AD_____

GRANT NUMBER DAMD17-94-J-4015

TITLE: Digital Image Database With Gold Standard and Performance Metrics for Mammographic Image Analysis Research

PRINCIPAL INVESTIGATOR: Kevin W. Bowyer, Ph.D.

CONTRACTING ORGANIZATION: University of South Florida
Tampa, Florida 33620-7900

REPORT DATE: August 1998

TYPE OF REPORT: Annual

PREPARED FOR: Commanding General
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1998	Annual (1 Jul 97 - 31 Jul 98)

4. TITLE AND SUBTITLE
Digital Image Database with Gold Standard and Performance Metrics for Mammographic Image Analysis Research

5. FUNDING NUMBERS
DAMD17-94-J-4015

6. AUTHOR(S)
Kevin W. Bowyer, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of South Florida
Tampa, Florida 33620-7900

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Digital Database for Screening Mammography (DDSM) is an infrastructure resource for use by the mammogram image analysis research community. It consists of digitized images for screening mammogram cases, associated information and associated software tools. Portions of the DDSM resource are currently available on the world-wide-web and are regularly being accessed by researchers. This report discusses the current sate of the DDSM resource, problems encountered and plans for completing the resource.

14. SUBJECT TERMS
Breast Cancer MAMMOGRAPHIC IMAGE ANALYSIS, DIGITAL IMAGE DATABASE, PERFORMANCE EVALUATION, SOFTWARE TOOLS, COMPARATIVE EVALUATION OF TECHNIQUES, BREAST CANCER

15. NUMBER OF PAGES
15

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

KWB Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Kevin W Bauguss 3 Sept 98

PI - Signature Date

Contents

4 Introduction

The Digital Database for Screening Mammography (DDSM) is an infrastructure resource for the mammogram image analysis research community. The purpose of the DDSM resource is to make it possible for researchers to conduct a more rigorous experimental comparison of the performance of different image analysis techniques. Previously, most research on computer image analysis for mammogram screening has used a "small" (10s to perhaps 100) number of images. Also, researchers have generally not been able to evaluate their work using the same images as used by other researchers. The DDSM infrastructure resource is meant to address these problems.

This annual report presents an update on (1) the current state of the DDSM resource, (2) problems encountered in developing the DDSM resource, and (3) plans for completion of the DDSM resource.

5 Current State of the DDSM Resource

The DDSM resource currently contains both image-related data and associated software tools. The image-related data is organized by case, where a "case" is the standard four images of a screening exam, plus information on patient age, a radiologist-specified BIRADS breast density rating, a radiologist-specified outline of the suspicious region(s) in an image, and a radiologist-specified subtlety rating for detection of the lesion(s) in the image.

Major improvements have been made to the web site for the DDSM resource in the last year. More data is available. A search engine has been added to allow users to determine what cases fit a given description. Additional supporting information has been added. The address for the DDSM web site is:

`http://marathon.csee.usf.edu/Mammography/Database.html`

A printed copy of this main page of the web site appears as the first appendix to this report. This should give an idea of the information provided and its basic organization.

Currently close to 1,000 cases of mammogram data can be browsed as image "thumbnails." By default, cases are organized into groups called "volumes." (In general, a "volume" is intended to be a data set of the size that just fits on one standard-format, 8-mm tape. This is for ease of exchanging data through the regular mail.) The cases in one "volume" are all instances of either normal, cancer or benign cases. Also, the images for the cases in one volume all were digitized by the same film scanner. The second appendix to this report is a printed copy of overview page for volume "cancer_01." This explains in more detail the content of the database. The third appendix to this report is a printed copy of the "thumbnail" overview of the first case in the volume "cancer_01."

Data in DDSM currently comes from three clinical sites: Massachusetts General Hospital in Boston, Wake Forest University School of Medicine in North Carolina and Sacred Heart Hospital in Pensacola, Florida. Three different types of film scanners have been used to date: DBA Systems, Lumisys, and Howtek. The DBA scanner is no longer in use for this project; this will be explained in more detail in a later section.

The search engine available on the web page allows the user to collect together thumbnails of all cases which satisfy a search query. For example, a user could collect together and browse the thumbnails of all cancer cases which have clustered calcifications, or of all density rating 4 cases which have a spiculated lesion.

The software tools available on-line through the DDSM resource include a utility ("DDSMView") for viewing the images and image-related data, a routine for matching the results of a CAD detection program to the radiologist-specified ground-truth location of a lesion, and a pointer to a lossless JPEG image compression utility.

Each month, there are accesses to the DDSM resource from a number of distinct sites around the world, including a wide variety of academic, government and commercial sites in the United States. In addition, a number of people have ordered tapes of various parts of the database, rather than downloading over the web. The company R2, which recently received FDA approval for their product, ordered several volumes of data.

6 Problems Encountered

As mentioned in previous reports, the DDSM project has proved to be a great challenge, and we have encountered many problems that were not originally anticipated.

One group of problems that has been tracked in the annual reports revolves around the use of the DBA M2100 film digitizer. This is the digitizer installed and initially used at Massachusetts General Hospital. Two years ago, we still had some plausible hope that DBA would make an honest attempt to provide a fully functional system. Last year it seemed to us that, while they did not formally state this, they no longer had any intentions of making the system work as originally described. In last year's annual report, we quoted this portion of a letter sent to DBA:

> We repeat the request made of our letter dated January 23, 1997:
> "To make sure that we at MGH/USF are being fair-minded, and so we can calibrate our experience to the broader world of customers, can you give us the contact name and telephone number of persons who have one of the M2100 ImageClear Digitizer systems in regular use at 21 microns for upward of 32 films per day and are happy with it?"
>
> Lastly, we consider that the warranty period has not started yet, since MGH and USF and DBA Systems seem to agree that the product does not yet function as represented at the time of purchase. If your position is now that the system cannot feasibly be made to function as originally represented, then a simple statement to that effect may help to move our discussions forward.

We have still not received any response to this letter (some 18 months later).

At this point in the project, the DBA digitizer is no longer functioning in a useful way. It is sitting unplugged at MGH. It has been replaced by a Howtek digitizer which seems to be much more reliable, in terms of both software and hardware. In addition, the technician operating the Howtek at MGH feels that the user interface is much better. MGH is investigating the possibility of returning the digitizer to DBA to seek some sort of refund.

Description of some of the problems encountered in attempting to use the DBA digitizer were described in the previous annual report, and are not repeated here.

The previously mentioned problem of the unavailability rate of cancer case films from the archives has been compensated for by collecting cases from a broader time frame than originally anticipated. The previously mentioned problem of permanent markings having been made on some cancer case films has been dealt with by more careful additional cleaning and by dropping cases from collection as a last resort.

One continuing problem is technically a minor thing, but in practical terms has generated an enormous unanticipated manual inspection of digitized images. This is the problem of deleting all patient identifier information from the digitized images. Patient identifier information appears typically in two places in a digitized image. The more obvious is the patient identifier information which is in the film image, due to lettering placed on the film cassette at the time of image acquisition. This identifier information is taped over before the film is digitized, using a polyester metalized tape. However, the tape does not effectively block all light transmission given the high sensitivity of the digitizer being used. In rare instances, this could allow the reconstruction of at least some of the patient identifier information. The other source of patient identifier information results from a typed gum label placed on the film after the study is done. When the film is digitized, this lettering on the label can be visualized in the digitized image. As a result, we have introduced a manual inspection step in which the digitized images are checked for any visualizable patient identifier information and such regions of the image are outlined and set to zero.

7 Plans for completion of the DDSM resource

We are now entering the one-year, no-cost extension to the original grant period. We are still well behind the originally projected schedule. However, we are hopeful that we will be able to complete all, or at least the major portion, of the originally planned project. The flow of data from acquisition through all of the processing steps has been good since the installation of the Howtek digitizer at MGH. In the past month, we have placed approximately 200 new cases in the database for distribution. We need to maintain a rate of approximately 200 per month for the next 10 months in order to complete the originally planned project. This allows the last month or two to package the last batches of data acquired.

8 Conclusions

Close to 1,000 cases of data can now be browsed in thumbnail form on the DDSM web site. We are making regular additions to the available data. The frequency of accesses to the database is continuing to grow. The traffic in download of data from the database can now be tracked on the web page in the form of a histogram of volume of data per week. We have received a number of positive comments about the quality of the data and are hopeful the database will facilitate higher quality research in mammogram image analysis.

University of South Florida Digital Mammography Home Page

DDSM: Digital Database for Screening Mammography

The Digital Database for Screening Mammography (DDSM) is a database established for use by the mammographic image analysis research community. This project is supported by a grant from the Breast Cancer Research Program of the U.S. Army Medical Research and Materiel Command. The DDSM project is a collaborative effort involving Massachusetts General Hospital, the University of South Florida, and Sandia National Laboratories. The primary purpose of the database is to facilitate sound research in the development of computer algorithms to aid in screening. Secondary purposes of the database may include the development of algorithms to aid in the diagnosis and the development of teaching or training aids. The database should eventually contain approximately 3,000 studies. Each study includes two images of each breast, along with some associated patient information and image information. Images containing suspicious areas will have associated pixel-level "ground truth" information about the locations and types of suspicious regions. Also provided will be software both for accessing the mammogram and truth images and for calculating performance figures for automated image analysis algorithms.

The Digital Database for Screening Mammography is organized into "cases" and "volumes." A "case" is a collection of images and information corresponding to one mammography exam of one patient. A "volume" is simply a collection of cases collected together for purposes of ease of distribution. The DDSM database is under construction. All volumes are available on 8mm tape, and at any given point in time, a number of volumes are also available on-line. The README file explaining "everything" about the database is available, and many answers to questions about the database are listed below.

- **What information is included in a case?**
  A case consists of between 6 and 10 files. These are an "ics" file, an overview "16-bit PGM" file, four image files that are compressed with lossless JPEG encoding and zero to four overlay files. Normal cases will not have any overlay files. Click here for more detailed information on the files contained in a case.

- **What is the difference between normal, cancer and benign volumes?**

Each volume is a collection of cases of the corresponding type. Normal cases are formed from a previous normal screening exam (pulled from a file) for a patient with a normal exam at least four years later. A normal screening exam is one in which no further "work-up" was required. Cancer cases are formed from screening exams in which at least one pathology proven cancer was found. Benign cases are formed from screening exams in which something suspicious was found, but was determined to not be malignant (by pathology, ultrasound or some other means).

- **What volumes are available?**
  This database is still growing. The table below lists the volumes that are currently part of the database:

VOLUME	CASES	SIZE	SCANNER	BITS	RESOLUTION	THUMBNAILS	NOTES	AVAI
normal_01	111	5.8 GB	DBA	16	42 microns	thumbnails	notes	
normal_02	117	6.6 GB	DBA	16	42 microns	thumbnails	notes	
normal_03	38	4.1 GB	DBA	16	42 microns	thumbnails	notes	
normal_04	57	5.1 GB	DBA	16	42 microns	thumbnails	notes	ftp
normal_05	47	4.3 GB	DBA	16	42 microns	thumbnails	notes	
normal_06	60	5.5 GB	DBA	16	42 microns	thumbnails	notes	
cancer_01	69	3.9 GB	LUMISYS	12	50 microns	thumbnails	notes	
cancer_02	88	5.7 GB	LUMISYS	12	50 microns	thumbnails	notes	
cancer_03	66	6.0 GB	DBA	16	42 microns	thumbnails	notes	ftp
cancer_04	31	2.8 GB	DBA	16	42 microns	thumbnails	notes	
cancer_05	83	6.6 GB	LUMISYS	12	50 microns	thumbnails	notes	ftp
cancer_06	56	6.3 GB	HOWTEK	12	43.5 microns	thumbnails	notes	
benign_01	70	5.6 GB	LUMISYS	12	50 microns	thumbnails	notes	

- **Do you have a "troubleshooting" section on you web pages?** NEW
  Yes. We have compiled a list of frequently asked questions and have provided answers to them. Click here to go to that information.

- **How do I acquire a volume?**
  Several volumes will be available by anonymous ftp at any given time (**figment.csee.usf.edu in pub/DDSM/cases**). You can download individual cases or entire volumes. Occasionally, we will change which volumes are available on line giving preference to the more recently released volumes. All volumes that are part of the database (whether they are on, or off line) can be ordered. Each is available on 8mm EXABYTE 160mXL data cartridges created using the UNIX tar command (and a model 8505XL 8mm drive). To order tapes, please specify the volume(s), and send a check of $30.00 for the first tape plus $20 for each additional tape. For international orders, add an additional $20. This is for customs and mailing. If we can find a cheaper way to do it, this may change in the future. Click here for an order form.

  Make check payable to: **University of South Florida** (Please be careful that the check is not made out to "University of Florida", "Florida Southern University", "University of Southern Florida" or other variations; this can cause problems at the bank.)
  Unfortunately, we are not set up to accept purchase orders or credit cards.
  Checks must be made in U.S. dollars, drawn on a U.S. bank. Mail to:

  Rachel Gadsden
  University of South Florida
  Department of Computer Science
  4202 E. Fowler Ave.
  ENB 118
  Tampa, FL 33620-5399

- **What software is available for working with this data?**
  We have software available for uncompressing image files, viewing cases, converting images to 16-BIT PGM format and utilities for comparing automated analysis results to ground truth. Source code for the lossless JPEG compression program is available from Stanford University. Documentation on the use of the viewing software, DDSMView, is also available.

- **Can I preview the cases in a volume?**
  Yes, we have made web pages that show "thumbnail" versions of the images. See the table for links to each volume of thumbnails. Each case has a separate web page. On each page, "thumbnail" images are displayed with all of the ground truth markings overlayed on them. The text information from the ics file and all of the overlay files is also provided. Please note that the colors for the overlayed ground truth markings are selected independently for each image. The color of each boundary can be used to index the associated textual information for that marking in the overlay table.

- **Can I search the cases in in the database?** NEW
  Yes, we have recently added a search capability to our database. Click here to search the database.

- **What is the "notes" link in the table of cases?**
  The table of cases has a link to a page for each volume. Each page contains additional information about cases and information on any changes made to the cases after they were released. Although each case is checked thoroughly (and re-checked) before being released, errors may rarely exist in released volumes. When any errors are found, they will be corrected and listed on the notes page for that volume.

- **How do I map grey levels to optical density?**
  In some situations, it may be useful to be able to map the grey levels in a mammogram image to optical density values. For example, you may want to run your image analysis software on data sets that were acquired on two different scanners. Since the grey levels in images acquired on different scanners will probably not correspond to the same optical density, you may want to "normalize" the images in some manner prior to processing them.
  Here's how to map grey levels to optical density for images digitized at:
  - DBA scanner
  - HOWTEK scanner
  - LUMISYS scanner

- **How can I keep myself informed on updates/additions to this database?**
  To place yourself on an electronic mailing list to receive updates about this project (including the eventual creation of mailing list discussion group), Click Here *Email: ddsm@bigpine.csee.usf.edu*

- **Do you have anonymous ftp access statistics available?** NEW
  Yes. We have a page displaying a graph showing the amount of data downloaded from DDSM (pub/DDSM/cases) by anonymous ftp each week. Click here to view the graph.

- **Are there other Mammography resources on this web site.**
  Yes. They have been moved to our "Other Resources" page.

Note: The Digital Database for Screening Mammography (DDSM) is supported through a grant from the DOD Breast Cancer Research Program, US Army Research and Material Command DAMD17-94-J-4015. The server for the DDSM is a dual processor Sun Sparc 20 with 520 Megs of RAM donated by Sun Microsystems through their Academic Equipment Grant (AEG) program, Grant #: EDUD-US-950408.

Please mail comments, suggestions and specific mammography questions to:
ddsm@bigpine.csee.usf.edu

Digital Database for Screening Mammography

Overview of Volume: cancer_01

This table contains links to overview pages for the cases in this volume. Each overview page contains the ics file, "thumbnails" of the images associated with the case, and the overlay markings of the abnormalities. The "thumbnail" images are meant for visual browsing of the cases only, and are not suitable for any form of experimental work. The thumbnail images were blurred, subsampled, sharpened and converted to eight bits per pixel. Thus, they contain artifacts introduced by this processing (i.e., aliasing, ringing, re-quantization, clipping etc.). Any experimental work should start with the full raw images for the case.

Each case in this volume of cancer cases has at least one path-proven cancer. Some cases contain more than one cancer in one breast, a cancer in each breast, or a cancer along with other abnormal/suspicious regions. The outlines of all regions have been transcribed from markings made by an experienced mammographer. In almost all cases, an abnormality is visible in both views of the breast. However, there are a cases where an abnormality seen in one view is not visible in the other view. This is most often due to the fact that the CC and MLO views do not image exactly the same tissue, but in some cases the abnormality may simply not present any visible signs in one view.

The "ics" header file contains the following information. The "DATE_OF_STUDY" field gives the date that the mammogram was performed. The "PATIENT_AGE" field gives the age of the patient at the time that the mammogram was performed. The "DENSITY" field gives an ACR BI-RADS breast density rating for the case, as assigned by an experienced mammographer. The "DIGITIZER" field indicates the equipment used to digitize the mammogram films. The "LEFT_CC," "LEFT_MLO," RIGHT_CC," and "RIGHT_MLO" fields specify the size of the corresponding images and whether or not a file exists for the overlay of abnormality outlines. The "FILM," "FILM_TYPE," "SEQUENCE" and "DATE_DIGITIZED" fields should be disregarded. They are being maintained for compatibility with earlier versions of the database, but do not contain useful information.

The individual overlay files contain keywords to describe each abnormality. They also contain an ACR BI-RADS assessment code for each abnormality. And they contain a mammographer-assigned "subtlety rating" on a scale of 1 to 5, where 1 is "subtle" and 5 is "obvious." The mammographer was not constrained to provide the same keyword description for a given abnormality in each view. Unusual objects (e.g., pacemaker, breast implant) may be visible in some images. These are not identified by an overlay file.

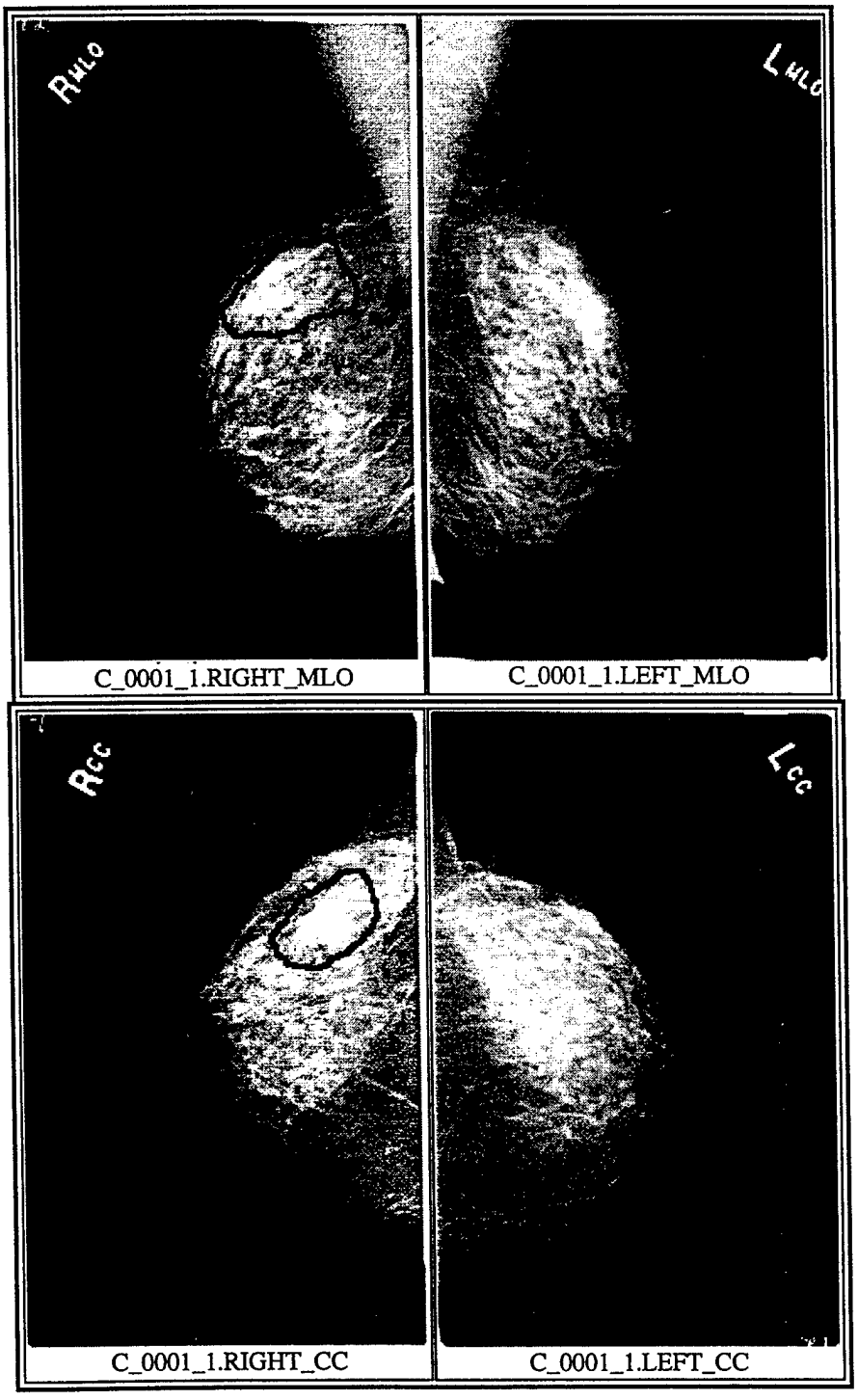
We strongly encourage that whenever research results are presented using cases from this database, a histogram of the subtlety rating of the images used in the research also be presented.

You can order the complete volume on tape. Click here for an order form. Please note that you should be using the lastest version of DDSMView to view the complete cases. You can get it here or from our ftp site (figment.csee.usf.edu) in the directory "pub/DDSM/software/bin" using anonymous ftp.

case0001	case0002	case0003	case0004	case0006	case0007	case0009	case0010
case0011	case0012	case0014	case0015	case0016	case0017	case0019	case0020
case3001	case3003	case3005	case3007	case3008	case3009	case3010	case3012
case3013	case3016	case3017	case3018	case3019	case3020	case3021	case3022
case3023	case3025	case3026	case3027	case3030	case3032	case3033	case3037
case3038	case3041	case3042	case3044	case3045	case3046	case3047	case3049
case3051	case3055	case3057	case3058	case3059	case3062	case3064	case3065
case3066	case3068	case3071	case3072	case3073	case3076	case3079	case3080
case3081	case3082	case3083	case3084	case3086			

Digital Database for Screening Mammography

Volume: cancer_01 Case: C-0001-1



```
ics_version 1.0
filename C-0001-1
DATE_OF_STUDY 15 10 1992
PATIENT_AGE 65
FILM
FILM_TYPE REGULAR
DENSITY 2
DATE_DIGITIZED 18 8 1997
DIGITIZER LUMISYS LASER
SEQUENCE
LEFT_CC LINES 4608 PIXELS_PER_LINE 2928 BITS_PER_PIXEL 12 RESOLUTION 50 NON_OVERLAY
LEFT_MLO LINES 4592 PIXELS_PER_LINE 2896 BITS_PER_PIXEL 12 RESOLUTION 50 NON_OVERLA
RIGHT_CC LINES 4584 PIXELS_PER_LINE 2920 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
RIGHT_MLO LINES 4616 PIXELS_PER_LINE 2888 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
```

```
FILE: C_0001_1.RIGHT_MLO.OVERLAY

TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE MASS SHAPE IRREGULAR MARGINS ILL_DEFINED
ASSESSMENT 4
SUBTLETY 3
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 1
BOUNDARY
```

```
FILE: C_0001_1.RIGHT_CC.OVERLAY

TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE MASS SHAPE IRREGULAR MARGINS SPICULATED
ASSESSMENT 5
SUBTLETY 5
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 1
BOUNDARY
```