Medical Informatics Europe '97
C. Pappas et al. (Eds.)
IOS Press, 1997

Medical image databases for CAD applications in digital mammography: Design issues

Maria Kallergi, Robert A. Clark, and Laurence P. Clarke
Department of Radiology, University of South Florida, Tampa, FL 33612, U.S.A.

Abstract. The evaluation of algorithms developed for computer assisted diagnosis in digital mammography requires image databases that allow relative comparisons and assessment of algorithms' clinical value. A review of the literature indicates that there is no consensus on the guidelines of how databases should be established. Image selection is usually done based on subjective criteria or availability. The generation of common database(s) available to the research community makes relative evaluations of algorithms with similar properties easier. However, questions regarding the "right database size," the "right image resolution," and the "right contents" remain. In this paper, database issues are reviewed and discussed and possible remedies to the various problems are proposed.

1. Common Databases

Computer assisted diagnosis (CAD) in radiology has sparked considerable research activity in the last decade. Mammography has claimed the majority of the applications, which focus on the automatic detection and/or classification of breast abnormalities such as calcification clusters and masses. The relative as well as clinical evaluation of CAD algorithms requires specific image databases the design of which faces significant problems and criticism. The criteria commonly used for the selection of test cases are defined by logistical issues related to the particular application, statistical significance, and/or one or more human experts [1-3].

The generation of *common databases* was proposed to address the increasing need for absolute or relative comparisons of CAD performances. First, in 1992, the Department of Radiology at the University of South Florida (USF) and the H. Lee Moffitt Cancer Center & Research Institute compiled a database of 100 mammograms, normal and cancer cases [4]. This effort initiated a panel discussion at the SPIE & IS&T 1993 Meeting regarding design issues of such a common image set with panelists from research groups around the world [5]. Since then, three more databases were made available to the research community: the MIAS database of 161 pairs of images with various abnormalities [6], the Nijmegen database of 40 images with microcalcifications [7], and the LLNL/UCSF database of 198 images with microcalcifications [8]. A large database is currently under development at USF's Computer Science and Engineering Department [9]. The selection of images for the existing common databases is usually based on criteria such as availability of films and pathology reports, and experts' evaluations of difficulty and degree of representation of clinical reality. These databases allow the relative evaluation of CAD methodologies but leave the general database design issues unresolved.

2. Review of Properties of Reported Databases

The issues of medical image databases have been addressed at several workshops [5,10,29] and by several researchers usually in relationship to observer studies [1,2]. In particular, Kroon *et al* [1] have provided a model for compiling databases of chest radiographs, Nishikawa *et al* [3] offered suggestions for mammograms with masses, and Kallergi *et al* [11] and Chan *et al* [12,13] discussed similar issues for mammograms with calcifications.

Literature suggests that five basic criteria should be considered: (a) truth of diagnosis, (b) quality of original films, (c) degree of difficulty of cases, (d) total number of cases and representation of clinical characteristics, and (e) digitizer and digitization conditions.

Pathological proof is the commonly used criterion for benign/malignant differentiation [14]. At least 2-year clinical follow-up without change is the criterion used for normals (negatives and nonbiopsied benign cases) [14]. Using the radiological and pathological reports, an expert usually indicates the size and location of an abnormality on a film copy or on the digital image. Truth files are generated and used to calculate the indices of performance. The limitations of this ground truth source are that it is subjective and provides limited quantitative information on the distribution and morphology of the abnormality. Electronic truth files are proposed and sometimes generated [4]. These files are still subjective and time consuming to create but offer some advantages, e.g., better and automatic assessment of the "correct decision" of an algorithm.

The rule of thumb regarding the quality of the films to be digitized is that mammograms that are acceptable in daily practice should be acceptable for the database [1]. However, this rule may sometimes allow inclusion of mammograms with common artifacts such as those due to dust or film processing, which can be easily recognized and disregarded by the human observer but pose significant problems in the application of CAD methods. Therefore, a stricter rule that excludes films with any artifacts is often applied at the risk of overestimating the performance of the method and having a questionable assessment of its clinical value.

The difficulty of the test images is the easiest target for criticism and an issue that has currently no objective solution [1-3,10]. There are several proposed parameters that are directly or indirectly related to the subtlety of the cases. For example, many researchers measure difficulty by the degree of visibility of an abnormality. The visibility is usually assessed by one or more experts using a 3-, 5-, or 10-point scale, from easily discernible to very hard to distinguish [15]. Lesion contrast, lesion size, and breast parenchyma measurements are also proposed as indirect measures of difficulty [3,11,16] because the majority of missed cancers are ≤10 mm in largest diameter (minimal cancers), of low contrast, and/or in dense breasts [17].

The number of images and the proportions of different abnormalities are issues that can have a relatively easy answer but with significant complications. For example, there are well known mathematical criteria dealing with error measurements [18] and statistical significance [2] that can be used to estimate the sample size needed for adequate training and/or testing of an algorithm. In addition, clinical proportions of normal and abnormal cases, parenchymal densities, and types of abnormalities are generally well known and can be followed in the establishment of a database. Such sample size estimates, however, are usually large and impractical. So, until now, the size seems to be determined more by the availability of the mammograms and the constraints on computation and less by the requirements for statistical accuracy or generalization. Small sample biases may be reduced by using different sets of images for training and testing the algorithms or, better, using resampling plans [19]. Observer studies (ROC) of CAD have been reported with up to 270 mammograms [11,20]. Non-observer studies with full mammograms are reported with sets of 25 to about 100 cases [3,21,22] while most studies are done on regions of interest (ROIs); as many as 672 ROIs have been used [15-17,23,24]. None of the used databases until now represents clinical distributions in terms of either benign:malignant, normal:abnormal, dense:fatty breast ratios. The reasons include unavailability, impractical proportions (only about 5 cancers per 1000 cases), specific requirements of ROC studies [1,2] or simply not needed for the study.

Until now, digital mammograms are generated from digitizing films. Hence image resolution and quality depends on the scanner and often the aim of the research. For microcalcifications, results have been reported for resolutions of 30-105 μm with 8-16 bits per pixel [11-13, 20,21,23-25]. For mass detection, results are reported for resolutions of 100-400 μm and 8-16 bits per pixel [3,16,22,26,27]. To our knowledge, no results are reported on image-based mass classification.

3. Proposed Criteria for Mammographic Database Generation

There are general guidelines on how to develop databases of mammograms and what should be avoided. Some requirements are relatively easy to satisfy, e.g., truth of diagnosis, film and digital image quality, matching normal and abnormal cases. Properties such as degree of difficulty, number of cases, relative proportions of different cases, and digitization conditions are more complex because they are observer, methodology, and task dependent and often impossible to define a priori. Possible solutions to the latter issues are discussed below.

Acutance, or sharpness, may be a better descriptor to the easy/difficult issue of cases that is task dependent, e.g., detection vs diagnosis. Measures of acutance proposed by Elkadiki and Rangayyan [28] may be modifiable to provide a quantitative measure of the degree of difficulty that relates the image properties to the performance of the human eye. Alternatively, the definition of acutance may include several variables such as perceptibility or visibility, size, contrast, and background parenchymal density, the estimation of which should involve more than one expert to account for variability. The contrast and size measurements of the lesions are useful parameters in the relative assessment of databases and can be related to difficulty if some ranges are established. False negative cases due to perception errors, cases where there is a discordance among radiologists, or minimal cancers are three groups that can be considered as sets of difficult cases and can have dual purpose: (a) be part of the test databases and (b) used to determine value ranges for the various parameters that describe the acutance of the data or provide insight in the definition of a measure for objective assessment of difficulty. An alternative proposed by our group to avoid the issue of "difficulty," is the use of consecutive cases that satisfy the film quality and biopsy criteria. This solution is attractive if many cases are to be studied but when a small number is selected, representation biases may be encountered.

The size of the database should be determined by the desired statistical and generalization power. For digital mammography, it is recommended that the generalization error is determined by the clinical requirements. If this is still impractical, then resampling techniques are recommended [19]. Wherever applicable, clinical distributions for breast density, lesion type, and normal:abnormal cases are preferred. Studies involving ROC or free response ROC tests, however, could use proportions determined by the model's requirements [2].

Digitization is an algorithm and task dependent issue, e.g., detection vs classification, CAD as a "pre-reader" vs CAD as an educational tool. Studies have shown that a pixel size $\leq$100 μm and depth $\geq$10 bits is adequate for calcification detection [11,12]. For masses, a pixel size of $\leq$200 μm and depth $\geq$8 bits per pixel may produce the desired results [22,27]. Classification of both abnormalities seems to demand pixel size $\leq$60 μm and depth $\geq$10 bits. It is recommended that films are digitized at the highest possible resolution, preferably one that matches the film's resolution. Then data can be reduced mathematically as needed. Finally, the arrival of direct digital mammography systems will make us revisit several issues including artifacts, resolution, and acutance, further increasing the need for objective and quantitative case selection criteria.

References

[1] Kroon HM, Steyerberg EW, Kool LJS, Hilkens CMU, and Seeley GW. Considerations in compiling a database of clinical test images. Invest. Radiol. 1992; 27:255-263.

[2] Metz CE. Some practical issues of experimental design and data in radiological ROC studies. Invest. Radiol. 1989; 24:234-245.

[3] Nishikawa RM, Giger ML, Doi K, Metz CE, Yin FF, Vyborny CJ, and Schmidt RA. Effect of case selection on the performance of computer-aided detection schemes. Med. Phys. 1994; 21(2):265-269.

[4] USF and H. Lee Moffitt Cancer Center mammography database - Questions can be addressed to Dr. Maria Kallergi (e-mail: kallergi@rad.usf.edu).

[5] Panel Discussion: Design of a common database for research in mammogram image analysis. SPIE 1993; 1905:534-553.

[6] Suckling J, Parke J, Dance DR, et al. The Mammographic Image Analysis Society Digital Mammogram Database. In *Digital Mammography,* Proc. 2nd Intern. Workshop on Digital Mammography, York, England,

10-12 July 1994. Gale AG, et al, Editors. Elsevier Science, B.V., pp. 375-378, 1994.

[7] Nijmegen mammography database - Questions can be addressed to Dr. Nico Karssemeijer (e-mail: nico@mbfys.kun.nl).

[8] Lawrence Livermore National Laboratories (LLNL) and University of California at San Francisco (UCSF) mammography database - Questions can be addressed to e-mail: mammo-db-hel@llnl.gov.

[9] USF Computer Science & Engineering Department mammography database -Questions can be addressed to e-mail: ddsm@bigpine.csee.usf.edu

[10] Zink S and Jaffe CC. Medical Imaging Databases: An NIH Workshop. Invest. Radiol. 1993; 28(4):366-372.

[11] Kallergi M, Clarke LP, Qian W, et al. Interpretation of calcifications in screen/film, digitized, and wavelet-enhanced, monitor displayed mammograms: An ROC study. Academic Radiology 1996; 3:285-293.

[12] Chan HP, Niklason LT, Ikeda DM, Lam KL, and Adler DD. Digitization requirements in mammography: Effects on computer-aided detection of microcalcifications. Med. Phys. 1994; 21(7):1203-1211.

[13] Chan HP, Sahiner B, Petrick N, Lam KL, and Helvie MA. Effects of pixel size on classification of microcalcifications on digitized mammograms. SPIE 1996; 2710:30-41.

[14] Feig SA. Decreased breast cancer mortality through mammographic screening: results of clinical trials. Radiology 1988; 167:659-665.

[15] Sahiner B, Chan HP, Wei D, et al. Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue. Med. Phys. 1996; 23(10):1671-1684.

[16] Wei D, Chan HP, Helvie MA, et al. Classification of mass and normal breast tissue on digital mammograms: Multi-resolution texture analysis. Med. Phys. 1995; 22:1501-1513.

[17] Reintgen D, Berman C, Cox C, Baekey P, Nicosia S, Greenberg H, Bush C, Lyman GH, and Clark RA. The anatomy of missed breast cancers. Surgical Oncology 1993; 2:65-75.

[18] Murata N, Yoshizawa S, and Amari S. Network Information Criterion - Determining the number of hidden units for an artificial neural network model. IEEE Trans. Neural Networks 1994; 5(6):865-872.

[19] Wagner RF, Chan HP, Mossoba J, et al. Finite sample effects and resampling plans: Applications to linear classifiers in computer-aided diagnosis. Proc. Medical Imaging 1997, Feb. 22-28, Newport Beach, CA.

[20] Nab HW, Karssemeijer N, Van erning L, Hendriks J. Comparision of digital and conventional and digital mammography, a ROC study of 270 mammograms. Med. Inform. 1992; 17(2):125-131.

[21] Qian W, Kallergi M, Clarke LP, Li HD, Venugopal P, Song D, and Clark RA. Tree structured wavelet transform segmentation of microcalcifications in digital mammography. Med. Phys. 1995; 22(8):1247-1254.

[22] Li H-D, Kallergi M, Clarke LP, Jain VK, and Clark RA. Markov Random Field for Tumor Detection in Digital Mammography. IEEE Trans Med Imag 1995; 14(3):565-576.

[23] Lefebvre F, Benali H, Gilles R, Kahn E, and Di Paola R. A fractal approach to the segmentation of microcalcifications in digital mammograms. Med. Phys. 1995; 22(4):381-390.

[24] Zhang W, Doi K, Giger ML, et al. An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. Med. Phys. 1996; 23(4):595-601.

[25] Huo Z, Giger ML, Vyborny CJ, Bick U, Lu P, Wolverton DE, and Schmidt RA. Analysis of spiculation in the comuterized classification of mammographic masses. Med. Phys. 1995; 22(10):1569-1579.

[26] Takehiro E, Doi K, Nishikawa R, et al. Image feature analysis and CAD in mammography: Reduction of false-positive clustered microcalcifications using local edge-gradient analysis. Med. Phys. 1995; 22(2):161.

[27] Yin FF, Giger ML, Doi K, et al. Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral subtraction technique. Med. Phys. 1994;21(3):445.

[28] Elkadiki SG and Rangayyan RM. Objective characterisation of image accutance. SPIE 1994; 2166:210.

[29] Windfield D, Silbiger M, Brown GS, et al. Technology transfer in digital mammography: Report of the Joint NCI-NASA Workshop of May 19-20, 1993. Invest. Radiol. 1994; 29(4):507-515.