Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM)

Reproduced from https://wiki.cancerimagingarchive.net

Summary

This CBIS-DDSM (Curated Breast Imaging Subset of DDSM) is an updated and standardized version of the **Digital Database for Screening Mammography (DDSM)**. 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. The CBIS-DDSM collection includes a subset of the DDSM data selected and curated by a trained mammographer. The images have been decompressed and converted to DICOM format. Updated ROI segmentation and bounding boxes, and pathologic diagnosis for training data are also included. A manuscript describing how to use this dataset in detail is available at https://www.nature.com/articles/sdata2017177.

Published research results from work in developing decision support systems in mammography are difficult to replicate due to the lack of a standard evaluation data set; 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. Few well-curated public datasets have been provided for the mammography community. These include the DDSM, the Mammographic Imaging Analysis Society (MIAS) database, and the Image Retrieval in Medical Applications (IRMA) project. 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 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. This causes an inability to directly compare the performance of methods or to replicate prior results. The CBIS-DDSM collection addresses that challenge by publicly releasing an curated and standardized version of the DDSM for evaluation of future CADx and CADe systems (sometimes referred to generally as CAD) research in mammography.

Please note that the image data for this collection is structured such that each participant has multiple patient IDs. For example, participant 00038 has 10 separate patient IDs which provide information about the scans within the IDs (e.g. Calc-Test_P_00038_LEFT_CC, Calc-Test_P_00038_RIGHT_CC_1). This makes it appear as though there are 6,671 patients according to the DICOM metadata, but there are only 1,566 actual participants in the cohort.

For scientific and other inquiries about this dataset, please contact the TCIA Helpdesk.

Data Access

Data Access

Data Type	Download all or Query/Filter	License
Images (DICOM, 163.6GB)	*Download* *Search* (Download requires the NBIA Data Retriever)	CC BY 3.0
Mass-Training-Description (csv, 755 kB)	*Download*	CC BY 3.0
Calc-Training-Description (csv, 904 kB)	*Download*	CC BY 3.0
Mass-Test-Description (csv, 212 kB)	*Download*	CC BY 3.0

Reproduced from https://wiki.cancerimagingarchive.net

Calc-Test-Description (csv, 186 kB)	*Download*	CC BY 3.0

Click the Versions tab for more info about data releases.

Detailed Description

Detailed Description

Collection Statistics	
Modalities	MG
Number of Participants	1,566*
Number of Studies	6775
Number of Series	6775
Number of Images	10239
Image Size (GB)	163.6

* The image data for this collection is structured such that each participant has multiple patient IDs. For example, pat_id 00038 has 10 separate patient IDs which provide information about the scans within the IDs (e.g. Calc-Test_P_00038_LEFT_CC, Calc-Test_P_00038_RIGHT_CC_1) This makes it appear as though there are 6,671 participants according to the DICOM metadata, but there are only 1,566 actual participants in the cohort.

The CBIS-DDSM contributors have provided the following additional options for subset download.

Data Type	Download
Mass-Training Full Mammogram Images (DICOM)	Mass-Training_full_mammogram_images_1-doiJNLP-wv6aeYDn.tcia (Download requires the NBIA Data Retriever)
Mass-Training ROI and Cropped Images (DICOM)	Mass-Training_ROI-mask_and_crpped_images_1-doiJNLP-07gmVj4b.tcia (Download requires the NBIA Data Retriever)
Calc-Training Full Mammogram Images (DICOM)	Calc-Training_full_mammogram_images_1-doiJNLP-PrQ05L6k.tcia (Download requires the NBIA Data Retriever)
Calc-Training ROI and Cropped Images (DICOM)	Calc-Training_ROI-mask_and_crpped_images-doiJNLP-kTGQKqBk.tcia (Download requires the NBIA Data Retriever)
Mass-Training-Description (csv)	mass_case_description_train_set.csv
Calc-Training-Description (csv)	calc_case_description_train_set.csv
Mass-Test Full Mammogram Images (DICOM)	Mass-Test_full_mammogram_images-doiJNLP-6ccCrb8t.tcia (Download requires the NBIA Data Retriever)

Reproduced from https://wiki.cancerimagingarchive.net

Mass-Test ROI and Cropped Images (DICOM)	Mass-Test_ROI-mask_and_crpped_images-doiJNLP-SmEOyQFn.tcia (Download requires the NBIA Data Retriever)
Calc-Test Full Mammogram Images (DICOM)	Calc-Test_full_mammogram_images-doiJNLP-SiXj6kpS.tcia (Download requires the NBIA Data Retriever)
Calc-Test ROI and Cropped Images (DICOM)	Calc-Test_ROI-mask_and_crpped_images-doiJNLP-PsjCfTdf.tcia (Download requires the NBIA Data Retriever)
Mass-Test-Description (csv)	mass_case_description_test_set.csv
Calc-Test-Description (csv)	calc_case_description_test_set.csv

The CBIS-DDSM was created from DDSM by undertaking the following specific procedures:

1) Removal of questionable mass cases

Not all DDSM ROI annotations include suspicious lesions. Due to this issue, a trained mammographer reviewed the questionable cases. In this process, 254 images were identified in which a mass was not clearly seen. These images were removed from the final data set.

2) Image Decompression

DDSM images are distributed as lossless 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. To address this the PVRG-JPEG codec was modified to successfully compile on an OSX 10.10.5 (Yosemite) distribution using Apple GCC clang-602.0.53. The decompression code outputs data in 8-bit raw binary bitmaps. Python tools were developed to read this raw data and store it as 16-bit gray scale TIFF files. These files were later converted to DICOM.

This process is entirely lossless and preserved all information from the original DDSM files.

3) 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. To address this the tools were re-implemented 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. The DDSM automatically clips optical density values to be between 0.05 and 3.0 for noise reduction. This clipping occurs in the CBIS-DDSM as well, but the new tools provide a flag to remove the clipping and retain the original optical density values.

4) Image Cropping

Reproduced from https://wiki.cancerimagingarchive.net

Several CAD tasks require only analyzing abnormalities (the portion of the image in the ROI) without needing the full mammogram image. A set of convenience images are also provided, which are focused crops of abnormalities. Abnormalities were cropped by determining the bounding rectangle of the abnormality with respect to its ROI. The square crops were created by extending the shorter edge of the rectangle to be the same size as the long edge. The centroid of the abnormality is located in the center of these square crops.

5) Updating for precision segmentation

Mass margin and shape have long been proven to be significant 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, a lesion segmentation algorithm (described below) was applied that is initialized by the general, original DDSM contours but is able to supply much more accurate ROIs. This 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 Vese11. Level set models follow a non-parametric deformable model, thus can handle topological changes during evolution11. 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. This 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 the original DDSM annotation was used as the level set segmentation initialization.

6) Standardized Train/Test splits

The data were split into a training set and a testing set based on the BIRADS category. This allows for an appropriate stratification for researchers working on CADe as well as CADx. 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 both the CC and MLO views.

Citations & Data Usage Policy

Citations & Data Usage Policy

Users must abide by the TCIA Data Usage Policy and Restrictions. Attribution should include references to the following citations:

Data Citation

Sawyer-Lee, R., Gimenez, F., Hoogi, A., & Rubin, D. (2016). **Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [Data set]**. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY

Publication Citation

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). **A curated mammography data set for use in computer-aided detection and diagnosis research**. In Scientific Data (Vol. 4, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/sdata.2017.177

Reproduced from https://wiki.cancerimagingarchive.net

TCIA Citation

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). **The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository**. In Journal of Digital Imaging (Vol. 26, Issue 6, pp. 1045–1057). Springer Science and Business Media LLC. https://doi.org/10.1007/s10278-013-9622-7

Other Publications Using This Data

TCIA maintains a list of publications which leverage TCIA data. If you have a manuscript you'd like to add please contact TCIA's Helpdesk.

Versions

Version 1 (Current): Updated 2017/09/14

Data Type	Download all or Query/Filter
Images (DICOM, 163.6GB)	*Download* *Search* (Requires the NBIA Data Retriever .)