

# Front-end data reduction in computer-aided diagnosis of mammograms: A pilot study

Hamed Sari-Sarraf<sup>\*a</sup>, Shaun S. Gleason<sup>a</sup>, Robert M. Nishikawa<sup>b</sup>

<sup>a</sup>Oak Ridge National Laboratory<sup>\*\*</sup>, P.O. Box 2008, Oak Ridge, TN 37831-6011

<sup>b</sup>The University of Chicago, Department of Radiology, Chicago, IL 60637

## ABSTRACT

This paper presents the results of a pilot study whose primary objective was to further substantiate the efficacy of front-end data reduction in computer-aided diagnosis (CAD) of mammograms. This concept is realized by a preprocessing module that can be utilized at the front-end of most mammographic CAD systems. Based on fractal encoding, this module takes a mammographic image as its input and generates, as its output, a collection of subregions called focus-of-attention regions (FARs). These FARs contain all structures in the input image that appear to be different from the normal background tissue. Subsequently, the CAD systems need only to process the presented FARs, rather than the entire input image. This accomplishes two objectives simultaneously: (1) an increase in throughput via a reduction in the input data, and (2) a reduction in false detections by limiting the scope of the detection algorithms to FARs only. The pilot study consisted of using the preprocessing module to analyze 80 mammographic images. The results were an average data reduction of 83% over all 80 images and an average false detection reduction of 86%. Furthermore, out of a total of 507 marked microcalcifications, 467 fell within FARs, representing a coverage rate of 92%.

**Keywords:** Computer-aided diagnosis, fractal encoding, microcalcification detection, data reduction, data throughput, breast cancer, computer vision

## 1. INTRODUCTION

The serious consequences of breast cancer among the female population, the importance of early detection in minimizing such consequences, and the role of mammography in effective screening and early diagnosis are all well known and well documented.<sup>1,2</sup> Also well documented, is the positive impact of utilizing CAD systems as aids to the radiologists.<sup>3</sup> The research in computer-assisted screening and diagnosis of mammographic abnormalities has been extensive over the past two decades,<sup>4</sup> leading to the development of commercially-available systems. There are, however, a number of challenging issues that are still under investigation. Two such issues, namely, the accuracy and the throughput of CAD systems, are the subject of the work that is presented herein.

The concept of front-end data reduction in mammographic CAD systems for the purpose of increasing the accuracy and the throughput was first conceived by the authors in 1995, and subsequently reported on in 1996.<sup>5</sup> This concept, which is based on fractal encoding, was first substantiated on a database of digitized mammograms provided by the Mammographic Image Analysis Society.<sup>6</sup> Since then, other researchers (Li et al.<sup>7</sup>) have also used fractal encoding as a means of modeling the normal background tissue in digitized mammograms. The work accomplished by Li et al., though aimed at a different goal of image enhancement rather than data reduction, is further evidence of the efficacy of fractal encoding in mammographic image analysis.

---

\*Correspondence: Email: sarisarrafh@ornl.gov; WWW: <http://www-ismv.ic.ornl.gov/~sarraf>; Telephone: 423 574 5542; Fax: 423 574 6663.

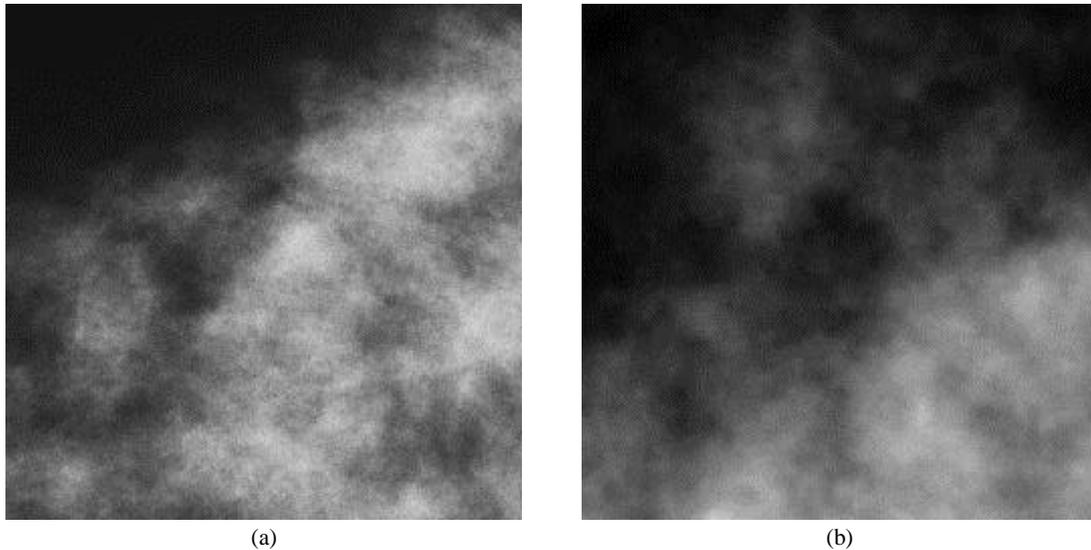
\*\*Managed by Lockheed Martin Energy Research Corporation for the U.S. Department of Energy under contract DE-AC05-96OR22464.

In this paper, the results of a second pilot study are reported. The primary objective of conducting this study was to further substantiate the efficacy of fractal-based, front-end data reduction for increased accuracy and throughput in mammographic CAD systems. The front-end, preprocessing module utilizes the fractal encoding scheme to reduce each mammographic image to a collection of FARs. These FARs contain all structures in the input image that appear to be different from the normal background tissue. Subsequently, the CAD systems need only to process the presented FARs, rather than the entire input image. This accomplishes two objectives simultaneously: (1) an increase in throughput via a reduction in the input data, and (2) a reduction in false detections by limiting the scope of the detection algorithms to FARs only. This recent study was conducted using a more extensive database of 80 mammograms that contain biopsy-proven clusters of microcalcifications.

We begin by including a brief summary of the theoretical basis of our approach. Then, in Section 3, details of the utilized methods, as well as our findings are presented.

## 2. FRACTAL ENCODING AND FAR GENERATION

The primary motivation for pursuing a fractal-based approach is that fractal encoding, and iterated function systems in general, are ideal for characterizing the cloud-like texture that represents the normal background tissue in mammograms, Fig. 1. This scheme can therefore be used to flag all structures that appear to be different from the normal background tissue. Fractal image encoding is the first of two steps executed in performing fractal image compression. Here, we describe fractal encoding in terms of its relationship to FAR generation. The interested reader is referred to prior publications<sup>5,8</sup> for a more detailed treatment of the subject matter.

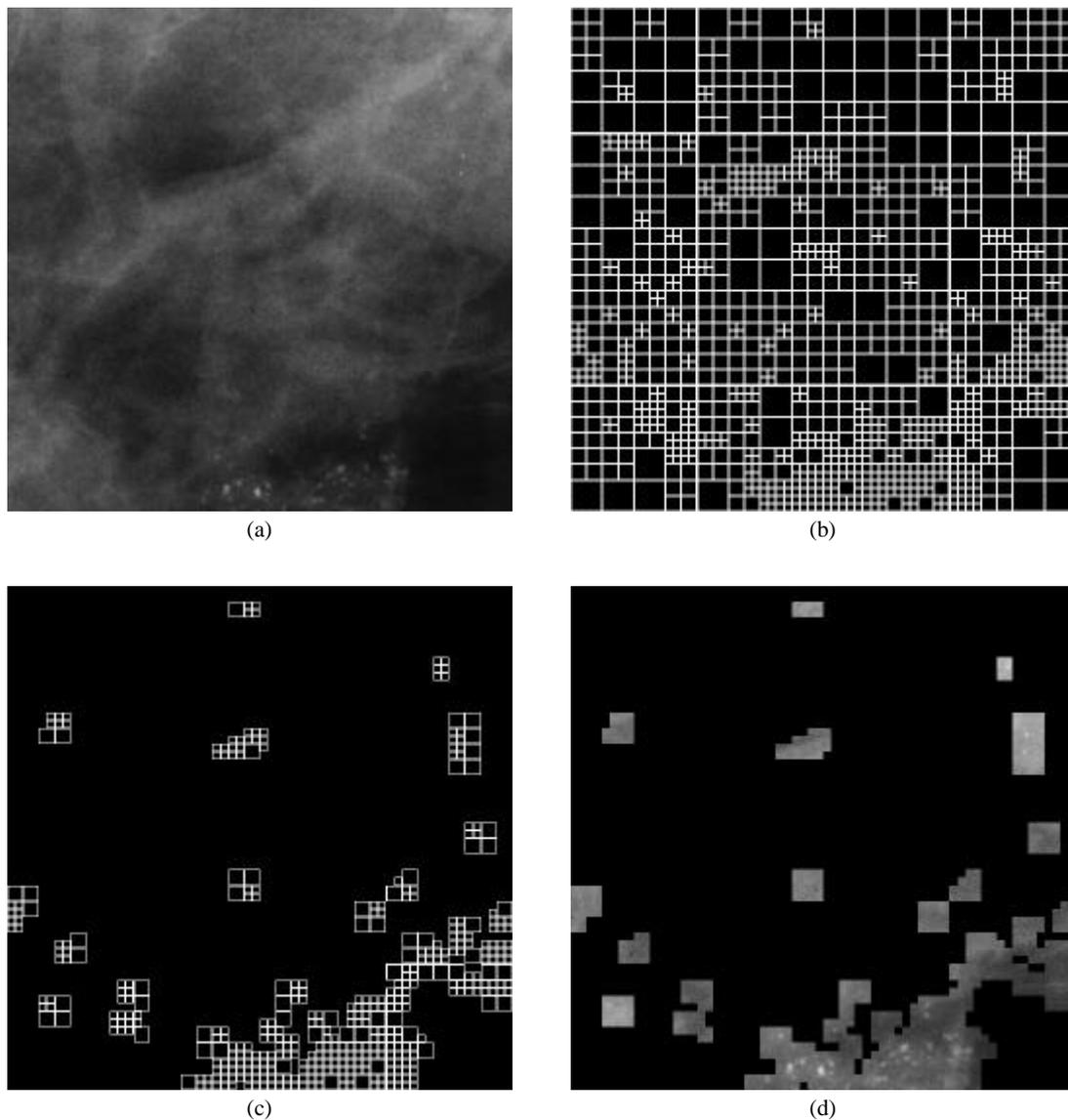


**Figure 1. These images demonstrate the textural similarity between a synthesized fractal and a digitized mammogram.**

Fractal encoding exploits the property of “partitioned” self-similarity of images. This means that instead of being formed of copies of its whole self (as is the case in exact self-similarity), the image, in an approximate sense, is composed of transformed parts of itself. In computing the coefficients of this transformation or map, it is assumed that because of the notion of partitioned self-similarity, each subregion of the image can be described (in the sense of minimizing a dissimilarity metric) in terms of another. The former subregion belongs to the range pool,  $R$ , while the latter belongs to the domain pool,  $D$ , of the map. If a given subregion in  $D$  cannot be mapped to any region in  $R$  (i.e., their measure of dissimilarity is above a specified threshold,  $T$ ), then  $R$  is further partitioned into smaller subregions. This process continues recursively until either a similar subregion from  $D$  is found or a specified maximum level of partitioning,  $L_{max}$ , is reached.

It is important to note that an exhaustive search to find matches between the subregions in the range and domain pools is computationally prohibitive. Generally, the subregions are classified into a manageable number of classes, and searches for similar regions occur only among those candidates that belong to the same class. The classification, as well as the range pool partitioning procedures, can take on a variety of forms. Thus far in this work, we have employed Fisher's classification technique and the commonly utilized quadtree partitioning scheme.<sup>9</sup>

In our first pilot study, we showed that during the fractal encoding process, for subregions in  $R$  that contain mammographic abnormalities,  $L_{max}$  will be reached. The reasoning is that the visual appearance of abnormalities such as microcalcifications is different from that of the coexisting normal structures. Therefore, subregions in  $D$  that can be mapped to those areas in the image with microcalcifications are expected to be nonexistent; see Figs. 2(a) and 2(b). Subregions for which  $L_{max}$  is reached along with their 8-neighbors make up the FARs, Fig. 2(c), (d).



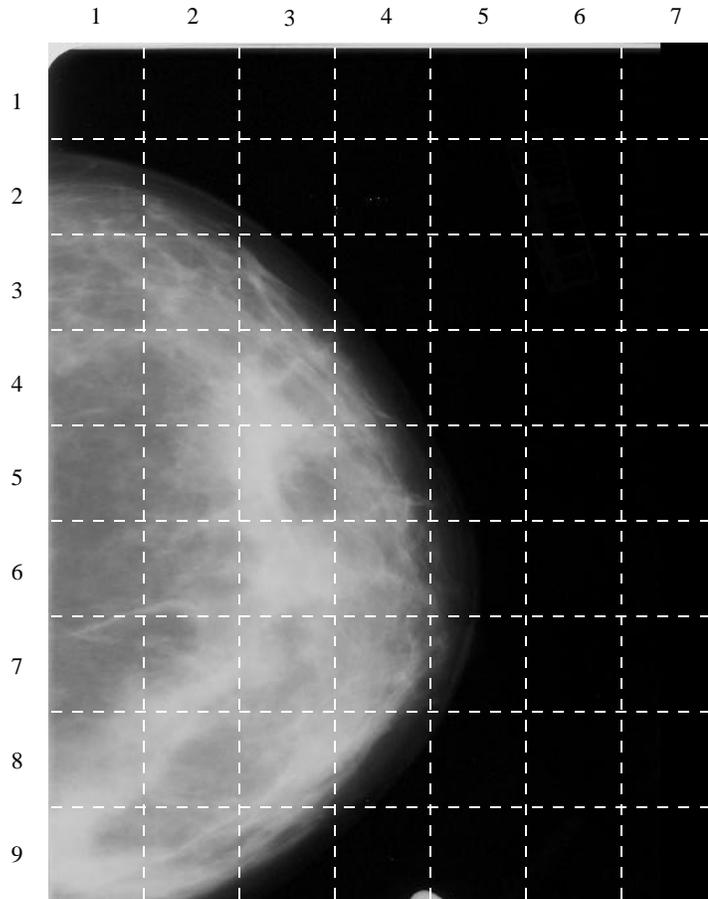
**Figure 2.** (a) A digital mammogram with clustered microcalcifications in the lower portion of the image. (b) Quadtree partitioning as a result of the fractal encoding of the image in (a) for  $L_{max} = 6$  (i.e., smallest subimages are  $8 \times 8$ ) and  $T = 3.4$ . (c) Those subregions and their 8-neighbors in (b) that never satisfied the similarity condition. (d) Generated FARs; note their concentration in the lower portion of the image.

### 3. PILOT STUDY

As mentioned above, by exposing an existing CAD system to FARs [e.g., regions in Fig. 2(d)] rather than the entire input image (e.g., Fig. 2(a)], two objectives are met: (1) an increase in throughput via a reduction in the input data, and (2) a reduction in false detections as a result of limiting the scope of the system's detection algorithms to FARs only. In what follows, we further substantiate this claim through an extensive pilot study involving 80 mammograms.

#### 3.1. Method

Eighty, 12-bit, 50 micron/pixel mammographic images were used to carry out this pilot study. These included 35 abnormal and 45 normal cases. The images were collected from 4 different hospitals. Cases containing biopsy-proven clustered microcalcifications were chosen randomly by radiologists at each center. The cluster locations were marked by an experienced radiologist at each site, and the location of individual microcalcifications within each cluster were marked by a medical physicist with over 15 years of experience in mammography. To mitigate problems caused by global nonuniformities, each image was divided into 512x512, nonoverlapping subimages (Fig. 3). Each subimage was then processed independently.



**Figure 3.** Prior to fractal encoding, each mammogram is divided into a number of 512x512, nonoverlapping subimages (63 in this case). Each subimage is then processed independently.

To study the impact of pixel depth on our approach, the FAR generation software was designed to handle both 8-bit and 12-bit pixels. A number of mammograms, though not all, were processed with both 8-bit and 12-bit pixels. The conversion from 12-bit to 8-bit pixels was performed using a straightforward linear mapping function. The results of this pilot study were quanti-

fied based on the following 3 criteria.

1. Reduction in data. That is, the percentage of input pixels not contained within FARs.
2. Reduction in false detections. Once a particular microcalcification detection algorithm is selected, it is applied to the entire input mammogram (or each subimage thereof). This performance measure is then quantified as the ratio of the number of detections that fall outside of FARs to the total number of detections for that mammogram (or each subimage thereof).
3. Coverage rate. That is, the percentage of marked microcalcifications that fall within FARs.

The detection algorithm devised by researchers at The University of Chicago<sup>10</sup> was used to quantify the second performance measure. This algorithm operates on 100-micron image data, obtained by pixel averaging the 50-micron images.

### 3.2. Results

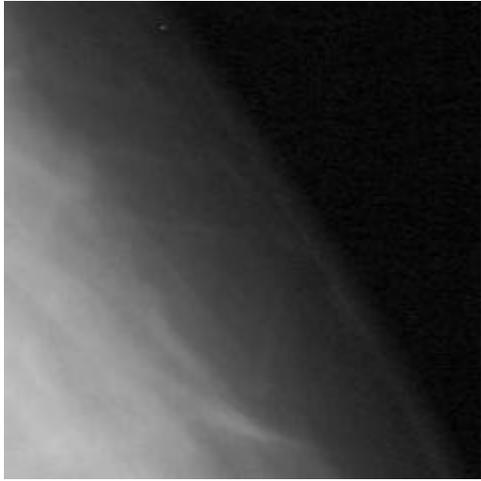
As we began to process the first set of both 8-bit and 12-bit images, the preliminary results indicated a problem, one which we had not encountered before. The reduction in data, going from the input images to their corresponding FARs, was far less than expected. A revealing clue was the fact that unexposed regions of the digitized film [e.g., subimage in the 4<sup>th</sup> row, 6<sup>th</sup> column of Fig. 3] were generating data reduction numbers well below the expected range of 95%-100%. These results were clearly unacceptable. A close scrutiny of some of the subimages revealed that they were contaminated with high levels of spatially-correlated, random noise, Fig. 4. The noise was deemed spatially correlated because its level seemed to increase as one moved left to right, from one column of subimages to the adjacent columns. This discovery was interesting because it implicated the film scanner as a potential source of this noise.

Whatever the source of the noise, its effects had to be minimized in order for fractal encoding to produce favorable results. It should be clear that fractal encoding cannot characterize a random pattern of noise efficiently. Because finding similar subregions in such a pattern is expectedly difficult, almost all subregions are partitioned to the maximum level and, therefore, included as FARs. This obviously results in a low data reduction number. To mitigate this problem, we began by employing the simplest noise removal technique available (i.e., neighborhood averaging), with the caveat that, depending on the outcome, a more complicated noise removal technique may be needed.

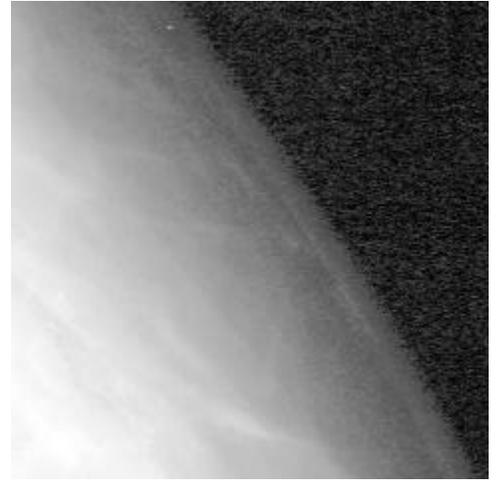
To minimize the loss of signal information (i.e., subtle microcalcifications) while removing sufficient amounts of noise, a 3x3 window size was selected, and all of the subimages were subjected to neighborhood averaging. Subsequent fractal encoding and FAR generation of these processed subimages yielded the following results.

- The preprocessing module produced an average data reduction of 83% (minimum of 64% and maximum of 94%) over all 80 mammograms. The image-by-image pattern of data reduction for the mammogram of Fig. 3 is presented pictorially in Fig. 5(a). The trend in data reduction for the first 15 abnormal mammograms is shown in Fig. 6. The average data reduction achieved within the breast areas (e.g., columns 1 through 4 in Fig. 3) of all mammograms was 76%.
- When applied to FARs alone, the selected detection algorithm produced 407 detections per image. Compared to 2984 detections per image when applied to the entire input images, this represents a 86% reduction in false detections. It should be noted that due to time limitations, the false detection numbers are generated using the first 15 abnormal mammograms only, Fig. 7. However, given the data reduction figure of 83%, we are confident that in the final analysis, the reduction in false detections for all 35 cases will fall within a few percentage points of the reported value of 86%.
- Out of a total of 507 marked microcalcifications, 467 fell within FARs, representing a coverage rate of 92%, Fig. 5(b), (c).
- In cases where both 8-bit and 12-bit versions of images were processed, we observed no significant differences between data reduction or coverage-rate numbers. It was then concluded that pixel depth provides no apparent advantage in our approach.

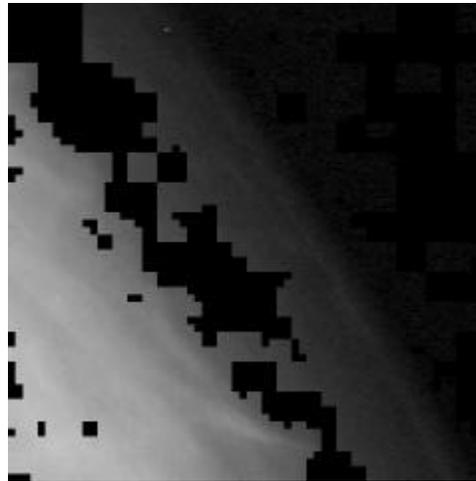
It should be mentioned that the parameters in the fractal encoding process were kept constant for all 50 mammograms. Considering the smoothing effect of the neighborhood averaging process, as well as the impact of the noise that persists even after averaging, the reported values of data reduction, false detection reduction, and coverage rate are remarkable.



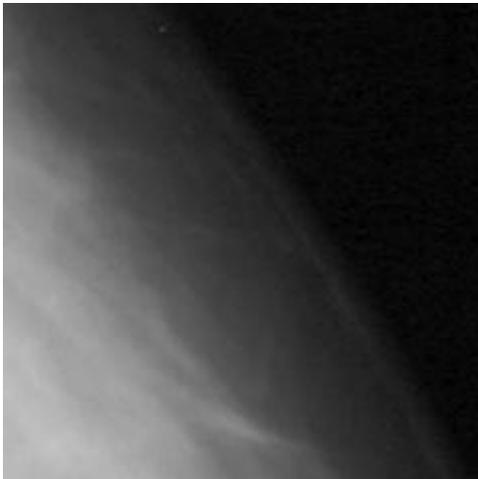
(a)



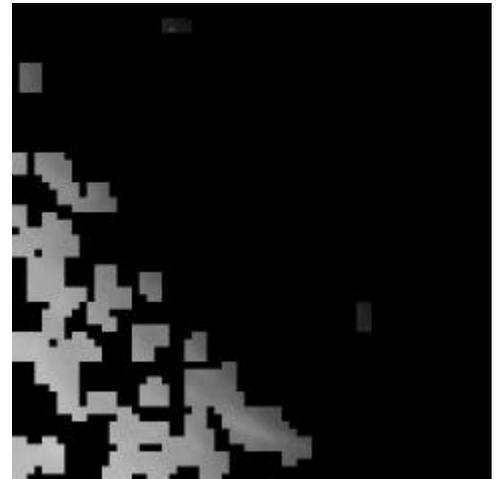
(b)



(c)



(d)



(e)

**Figure 4. (a) Input image corresponding to the subimage in the 4<sup>th</sup> row, 4<sup>th</sup> column of the mammogram in Fig. 3. (b) Histogram equalized version of the image in (a) demonstrating noise contamination. (c) FARs generated from the image in (a), representing only a 29% reduction in data. (d) Neighborhood averaged version of the image in (a). (e) FARs generated from the image in (d), representing an 85% reduction in data.**

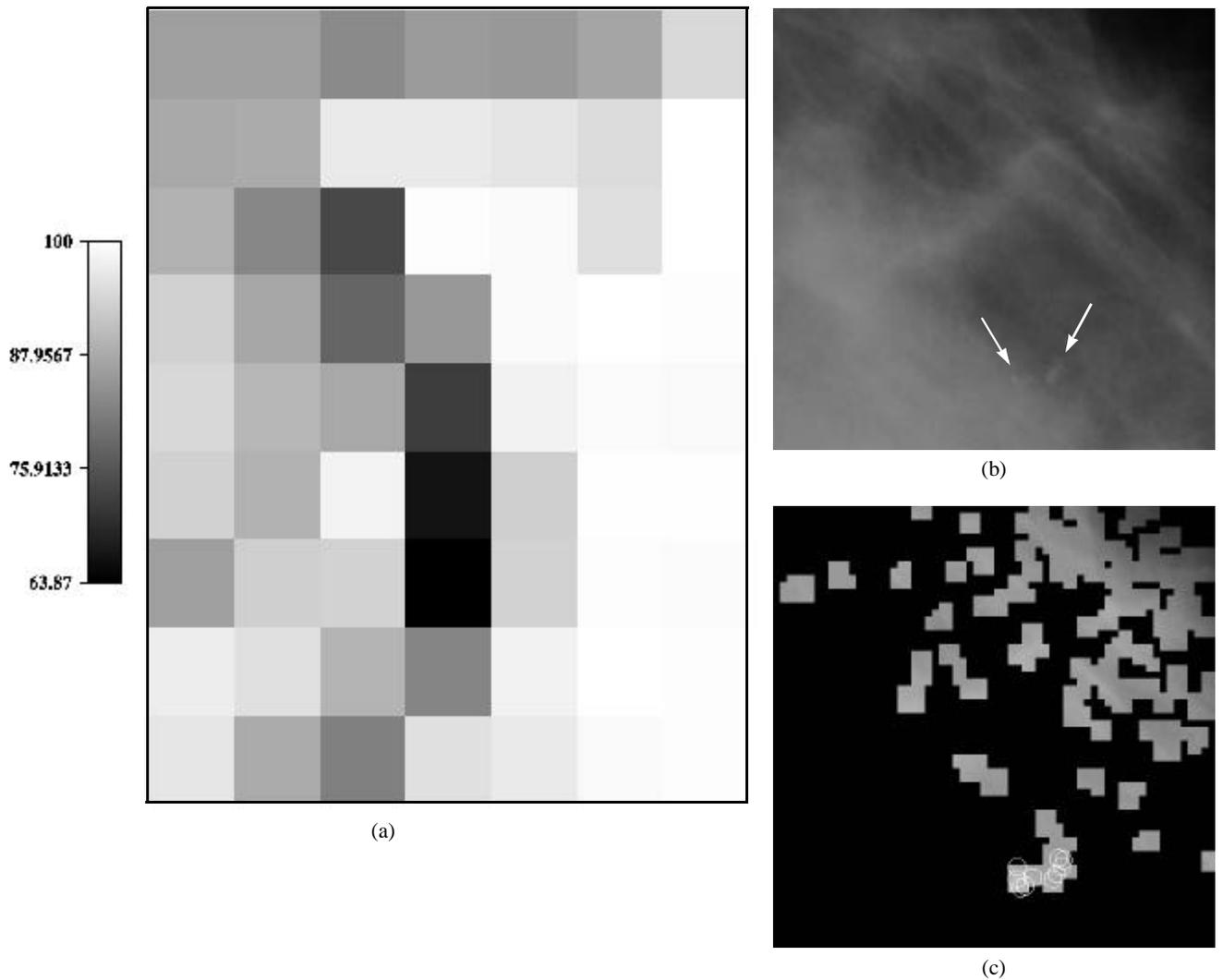


Figure 5. (a) Pattern of data reduction in the subimages of the mammogram in Fig. 3. Indicated gray levels and their corresponding values represent percent data reduction. Average reduction in data over all subimages was 92%. (b) Subimage in the 4<sup>th</sup> row, 3<sup>rd</sup> column of the mammogram in Fig. 3, which contains subtle, clustered microcalcifications near the indicated arrows. (c) FARs generated from the image in (b), indicating a perfect coverage rate of the marked microcalcifications (encircled).

#### 4. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented the results of a second pilot study with the objective of further substantiating the efficacy of front-end data reduction in CAD of mammograms. This concept is realized by a preprocessing module that can be incorporated into the front-end of most mammographic CAD systems. Based on fractal encoding, this module takes a mammographic image as its input and generates, as its output, a collection of subregions, namely FARs. These FARs contain all structures in the input image that appear to be different from the normal background tissue. Subsequently, the CAD systems need only to process the presented FARs. This accomplishes two objectives simultaneously: (1) an increase in throughput via a reduction in the input data, and (2) a reduction in false detections by limiting the scope of the detection algorithms to FARs only. The pilot study consisted of using the preprocessing module to analyze 80 mammographic images. The results were an average data reduction of 83% over all 80 images and an average false detection reduction of 86%. Furthermore, 467 of the 507 marked microcalcifications fell within FARs, representing a coverage rate of 92%.

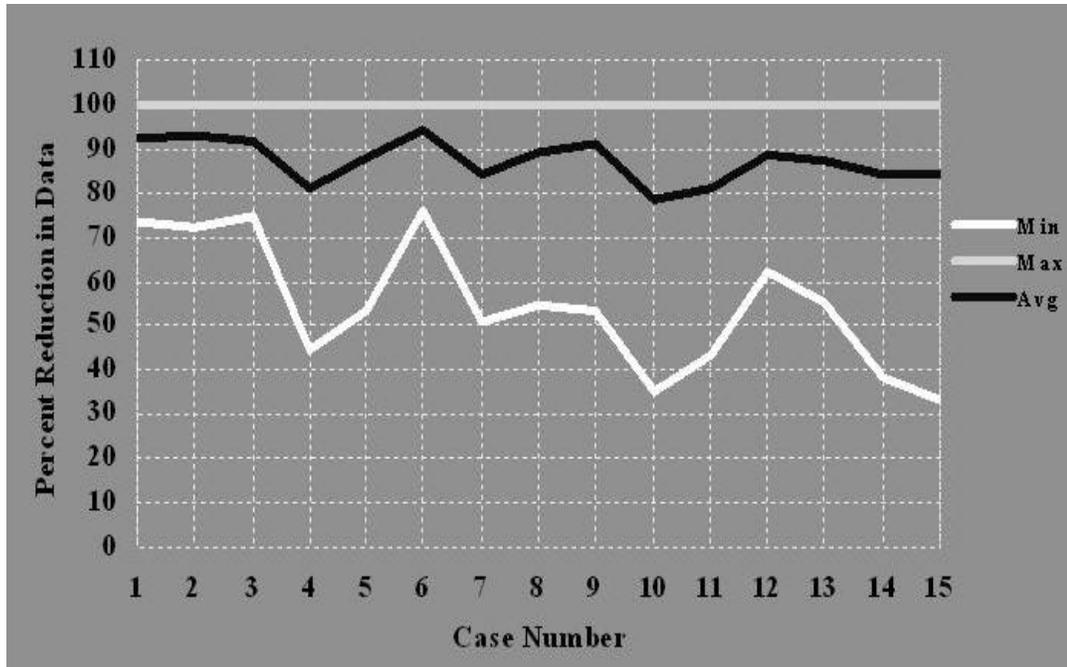


Figure 6. The trend in data reduction for the first 15 abnormal mammograms.

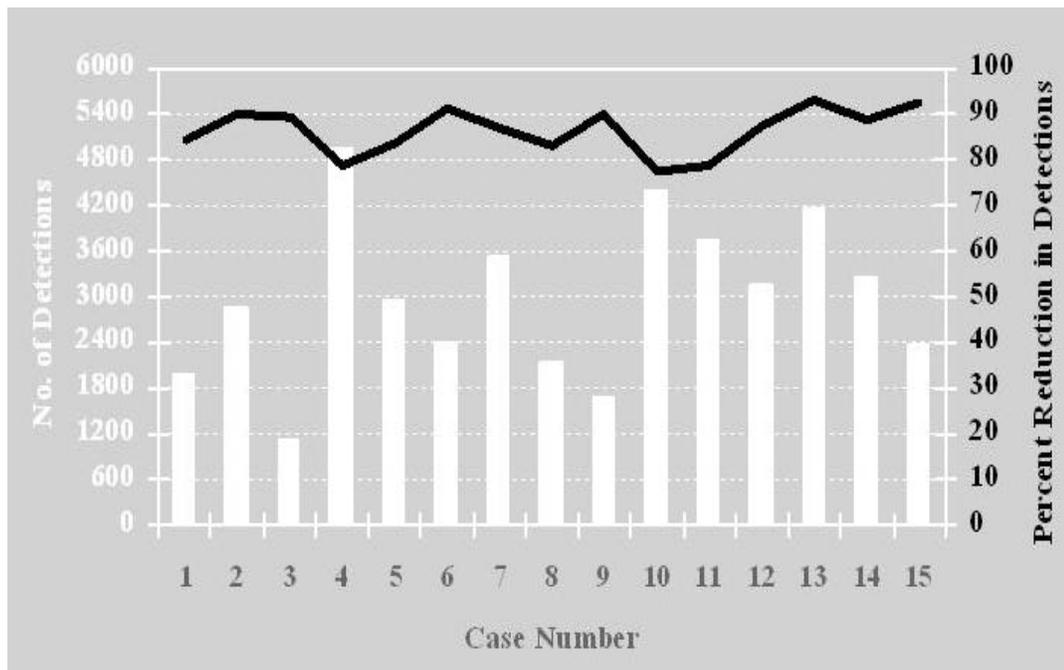


Figure 7. Number of false detections produced by the selected detection algorithm for the first 15 abnormal mammograms, and the percent reduction in false detections for each case.

The research in this area of front-end data reduction will continue on two fronts. First, we would like to tune the fractal encoding process for an optimized response to mammographic images. Thus far, we have used a standard encoding process that is generally used for image compression. Second, we would like to evaluate the performance of this approach as an integral com-

ponent of a CAD system. That is, to study its impact on the overall sensitivity and the specificity of a given CAD system.

## REFERENCES

1. The Breast Cancer Resource Center of the American Cancer Society (<http://www.cancer.org>).
2. G. Cardenosa, "Mammography: An overview," *Digital Mammography '96*, K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, eds., pp. 3-10, Elsevier, New York, 1996.
3. H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis," *Invest. Radiol.* **25**, pp. 1102-1110, 1990.
4. S. S. Gleason, H. Sari-Sarraf, K. T. Hudson, and K. F. Hubner, "Higher accuracy and throughput in computer-aided screening of mammographic microcalcifications," *IEEE Medical Imaging Conf.*, 1997.
5. H. Sari-Sarraf, S. S. Gleason, K. T. Hudson, and K. F. Hubner, "A novel approach to computer-aided diagnosis of mammographic images," *3rd IEEE Workshop on Applications of Computer Vision*, 1996.
6. The Mammographic Image Analysis Society (<http://s20c.smb.man.ac.uk/services/MIAS/MIASweb.html>).
7. H. Li, K. J. R. Liu, and S.-C. B. Lo, "Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms," *IEEE Trans. Med. Imaging* **16**, pp. 785-798, 1997.
8. Y. Fisher, "Mathematical background," *Fractal Compression: Theory and Application to Digital Images*, Y. Fisher, ed., pp. 25-53, Springer Verlag, New York, 1994.
9. Y. Fisher, "Fractal image compression with quadtrees," *Fractal Compression: Theory and Application to Digital Images*, Y. Fisher, ed., pp. 55-77, Springer Verlag, New York, 1994.
10. R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computer-aided detection of microcalcifications in mammograms: Efficient and effective method for computerized grouping of microcalcifications," *Med. Phys.* **20**, pp. 1661-1666, 1993.