

Computer Aided Diagnosis in Mammography: Its Development and Early Challenges

Brian Dolan

University of California, San Francisco
Anthropology, History & Social Medicine
3333 California Street, Suite 485
San Francisco, CA 94143-0850

Abstract: This article discusses the development of a technology of “computer vision” designed to assist medical practitioners diagnosing breast cancer in the 1980s and 1990s. A CAD system was designed to augment human vision by digitizing mammograms, enhancing computer-selected “regions of interest” and offering a protocol to recommend a course of action (follow up examination, biopsy, etc.). One issue that emerged following the introduction of CAD was that human vision—previously the “gold standard” for diagnostic accuracy—was influenced by the prompts the computer provided the interpreter, illuminating the paradoxes of de-skilling and the problems mediating visualization and expert decision making.

I. INTRODUCTION

This study is part of a larger investigation of the development and impact of medical imaging technologies on clinical skills and decision making. The research is framed by broad questions that seek to illuminate trends and challenges in the production of modern biomedical knowledge: How do scientific practices appear to become standardized and travel between different contexts and social arenas? What do illustrations obscure or hide of what they claim to represent? Does the use of illustrations to convey knowledge deny the user from interpreting or understanding the meaning of the claims in any other way?

Examining moments when new technologies are used to provide new representations of disease provides case studies to investigate how the technologies come to be accepted (or rejected) by the medical profession. This provides insight to the early collaboration between designers of technologies and their users (physicians, radiologists, etc.) and the impact of such technology on clinical actions and patient management. The particular case study discussed here developed from related research on the uses of breast MRI and a comparison between radiologists’ assessments regarding specificity and sensitivity using breast MRI as opposed to conventional X-ray mammography. It was discovered that in the period between 1995-2000, the techniques of breast MRI were being evaluated by the medical community at the same time that computer-aided diagnosis (CAD) in mammography, which had been developing over roughly the previous decade, was being evaluated. This illustrated how rapidly one technology develops and comes to compete with another before either is fully assessed or accepted as a stan-

dard of practice for a particular clinical investigation. This study sought to uncover what the perceived benefits of CAD were over conventional mammographic practices to understand the relative strengths and weaknesses of each technique which could then be used in later comparisons with alternative techniques such as breast MRI.

II. BACKGROUND TO MAMMOGRAPHY

The hypothesis for this study was that computer-aided diagnosis was adapted in response to a perceived problem with existing techniques in mammography. The first part of the study therefore sought to identify what the specific social and professional demands were to change existing practices.

The skill at the heart of radiologic practice is interpreting images. Depending on the technology used to produce them—whether X-rays, ultrasound, MRI, etc.—this involves differentiating potentially pathological from normal tissue based on variations in shades of grey on film, sound waves’ echoes on a screen, or colored pixels on a monitor. This information is then correlated with other data. The radiologist’s interpretation of the image will predominately determine whether or not the patient will be subjected to further tests, and possibly to surgery.

The history of mammography, its techniques of film-screen processing and the attention paid to the procedures by the American Cancer Society, the National Cancer Institute and the American College of Radiology, especially after the Breast Cancer Detection Demonstration Project in the 1970s is treated at length elsewhere [1]. It is worth mentioning here a couple of points regarding the recommendations that women regularly have biennial base-line mammograms after age thirty-five and annually after age fifty [2]. They point to a context that affected radiologists’ views about the applicability of enrolling new computer technology for assistance in their work.

First, radiologists expressed concern about potential problems of an increased workload generated by a national screening program. With an estimated 47 million women of screening age in the United States in 1990, it was stated that “every radiologist, regardless of expertise, would interpret 2,350 mammograms per year, or 9 per day” [3]. Unlike

other screening tests for cancer, such as those for cervical (Pap), colon (stool guaiac) or prostate (prostate-specific antigen), where the prior probability of disease was pre-screened by clinicians, breast cancer was primarily an imaging test, requiring the interpreter not only to have “a talent for perception of abnormalities” which nonetheless carried “a degree of difficulty that is underestimated by both our critics and our colleagues” [3, p. 31]. The challenge of a national screening program and the work-load that would generate was a major concern. According to a radiologist writing in the *New England Journal of Medicine* in 1986, because of the amount of work necessary to properly read mammograms, “it will be necessary, in my opinion, to use non-physician radiology assistants to interpret uncomplicated examinations. The potential political and legal ramifications of doing this are obvious” [4, p. 53]. Other than legal, there were also practical and ethical ramifications of dividing labor in this way, not least since what constituted an “uncomplicated examination” was of course not self-explanatory.

Second, there were few standards for the laboratory preparation of X-ray film and conditions of interpretation. Despite the establishment of the Mammography Accreditation Program by the American College of Radiology in 1987, which was meant to be the standard by which radiological facilities were deemed acceptable, in 1990 only 500 out of 8,000 such facilities across America had met the criteria [5]. This situation was exacerbated by the findings of a major study the following year that compared the results of 150 radiologists’ independent interpretation of the same mammograms which were found to vary widely. One thing they were being asked to identify was the presence of microcalcifications which appear as tiny white dots on film. Identifying this in the tissue was considered crucial to early detection and treatment of breast cancer, but locating them by looking at conventional mammograms was extremely difficult because of their small size and because they were often cloaked by the white mass of other dense tissue matter, hence they are often referred to as “occult breast tumors.”

Asked to select from three diagnostic categories, “normal,” “abnormal, probably benign,” and “abnormal, suggestive of cancer,” the results were alarming. In 25% of the cases as a whole, there was substantial disagreement in patient management recommendations, in which half the radiologists recommended a later follow up while the other half recommended a biopsy. In 9% of the cases, the radiologists all agreed on a biopsy, but disagreed on whether it was for the left or the right breast. These findings came out alongside seven other studies that showed that up to 30% of carcinomas were missed by radiologists during routine screening [6].

It was at this point that a possible solution to radiologists’ and patients’ problems was presented. The concept was to remove much of the complexity of interpretation, and the time spent reviewing thousands of images, by transferring human skill to another agency. Precisely because diagnosis concentrated on pattern recognition in images, it was sug-

gested that a combination of new digital technologies could be implemented to allow a computer to perform the job of an expert radiologist. The introduction of a new “expert system” in medicine was hailed as the beginnings of a possible revolution in CAD.

III. EARLY DEVELOPMENT OF CAD IN MAMMOGRAPHY

Throughout the early 1980s, the literature on computer-aided diagnosis was remarkably scarce, but with new developments in computer technology, especially laser scanning and printing, mammography entered a phase of rapidly “going digital.” This involved scanning conventional film-screen mammograms, using a computer program to enhance clusters of pixels with shapes and colors of interest (colors being shades of grey), and print out a high resolution image for the radiologist to review. In 1987, a group from the Laboratories for Radiologic Image Research, at the University of Chicago, published the first of a number of articles that discussed CAD with reference to the potential uses of new computer technology. As they stated, “The efficiency and effectiveness of the screening process may be greatly increased if an automated computer system can be successfully employed for the detection of microcalcifications” [7, p. 538].

The research group’s computer-aided approach involved multiple stages of image acquisition and enhancement. In brief, the first step was to obtain a digitalized screen-film mammogram with a Fuji drum scanner which produced an image with 1024 gray levels with the size of the region that contained the breast image being approximately 1000 X 1800 pixels. Once the digitalized image was obtained, it was then re-processed with two filters. The first filter produced an image with enhanced characteristics: enhanced pixels that matched the pre-programmed size and contrast variations of “a typical breast microcalcification.” But, as the team noted, “since the size and shape of microcalcifications vary, it is not possible to design filters that exactly match each different microcalcification” [7, p. 539]. Thus the “match” between a cluster of pixels with a certain size, shape and gray-scale value and a microcalcification necessarily relied upon an approximation of what a “typical” microcalcification looked like in the mammograms the radiologists had previously studied.

The second filter did the opposite to the first. Using a signal suppression filter, the pixels that had values representing all non-microcalcification characteristics were kept, while everything else—which corresponded to what might be microcalcifications—was eliminated from the image. At this point, the digital image is only as useful as the programming that guided the double-filtering process. Because the “difference image” is enhanced and altered according to the way the computer is programmed to “see” (or ignore) microcalcifications, any other characteristics of the breast that existed can no longer be considered part of the image. This, of course, is the point of computer-aided diagnosis centered on identifying microcalcifications: the program follows instructions rigidly, and is allegedly never “distracted” from its

field of vision. Indeed, the next stage in digital enhancement is to produce a threshold image, in which groups of two or three pixels with values corresponding to what approximates microcalcifications (which “generally” are less than half a millimeter in length) are superimposed on “an absolutely uniform background” [7, p. 540].

What remained to be tested, however, was whether the programming (which assigned values to the shape, color, and size of pixel groups) provided a reliable and accurate guide to identifying microcalcifications. To determine the true-positive and false-positive results, the research team used a computer program which generated simulated microcalcifications and placed them on the images for the computer algorithm to detect. The results were as high as 90% true-positive detection (whereas unaided radiologists fell between 70-90%). But that was a simulation. Further tests needed to be done to determine, for instance, how to prevent the computer from misidentifying other breast structures such as fibrous strands, ducts or skin folds which have similar appearance (and therefore might be assigned the same values) to microcalcifications. This would involve substantial amounts of programming, and at the time, the “computational requirements of digital mammography and computer analysis of mammographic images limited the practical application of their techniques.” [8, p. 701] However, it did not take long for fresh enthusiasm to emerge, which occurred when a new technique was developed in the field of artificial intelligence in medicine (AIM).

Adding Intelligence

Three years after the 1987 publication describing the uses of computer-aided diagnosis in mammography (described above), the same research group at the University of Chicago published the results of a number of tests assessing the potential usefulness of artificial neural networks (ANN) to assist diagnosis—aiding observation in the areas of chest radiology and digital mammography. A neural network was, the research group explained, “a computational model based on the brain; it is a powerful tool for pattern recognition” [9, p. 857]. In much the same way that doctors were understood to make decisions—by weighing evidence presented to them, drawing on past experience with similar cases, then making a diagnostic prediction—they explained that a neural network could be programmed to evaluate evidence and even learn from its own experience.

ANN consists of a set of processing units (called nodes) which are interconnected via paths that allow inputted data of different “weights” to travel through the network in parallel as well as serially, performing a non-linear calculation (analogous to neurons and synaptic connections in the human nervous system). Each incoming signal to a node—each piece of information from a dataset inputted to the neural network—is given a numeric value (a weight) assigned by certain highly skilled medical professionals (who become the gold standard). The neural network is initially programmed by inputting a certain number (the higher the better) of examples whereby each piece of clinical information

that went into making a correct diagnosis (patient’s age, sex, symptoms, and array of radiographic signs) is weighted and the neural paths accordingly programmed to yield a correct diagnosis. Then, each time the computer is fed a dataset, it (like a young medical student) can weigh the evidence according to what it has learned from the previously inputted examples and provide a diagnosis. If it encounters information that it has not been programmed to weigh, or the weight of the sum of evidence does not have a predetermined output path, and it provides an incorrect diagnosis, “these new data can be incorporated into the data base along with the correct diagnosis so that very similar cases, which may be encountered subsequently, will be correctly identified” [9, p. 860]. In this way, the neural network is said to learn from its own mistakes.

In a number of studies on the computerized detection of clustered microcalcifications in digital mammograms [10, 11] the team from the University of Chicago had an expert radiologist locate “true” microcalcifications in digitized mammograms. Forty-three radiographic features defining calcifications were selected as weighted inputs to the neural network. These features ranged from the shape, size, and pattern of breast masses to the number, uniformity, and distribution of calcifications. Then, 133 “textbook cases” were selected from a published mammography atlas as a training data base, whereupon an experienced mammographer assigned a value to each of the forty-three features and programmed the correct diagnosis (the “truth”). To test the system, sixty separate clinical cases showing abnormalities that were independently proved to be either a mass, a cluster of microcalcifications, or another abnormality, were fed into the computer system, and its performance was compared to the interpretation of attending radiologists. The neural network performed with higher sensitivity (probability of diagnosing a malignant lesion), higher specificity (probability of correctly diagnosing a benign lesion), and higher positive predictive value than the average performance of the radiologists. “Therefore,” the research team concluded, “the neural network, working with features extracted by an experienced mammographer, appeared to be able to recommend an appropriate course of action better than the average radiologist, the average resident, or the experienced mammographer himself” [11, p. 86].

The results were encouraging to advocates of CAD. It promised to put the process of identifying diseases and acting on their treatment into warp drive. The message resonating in dozens of articles published between 1990 and 1995 was that ANN was simply far smarter than any human.

IV. CHALLENGES TO THE CAD SYSTEM

Although the use of ANN as an automated classifier was celebrated by some, there remained drawbacks and untested parameters, most significantly being the fact that the values of the extracted features used for training were, first, not exhaustive, and second, they relied on the subjective interpretation of the inputting radiologist. In other words, there was no comprehensive or standardized way to classify the

digital representation of disease. This issue had emerged when mammograms were first digitized for computer analysis, where the brightness of each pixel was assigned a value (the “grey level”) which was recognized and analyzed by the computer for the “enhancement of *meaningful structure*, the quantitative description of image characteristics and features, or the detection of *abnormalities*” [12, emphasis added]. The introduction of apparently self-learning artificial neural networks did little to draw agreement about what “meaningful structures” and “abnormalities” look like. How does one test the system if the gold standard is still the human interpreter whose own limitations are precisely what the computer is to transcend?

Nevertheless, some researchers have attempted to derive classification systems informing risk assessment based on pixel analysis, none of which are universally accepted. However, a relatively simple method was developed in 1994 for a computer to sort images automatically into two categories: those which are “easy” and those which are “difficult” to interpret, referring both to the interpretive capabilities of humans and the machines they program [12]. The usefulness of sorting images in this way was meant to quickly facilitate the decision of which cases should be referred to “the most experienced readers” and those which an assistant radiologist could interpret. While it was recognized that this “could permit the better use of the time and skills of expert radiologists,” what made the category of “difficult” images additionally useful was that it would provide research materials to continually test the capabilities of the artificial neural networks by re-entering new data. This was a step along the way to creating what a research group in 1999 considered a diagnostic aid that went beyond human perceptual features in the first study to evaluate the clinical potential of computer classification based on computer extracted features completely independent from radiologists’ interpretation of mammograms [13].

Such work, which other radiologists have described in terms that make it clear that computers will always supplement, and not replace, human experts, nevertheless suggests a possibility that computer-aided diagnosis could assume a status whereby their operations will be taken for granted by virtue of the fact that they can assimilate and calculate more quickly than any human brain, and work “independently” and without fatigue.

However, those who did not accept that computers were superior to humans—just as technologists were not considered to be *better* than expert radiologists at decision making (but were good at catching the occasional error)—expressed another reason why caution should be taken in the use of such systems which pointed to the ethics of deferring diagnostic logic or clinical cognition to another agency. “Merely feeding clinical data into a computer and reading the result,” explained a clinician writing in the *Lancet*, “irrespective of the method used to derive it, could undermine the clinician’s ability to take personal responsibility for clinical decisions” [14]. In particular, according to another contributor in the same journal, “in negligence cases courts may consider

black-box systems as products not services, forcing developers to take on strict liability because they interfere with the ability of a professional to act as a ‘learned intermediary’” [15, p. 1176]. Thus, however purportedly intelligent such ‘black boxes’ were, they faced difficulty gaining acceptance since the designers conceived of the benefits of the support tools according to their own idealized view of how medical decision making should be done, rather than being based on the needs of the clinicians and their patients.

Unintended Consequences

Rather than embraced as a welcome aid to their practice, some radiologists considered such systems an “interference” and more burdensome on their time. Even the foremost advocates of CAD acknowledged that “techniques for the computer detection of mammographic abnormalities vary markedly in their structure and execution ... [and] require that a number of empirical decisions be made regarding parameters that occur during the execution of the program” [8, p. 705]. And finally, once all the work had been done in attempt to standardize and calibrate the machines to enhance their performance, the concern then becomes one of how much influence the “computer vision” will have on human vision. While designers of intelligent systems assured the radiology community that “the radiologist using the workstation will always make the final decision” [16], new studies were beginning to find that computer-aided “prompting may affect both the performance and the visual search behavior of radiologists interpreting mammograms” [17]. Once the computer suggests a “region of interest” to examine, a behavioral pattern emerged whereby human interpreters merely followed the machine’s instructions for where to look, spending little time elsewhere on the film, leading to potential oversight of other features that would be searched for in conventional mammography.

V. CONCLUSION

By 1995, a number of research groups in America and Europe were testing artificial neural networks (programs designed by different computer software companies) for analyzing a variety of medical datasets. However, a survey of over 200 articles published between 1990-1995 which discussed ANN yielded from a Medline search reveals that none of them described an actual clinical trial using neural networks. Assessments of the uses of neural networks in computer-aided diagnosis remained centered on controlled trials of ‘textbook’ cases. But even in these simulated settings, findings showed that the performance of neural networks varied considerably, alerting programmers and medical practitioners alike to the fact that the selection of data inputted to create an optimal network was no trivial task. “Despite claims made by software vendors,” stated one author in the *Lancet*, “building and testing any decision aid takes great skill, and statistical support is essential” [15, p. 1177].

Rather than freeing up experts’ time and facilitating complex decision making, these early findings suggested that

more statistical calculations and diligence was required to make the systems work. One area that was singled out as being particularly problematic for machine learning was in the case of training a computer to see morphological features on a cervical cytology slide. As another article in the series in the *Lancet* which evaluated ANN reported, "Not only is there no absolute consensus on what features of a cell or of a cytological smear contribute to an assessment of normality but also it is unclear exactly where the bounds of such 'normality' begin and end. Histological images are so complex that their automated interpretation requires a highly adaptive mathematical procedure" [18, p. 1203].

One explanation for the variability in neural networks (and other computer-aided diagnosis systems) was the fact that no standards had been set for the data inputted to the computer which allowed it to "see" correctly and make appropriate recommendations to the practitioner. Ironically, this is analogous to the lack of standards for mammography film developing and viewing conditions that challenged radiologists' abilities until the Standards Act of 1992. Furthermore, when reducing the decision-making process to a rule-bound formula, it was unclear which, if any, features of a digitized image should be used to indicate the existence of an actual pathological problem.

While one problem with human observers was their variable skills, neural network performance was likewise degraded by the input of poor data. "Experience," it was stated, "is just as important for a neural network as it is for man" [18, p. 1204]. This point was reiterated in another article that pointed out that "there is not one ANN, but an infinite number, and art, logic, and luck are all involved in selecting a useful one. The careful fitting, pruning, and general tweaking of an ANN when applied to a real problem requires just as much experience and training as any other analytical process. ANNs therefore, rather than removing the need for statistical modeling skills merely replace them with a different form of experience" [19].

These issues are central to the application of computer-aided diagnosis and neural networks to mammography since feature extraction, defining representational 'normality' and assigning pixel values are essential to image interpretation. But in the process of standardizing software, datasets, and pathological features, new sets of skills and criteria of decision-making emerge which place new demands on the human experts for whom computer vision was intended to assist.

ACKNOWLEDGEMENTS

I would like to thank Dr. Mia Markey for organizing the session at Asilomar, and my colleagues at UCSF for their helpful comments.

REFERENCES

- [1] B. Lerner, *The Breast Cancer Wars*. Oxford: Oxford University Press, 2003.
- [2] American Cancer Society, "Proceedings of the workshop on cost of screening mammography," *Cancer* vol. 60, pp. 1669-1702, 1987.
- [3] B. Monsees, "Screening mammography: who will meet the need?" *Radiology*, vol. 184, pp. 30-31, 1992.
- [4] F. Hall, "Screening mammography: potential problems on the horizon," *N. Eng. J. Med.*, vol. 134, pp. 53-55, 1986.
- [5] F. Houn, K. Franke, M. Elliott, C. Finder, R. Burkhart, R. Fischer, "The mammography Quality Standards Act of 1992: history and process," *Food and Drug Law Journal*, vol. 50, pp. 485-492, 1995.
- [6] J. Elmore, C. Wells, C. Lee, D. Howard, A. Feinstein, "Variability in radiologists' interpretations of mammograms," *N. Eng. J. Med.*, vol. 331, pp. 1493-1499, 1994.
- [7] H.-P. Chan, K. Doi, S. Galhotra, C. Vyborny, H. MacMahon, P. Jokich, "Image feature analysis and computer-aided diagnosis in digital radiography: automated detection of microcalcifications in mammography," *Med Physics*, vol. 14, pp. 538-548, 1987.
- [8] C. Vyborny and M. Giger, "Computer vision and artificial intelligence in mammography," *AJR: Am J of Roentgenology*, vol. 162, pp. 699-708, 1994.
- [9] N. Asada, K. Doi, H. MacMahon, "Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung disease: pilot study," *Radiology*, vol. 177, pp. 857-860, 1990.
- [10] Y. Wu, K. Doi, M. Giger, R. Hishikawa, "Computerized detection of clustered microcalcifications in digital mammograms: application of artificial neural networks," *Med Physics*, vol. 19, pp. 555-560, 1992.
- [11] Y. Wu, M. Giger, K. Doi, C. Vyborny, R. Schmidt, C. Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81-87, 1993.
- [12] P. Taylor, S. Hajnal, M. Dilhuydy, B. Barreau, "Measuring image texture to separate 'difficult' from 'easy' mammograms," *Brit J Radiology*, vol. 67, pp. 456-463, 1994.
- [13] Y. Jiang, R. Nishikawa, R. Schmidt, C. Metz, M. Giger, K. Doi, "Improving breast cancer diagnosis with computer aided diagnosis," *Academic Radiology*, vol. 6, pp. 22-23, 1999.
- [14] S.R. Dadds, "Neural Networks," *Lancet*, vol. 346, pp. 1500-1501, 1995.
- [15] J. Wyatt, "Nervous about artificial neural networks?" *Lancet*, vol. 346, pp. 1175-1176, 1995.
- [16] R. Nishikawa, R. A. Schmidt, J. Papaioannou, R. Osnis, R. A. Halde mann Heusler, M. L. Giger, D. E. Wolverton, C. Comstock, K. Doi, "Performance of a prototype clinical 'intelligent' mammography workstation," in *Digital Mammography '96*, K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, Eds. Amsterdam: Elsevier, 1996.
- [17] M. Mugglestone, R. Lomax, A. Gale, A. Wilson, "The effect of prompting mammographic abnormalities on the human observer," in *Digital Mammography '96*, K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, Eds. Amsterdam: Elsevier, 1996.
- [18] R. Dybowski and V. Gant, "Artificial neural networks in pathology and medical laboratories," *Lancet*, vol. 346, pp. 1203-1207, 1995.
- [19] D. Signorini and J. Slattery, "Neural networks," *Lancet*, vol. 346, p. 1500, 1995.