn ROI, and the green is the automatically segmented ROI. The $D_{H,C}$ for each of these ROIs with respect to the newly hand-drawn annotations were 0.904, 0.886, 0.749, and 0.808, and the $D_{H,D}$ were 0.237, 0.423, 0.797, and 0.682, respectively. As expected, the accuracy of the computer method decreases with increase in breast density. Though $D_{H,C}$ and $D_{H,D}$ are comparable in some cases, particularly in higher density cases, the computer-generated ROIs were in much higher agreement with the hand-drawn ROIs and are overall much better segmentations.

References

1. Nishikawa, R. M. & Gur, D. CADe for early detection of breast cancer-current status and why we need to continue to explore new approaches. *Acad. Radiol.* **21,** 1320–1321 (2014).
2. Doi, K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Comput. Med. Imaging Graph.* **31,** 198–211 (2007).
3. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
4. LeCun, Y., Cortes, C. & Burges, C. MNIST handwritten digit database (1998). Available at http://yann.lecun.com/exdb/mnist/ (Accessed: 29th September 2015).
5. Griffin, G., Holub, A. & Perona, P. Caltech-256 object category dataset (2007). Available at http://www.vision.caltech.edu/Image_Datasets/Caltech256/(Accessed: 29th September 2015).
6. Heath, M., Bowyer, K., Kopans, D., Moore, R. & Kegelmeyer, W. P. The Digital Database for Screening Mammography. *Proceedings of the Fifth International Workshop on Digital Mammography* 212–218 (2001). Available at http://marathon.csee.usf.edu/Mammography/software/HeathEtAlIWDM_2000.pdf (Accessed: 29th September 2015).
7. Suckling, J. *et al.* The Mammographic Image Analysis Society digital mammogram database. *Exerpta Medica* 375–378 (1994). Available at http://peipa.essex.ac.uk/info/mias.html (Accessed: 29th September 2015).
8. Lehmann, T. M. *et al.* IRMA—Content-based image retrieval in medical applications. *Methods Inf. Med.* **43,** 354–361 (2004).
9. Song, E. *et al.* Breast mass segmentation in mammography using plane fitting and dynamic programming. *Acad. Radiol.* **16,** 826–835 (2009).
10. Chan, T. F. & Vese, L. A. Active contours without edges. *IEEE Trans. Image Process* **10,** 266–277 (2001).
11. Hoogi, A. *et al.* Adaptive local window for level set segmentation of CT and MRI liver lesions. *Med. Image Anal.* **36,** 47–55 (2017).
12. Hoogi, A., Subramaniam, A., Veerapaneni, R. & Rubin, D. L. Adaptive estimation of active contour parameters using convolutional neural networks and texture analysis. *IEEE Trans. Med. Imaging* **36,** 781–791 (2017).
13. Karssemeijer, N. & te Brake, G. M. Detection of stellate distortions in mammograms. *IEEE Trans. Med. Imaging* **15,** 611–619 (1996).
14. Mudigonda, N. R., Rangayyan, R. M. & Desautels, J. E. L. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. *IEEE Trans. Med. Imaging* **20,** 1215–1227 (2001).
15. Liu, S., Babbs, C. F. & Delp, E. J. Multiresolution detection of spiculated lesions in digital mammograms. *IEEE Trans. IMAGE Process* **10,** 874–884 (2001).
16. Li, L., Clark, R. A. & Thomas, J. A. Computer-aided diagnosis of masses with full-field digital mammography. *Acad. Radiol.* **9,** 4–12 (2002).
17. Baum, F., Fischer, U., Obenauer, S. & Grabbe, E. Computer-aided detection in direct digital full-field mammography : initial results. *Eur. Radiol* **12,** 3015–3017 (2002).
18. Kim, S. J. *et al.* Computer-aided detection in digital mammography: Comparison of craniocaudal, mediolateral oblique, and mediolateral views. *Radiology* **241,** 695–701 (2006).
19. Yang, S. K. *et al.* Screening mammography—detected cancers : Sensitivity of a computer-aided detection system applied to full-field digital mammograms. *Radiology* **244,** 104–111 (2007).
20. The, J. S., Schilling, K. J., Hoffmeister, J. W. & Mcginnis, R. Detection of breast cancer with full-field digital mammography and computer-aided detection. *Am. J. Roentgenol* **192,** 337–340 (2009).
21. Sadaf, A., Crystal, P., Scaranelo, A. & Helbich, T. Performance of computer-aided detection applied to full-field digital mammography in detection of breast cancers. *Eur. J. Radiol.* **77,** 457–461 (2011).
22. Chu, J., Min, H., Liu, L. & Lu, W. A novel computer aided breast mass detection scheme based on morphological enhancement and SLIC superpixel segmentation. *Med. Phys.* **42,** 3859–3869 (2015).
23. Brzakovic, D., Luo, X. M. & Brzakovic, P. An approach to automated detection of tumors in mammograms. *IEEE Trans. Med. Imaging* **9,** 233–241 (1990).
24. Huo, Z. *et al.* Automated computerized classification of malignant and benign masses on digitized mammograms. *Acad. Radiol.* **5,** 155–168 (1998).
25. Rangayyan, R. M., Mudigonda, N. R. & Desautels, J. E. Boundary modelling and shape analysis methods for classification of mammographic masses. *Med. Biol. Eng. Comput.* **38,** 487–496 (2000).
26. Mudigonda, N. R., Rangayyan, R. M. & Desautels, J. E. Gradient and texture analysis for the classification of mammographic masses. *IEEE Trans. Med. Imaging* **19,** 1032–1043 (2000).
27. Sahiner, B., Chan, H. P., Petrick, N., Helvie, M. A. & Hadjiiski, L. M. Improvement of mammographic mass characterization using spiculation measures and morphological features. *Med. Phys.* **28,** 1455–1465 (2001).

28. Timp, S., Varela, C. & Karssemeijer, N. Temporal change analysis for characterization of mass lesions in mammography. *IEEE Trans. Med. Imaging* **26,** 945–953 (2007).
29. Ganesan, K., Acharya, U. R., Chua, C. K., Min, L. C. & Abraham, T. K. Automated diagnosis of mammogram Images of breast cancer using discrete wavelet transform and spherical wavelet transform features: A comparative study. *Technol. Cancer Res. Treat.* **13,** 605–615 (2014).
30. Görgel, P., Sertbas, A. & Uçan, O. N. Computer-aided classification of breast masses in mammogram images based on spherical wavelet transform and support vector machines. *Expert Syst* **32,** 155–164 (2015).
31. Qiu, Y. *et al.* Computer-aided classification of mammographic masses using the deep learning technology: a preliminary study. in *SPIE, Medical Imaging 2016: Computer-Aided Diagnosis* **9785,** (2016).
32. Choi, J. Y., Kim, D. H., Plataniotis, K. N. & Ro, Y. M. Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography. *Expert Syst. Appl.* **46,** 106–121 (2016).

Data Citation

1. Lee, R. S., Gimenez, F., Hoogi, A. & Rubin, D. L. *The Cancer Imaging Archive*. http://dx.doi.org/10.7937/K9/TCIA.2016.7O02S9CY (2016).

Acknowledgements

We would like to thank Dr Jafi Lipson for her input in mass segmentation. This work was supported by the National Cancer Institute, National Institutes of Health, under Grants U01-CA142555 and U01-CA190214. This project has also been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views of policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Daniel Rubin had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Author Contributions

R.S.L. discovered the need for and the potential of DDSM as a standard mammography data set for evaluation of computer-aided diagnosis and detection methods. R.S.L. also wrote the majority of the paper, extracted and organized the metadata, converted images to DICOM, and performed the technical evaluation. F.G. discovered the need for and the potential of DDSM as a standard mammography data set for evaluation of computer-aided diagnosis and detection methods. F.G. also rewrote the image decompression code in Python and contributed to the writing and editing of the paper. A.H. developed the mass segmentation algorithm and contributed to the writing and editing of the paper. K.K.M. provided hand-drawn ROIs for our segmentation gold standard and reviewed the paper. M.G. for examined mammograms for annotations to unseen masses and reviewed the paper. D.L.R. directed the project, edited the paper, and is guarantor of the study.

Additional Information

Competing interests: The authors declare no competing financial interests.

How to cite this article: Lee, R. S. *et al.* A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* 4:170177 doi: 10.1038/sdata.2017.177 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.