ABSTRACT

Title of dissertation: Similarity Classification and Retrieval in Cancer Images and Informatics

David Alan Tahmoush, Doctor of Philosophy, 2008

Dissertation directed by: Professor Hanan Samet
Computer Science Department

Techniques in image similarity, classification, and retrieval of breast cancer images and informatics are presented in this thesis. Breast cancer images in the mammogram modality have a lot of non-cancerous structures that are similar to cancer, which makes them especially difficult to work with. Only the cancerous part of the image is relevant, so the techniques must learn to recognize cancer in noisy mammograms and extract features from that cancer to classify or retrieve similar images. There are also many types or classes of cancer with different characteristics over which the system must work. Mammograms come in sets of four, two images of each breast, which enables comparison of the left and right breast images to help determine relevant features and remove irrelevant features. Image feature comparisons are used to create a similarity function that works well in the high-dimensional space of image features. The similarity function is learned on an underlying clustering and then integrated to produce an agglomeration that is relevant to the images. This technique diagnoses breast cancer more accurately than commercial systems and other published results. In order to collect new data and capture the medical diagnosis used to create and improve these methods, as well as develop relevant feedback, an innovative image retrieval, diagnosis capture, and multiple image viewing tool is presented to fulfill the needs of radiologists. Additionally, retrieval and classification of prostate cancer data is improved using new high-dimensional techniques like dimensionally-limited distance functions and dimensional choice.

Similarity Classification and Retrieval in Cancer Images and Informatics

by

David Alan Tahmoush

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:

Professor Hanan Samet, Chair
Professor Jim Purtilo
Professor Amitabh Varshney
Professor Mike Boyle
Associate Professor David Jacobs

Table of Contents

List of Tables

List of Figures

Acronyms, Abbreviations, Definitions

Digital Imaging and Communications in Medicine – A standard developed by the American College of Radiology-National Electrical Manufacturer's Association (ACRNEMA) to aid in the distribution and viewing of medical images, such as CAT scans, MRI results, and ultrasound images. Hereafter referred to as the DICOM standard.

Feature – An aspect of an image that is identifiable in imaging software. In Archimedes, a feature is a user-defined graphical addition to an image that can be used later as a criteria for a search.

Health Insurance Portability and Accountability Act of 1996 – Legislation passed to protect patient personal information, streamline inefficient industry practices, reduce paperwork, help detect fraud and abuse, and allow workers to change health insurance between jobs. Hereafter referred to as HIPAA.

Deidentified Image – A term used in medical fields referring to an image with all information regarding patient identity removed.

Protected Information – Information which must be kept private according to HIPAA standards. In Archimedes, this information is only accessible to designated user groups.

Metadata – Metadata refers to data which is associated with another, central piece of data, and provides additional information about the central data. In Archimedes, images are associated with metadata.

1 Introduction

A technique that radiologists use to diagnose breast cancer involves first finding suspicious sites in the mammograms and then comparing the left and right breasts to reduce the number of false positives. The symmetry of the human body is utilized to increase the accuracy of the diagnosis through visual registration of the mammograms.

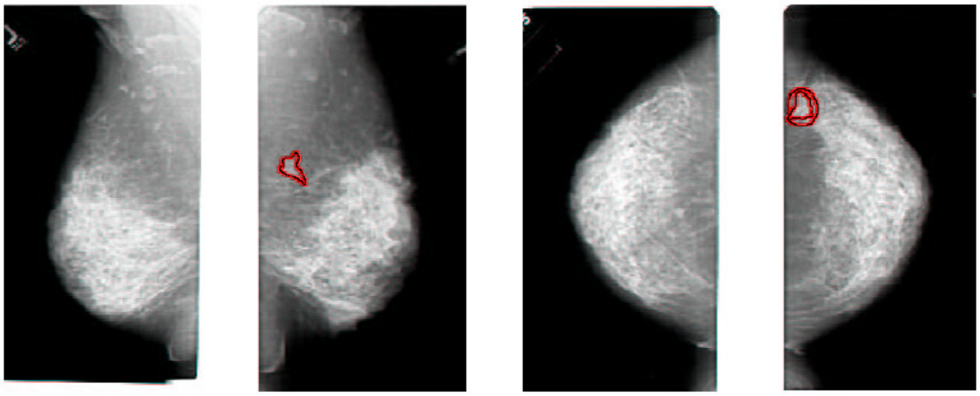
This technique is emulated by combining both computer vision and learning techniques, attempting to capture the diagnosis of the radiologist. Therefore this thesis is motivated not only by computer science theory and technique, but also by domain-specific knowledge and theory. These ideas were verified through a simple approach that has been completed with surprisingly good results at diagnosing breast cancer that is described in Chapter 3. It is hoped that this thesis will improve techniques in image similarity and CBIR, as well as provide insights into medical imaging and especially into the imaging of breast cancer.

Breast cancer remains a leading cause of cancer deaths among women in many parts of the world. In the United States alone, over forty thousand women die of the disease each year [5]. Mammography is currently the most effective method for early detection of breast cancer [77], and example mammograms are shown in Figure 1. For two-thirds of the women whose initial diagnosis of their mammogram is negative but who actually have breast cancer, the cancer is evident upon a second diagnosis of their mammogram [77]. Computer-aided detection (CAD) of mammograms could be used to avoid these missed diagnoses, and has been shown to increase the number of cancers detected by more than nineteen percent [41], so there is hope that improving techniques in computerized detection of breast cancer could significantly improve the lives of women across the globe. Asymmetry, which consists of a comparison of the left and right breast images [39], is a technique that has been neglected in CAD but could be used to significantly improve the results. An automated prescreening system only classifies a mammogram as either normal or suspicious, while CAD picks out specific points as cancerous [12]. One of the most challenging problems with prescreening is the lack of sensitive algorithms for the detection of asymmetry [11]. Image similarity methods can capture the asymmetry properties, and then improve both CAD and prescreening of breast cancer.

Contextual and spatial comparisons can be combined to determine image similarity, which has been often utilized for content-based image retrieval (CBIR) from image databases [34, 45, 49, 103]. Medical image databases have also used image similarity, ranging from rule-based systems for chest radiographs [107] to anatomical structure matching for 3-D MR images [50] to learning techniques [48]. However, the focus is often on the non-cancerous structures, while it is the cancerous structures that are of principle interest. This thesis applies image similarity concepts to the problem of detecting breast cancer in mammograms and CBIR.

Detecting breast cancer in mammograms is challenging because the cancerous structures have many features in common with normal breast tissue. This means that a high number of false positives or false negatives are possible. Asymmetry can be used to help reduce the number of false positives so that true positives are more obvious. Previous work utilizing asymmetry has used wavelets or structural clues to detect asymmetry with correct results as often as 77% of the time [39, 83]. Additional work has focused on bilateral or temporal subtraction, which is the attempt to subtract one breast image from the other [114, 123]. This approach is hampered by the necessity of exact registration and the natural asymmetry of the breasts. Bilateral subtraction tries to utilize the multiple images taken with the same machine by the same technician and analyzed using the same process in an effort to reduce the systematic differences that can be introduced. Developing ways to better utilize asymmetry is consistent with a philosophy of trying to use methods that can capture measures deemed important by doctors thereby building upon their knowledge base, instead of trying to supplant it. However, measuring asymmetry involves registration and comparing multiple images, and thus it is a more complicated process.

Registration is the matching of points, pixels, or structures in one image to another image. Registration of mammograms is difficult because mammograms are projections of compressed three-dimensional structures. Primary sources of misregistration are differences in positioning and compression, which manifests itself in visually different images. The problem is more complex because the breast is elastic and subject to compression. Additional sources of difficulties include the lack of clearly defined landmarks and the normal variations between breasts. Strictly speaking, precise mammogram registration is intractable. However, an approximate solution is possible [110]. Warping techniques have been used [96], as well as statistical models [121] or mutual information as a basis for registration [115]. The technique used in this thesis learns image comparison models based upon clustering that encapsulate an approximate registration and uses them to compare the mammograms of the left and right breasts. This method also avoids direct registration by applying a similarity technique to measure the image similarity and build a CBIR system.



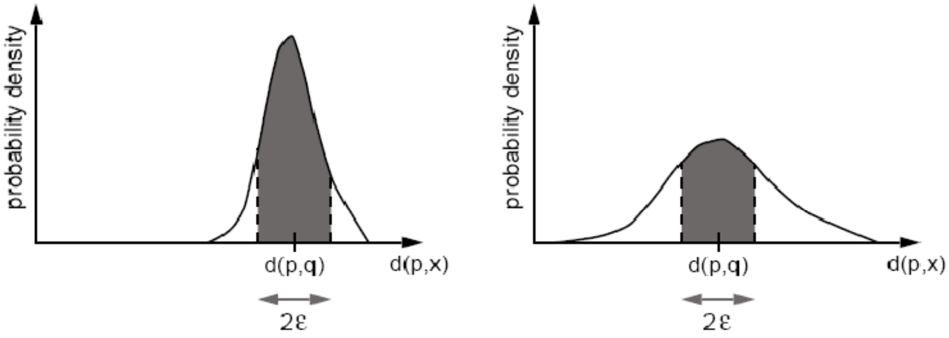
(a) (b) (c) (d)

Figure 1: The typical set of four images that make up a mammogram, the side view of the left breast in (a), the side view of the right breast in (b), the top view of the left breast in (c), the top view of the right breast in (d). The cancerous areas are outlined in red. Since the images come in sets, the non-cancerous cases are examples of similar images, while the cancerous cases are examples of dissimilar images, and these examples can be used to determine image similarity. Note that the textures of the cancer are very similar to non-cancerous areas, which is why image comparisons are so important in the analysis of mammograms. Also note that the cancer is apparent in both images of the same breast, which provides additional information for the analysis. This image set was correctly classified by the method described in Chapter 3.

In order to create an effective CBIR and cancer diagnosis technique for breast cancer images, methods are needed to efficiently retrieve data based on similarity to a given exemplar or set of exemplars. The retrieval is generally in response to queries such as the following:

1. Finding objects having particular feature values (point queries).
2. Finding objects whose feature values fall within a given range or where the distance from some query object falls into a certain range (range queries).
3. Finding objects whose features have values similar to those of a given query object or set of query objects (nearest neighbor queries).
4. Finding pairs of objects from the same set or different sets which are sufficiently similar to each other ('all closest pairs' queries). This is also a variant of a more general query commonly known as a spatial join query.

These queries are collectively referred to as similarity retrieval, and supporting them is the subject of this proposal. Of these queries, the nearest neighbor query is particularly important, and it is the one that is emphasized. An apparently straightforward solution to finding the nearest neighbor is to compute a Voronoi diagram for the data points (i.e., a partition of the space into regions where all points in the region are closer to the region's associated data point than to any other data point), and then locate the Voronoi region corresponding to the query point. The problem with this solution is that the combinatorial complexity of the search process in high dimensions, expressed in terms of the number of objects, is prohibitive thereby making it virtually impossible to store the Voronoi diagram which renders its applicability moot.



(a) (b)

Figure 2: A probability density function (analogous to a histogram) of the distances d(p,x) with the shaded area corresponding to |d(q,p)-d(p,x)|/ε. (a) indicates a density function where the distance values have a small variation, while (b) indicates a more uniform distribution of distance values thereby resulting in a more effective use of the triangle inequality to prune objects from consideration as satisfying the range search query.

The problem described above is typical of the issues that must be faced when dealing with high-dimensional data. Generally speaking, multidimensional problems such as these queries become increasingly more difficult to solve as the dimensionality increases. The difficulties that are encountered are attributed to the curse of dimensionality which surfaces in a number of different forms. In essence, the term was coined by Bellman [20] to indicate that the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with the number of variables (i.e., dimensions) that comprise it. For similarity searching (i.e., finding nearest neighbors), this means that the number of objects (i.e., points) in the data set that need to be examined in deriving the estimate (i.e., the nearest neighbor) grows exponentially with the underlying dimension.

The curse of dimensionality has a direct bearing on similarity retrieval in high dimensions in the sense that it raises the issue of whether or not nearest neighbor searching is even meaningful in such an environment. In particular, it has been shown that for data and queries drawn from a uniform distribution, the distance to the nearest neighbor and the distance to the farthest neighbor tend to converge as the dimension increases [38]. This is why dimension reduction is an important issue in classification.

Assuming that the distance d is a distance metric (which is the case for the commonly used Minkowski metric L_p), and hence that the triangle inequality holds, an alternative way of understanding the ramifications of the curse of dimensionality is to observe that when dealing with high-dimensional data, the probability density function (analogous to a histogram) of the distances of the various elements is more concentrated and has a larger mean value. This means that similarity searching algorithms will have to perform more work. In the worst case, for an arbitrary object x, there is the situation where $d(x,x)=0$ and $d(x,y)=1$ for all y x, which means that a similarity query must compare the query object with every object of the set. One way to see why more concentrated probability densities lead to more complex similarity searching is to observe that this means that the triangle inequality cannot be used so often to eliminate objects from consideration. In particular, the triangle inequality implies that every element x such that $|d(q,p)-d(p,x)|> \varepsilon$ cannot be at a distance of ε or less from q (i.e., from $d(q,p) < d(p,x)+d(q,x)$). For the probability density function of d(p,x), when ε is small while the probability density function is large at d(p,q), then the probability of eliminating an element from consideration via the use of the triangle inequality is the remaining area under the curve, which is quite small (see Figure 2a in contrast to Figure 2b where the density function of the distances is more uniform).

The high dimensionality of the data also has an effect on the search process which is aided by the presence of indexes. In particular, for uniformly distributed high-dimensional data, most of the data lies near the boundary of the underlying space (e.g., [15]) and thus most indexes result in visiting all of the index blocks. This has led to the use of methods based on a sequential scan (e.g., [16, 38, 112]). However, these methods also make use of a variant of an index in the sense that they resort to the use a compressed index on the data to speed up the sequential scan.

A number of methods have been proposed to overcome the curse of dimensionality. One approach is to observe that the data is rarely uniformly distributed which leads to pointing out that some dimensions are more significant than others thereby focusing on them (e.g., [44, 55, 66]). Such methods are also known as dimension-reduction techniques and some examples include SVD [47] and the Discrete Fourier Transform (DFT) [43]. The traditional and the state-of-the-art dimensionality reduction methods can be generally classified into feature extraction [78, 80, 86] and feature selection [19, 29, 120] approaches. In general, feature extraction approaches are more effective than the feature selection techniques [32, 111, 117] and they have shown to be very effective for real-world dimensionality reduction problems [37, 60, 78, 80]. Many scalable online FE algorithms have been proposed. Incremental PCA (IPCA) [7, 81] is a well-studied incremental learning algorithm. The latest version of IPCA is called Candid Covariance-free Incremental Principal Component Analysis (CCIPCA) [113]. However, IPCA ignores the valuable label information of data and is not optimal for general classification tasks. The Incremental Linear Discriminant Analysis (ILDA) [56] algorithm has also been proposed recently. Another feature extraction algorithm is called Incremental Maximum Margin Criterion (IMMC) [116].

The above discussion about the effects of the curse of dimensionality implicitly assumed that the data lies in a vector space. In other words, is is based on the premise that the features that describe the objects (and hence the dimensionality of the underlying feature space) are known. In fact, it is often quite difficult to identify the features and in such a case the only available information is a distance function d that indicates the degree of similarity (or dis-similarity) between all pairs of objects, given a set of N objects. Usually d is required to obey the triangle inequality, be non-negative, and be symmetric, in which case it is known as a metric and also referred to as a distance metric. Some examples of distance functions that are distance metrics include edit distances such as the Levenshtein [76] and Hamming [51] distances for strings and the Hausdorff distance for images (e.g., [65]. There are two ways of performing similarity retrieval using such data. The first is to embed the data in a vector space using an embedding method such as FastMap [35], SparseMap [64], etc. [59], and then apply one of the classical spatial indexing methods similar to what is done in dimension reduction. The second is to use a distance-based index [58] such as a vp-tree [108, 122], mvp-tree [22], gh-tree [108], GNAT [23], M-tree [27], sa-tree [84], kNN graph [100], etc.

The rest of this thesis is organized as follows. Chapter 2 describes ultrasound and feature extraction from ultrasound images, as well as mammogram images and feature extraction from mammograms. Chapter 3 details a supervised k-means approach taken as an initial data exploration and classification approach, and the development of a similarity function. Chapter 4 describes the medical image database capabilities and techniques. The some work on content-based image retrieval of medical images are in Chapter 5. Appendix A contains a selection of relevant Mammographic images.

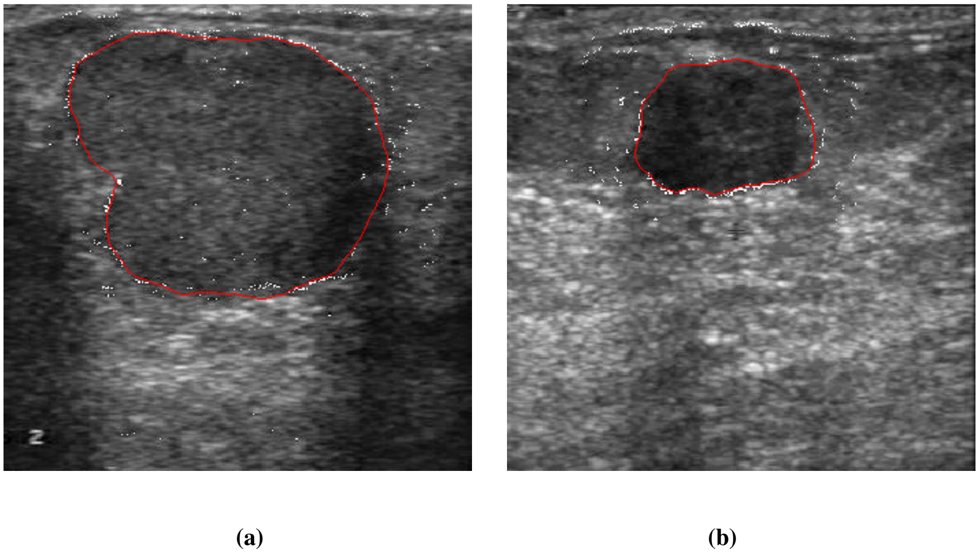
2 Data and Image Description

This Chapter gives a brief introduction to various data types and images used in this thesis.

2.1 Ultrasound

Ultrasound, also known as sonography, is an imaging method in which high-frequency sound waves are used to outline a part of the body. High-frequency sound waves are transmitted through the area of the body being studied and the sound wave echoes are picked up and translated by a computer into an image. Breast ultrasound is sometimes used to evaluate breast problems that are found during a screening or diagnostic mammogram or on physical exam, but it is not yet routinely used for screening. Ultrasound may be a helpful addition to mammography when screening women with dense breast tissue, which is difficult to evaluate by mammogram. Ultrasound is useful for evaluating some breast masses and can determine if a suspicious area is a cyst without placing a needle into it to aspirate fluid.

This thesis builds upon the experience of doctors by capturing the most relevant medical criteria for a particular imaging type, which for ultrasound are shape type, margin type, and width-to-anteroposterior (AP) dimension ratio [91]. Examples are in Figure 3. Other criteria from the literature like the Stavros Criteria [40, 106] include posterior echoes (enhanced, unaffected, or decreased), echogenicity (intensity of internal echoes), echotexture (homogeneity of internal echoes), the presence of calcifications, the presence of lateral edge refraction, and presence of a pseudocapsule. These may be included, but have not been shown to be as important as the characteristics of shape, margin, and width-to-AP ratio [91].



(a) **(b)**

Figure 3: Ultrasound images of two different cancers. These have had their margin, or boundary, determined by a collaborator. The white dots are edges that have been detected, and the red line is a boundary that has been determined. Image (a) has a lobulated shape, while image (b) has a microlobulated margin.

2.2 Mammograms

A mammogram is an x-ray exam of the breast. It is used to detect and diagnose breast cancer, both in women who have no breast complaints or symptoms and in women who have breast symptoms (problems such as a lump, pain, or nipple discharge). The special type of x-ray machine used for the breasts is different than for other parts of the body and produces x-rays that do not penetrate tissue as easily as that used for routine chest films

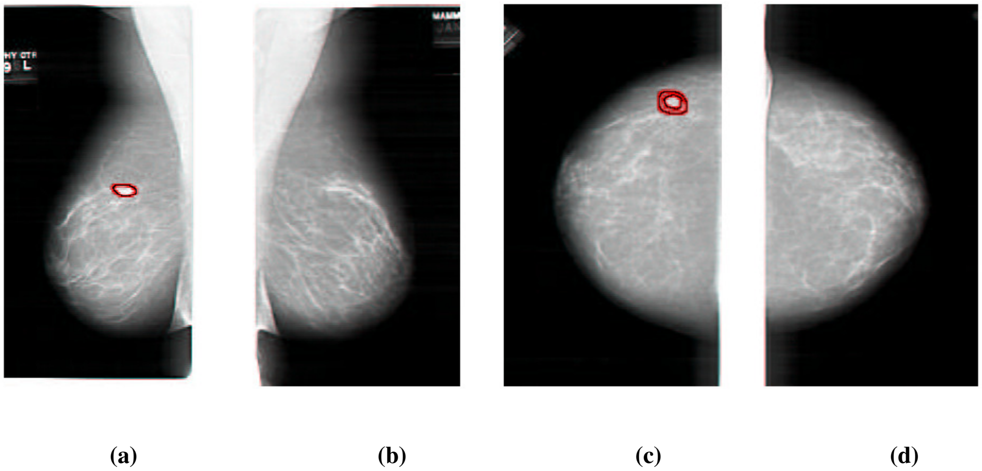


Figure 4: The typical set of four images that make up a mammogram, the side view of the left breast in (a), the side view of the right breast in (b), the top view of the left breast in (c), the top view of the right breast in (d). The cancerous areas are outlined in red. Comparing this mammograms set with Figure 1 shows some of the variation in the size and morphology of spiculated lesions. This image set was correctly classified by the method described in Chapter 3.

or x-rays of the arms or legs, but gives a better image of variations in tissue density. For a mammogram, the breast is squeezed between two plastic plates attached to the mammogram machine unit in order to spread the tissue apart. This squeezing or compression ensures that there will be very little movement, that the image is sharper, and that the exam can be done with a lower x-ray dose. However, it also makes 3-D reconstruction of the breast structure much more difficult.

Mammography produces a black and white image of the breast tissue on a large sheet of film as shown in Figure 4, which is interpreted by a radiologist, though modern mammography machines are digital. Radiologists have special training in diagnosing diseases by looking at images of the inside of the body produced using x-rays, sound waves, magnetic fields and other methods. Reading mammograms can be challenging because the appearance of the breast on a mammogram varies a great deal from woman to woman. Some breast cancers produce changes in the mammogram that are difficult to notice and can be overlooked. One of these types of cancers is spiculated lesions, which are highly malignant, and upon which this thesis focuses. It is very important for the radiologist to have the x-ray films from previous mammograms and not just the report for comparison. This helps the radiologist find small changes and detect a cancer as early as possible. This type of multiple image comparison is incorporated into the image viewing tool that this thesis developed for radiologists. Simultaneous and independent zooming into the details of the images in allows better comparisons.

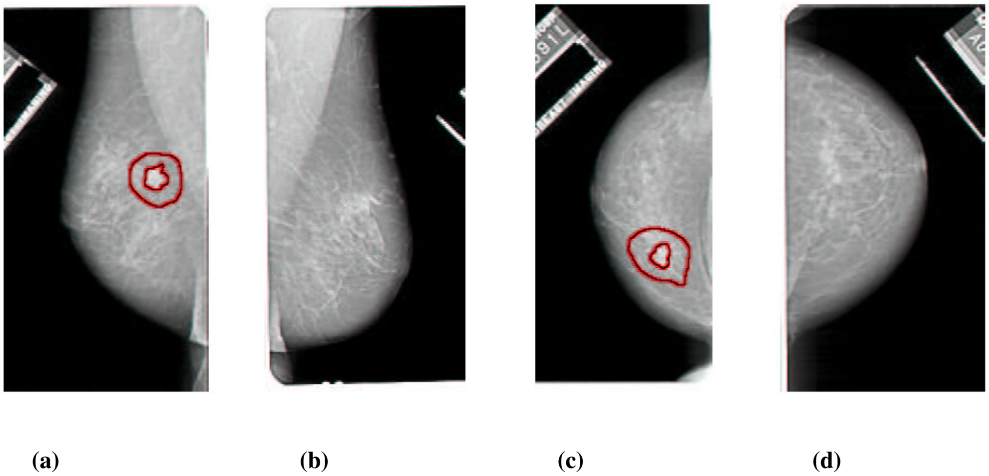


Figure 5: The typical set of four images that make up a mammogram. Comparing this mammograms set with Figures 1 and 4 shows some of the variation in the size and morphology of spiculated lesions. This image set was correctly classified by the method described in Chapter 3.

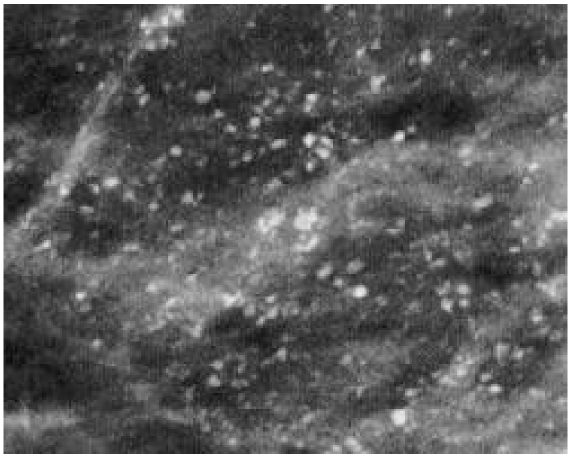


Figure 6: Mammographic images of microcalcifications. Clusters of microcalcifications can be indicative of malignant cancer.

Breast cancer takes years to develop. Early in the disease, most breast cancers have none of the obvious symptoms like lumps. When breast cancer is detected in a localized stage when it has not spread to the lymph nodes, the five year survival rate is 98%. If the cancer has spread to the auxiliary lymph nodes, the rate drops to 80%. If the cancer has metastasized to distant organs such as the lungs, bone marrow, liver, or brain, the five-year survival rate is only 26% [5]. A screening mammogram is an x-ray exam of the breast in a woman who has no symptoms, and example mammograms are shown in Figures 1, 4, and 5. The goal of a screening mammogram is to find cancer when it is still too small to be felt by a woman or her doctor. Finding small breast cancers early by a screening mammogram greatly improves a woman's chance for successful treatment. A screening mammogram usually takes two x-ray images of each breast. For some patients, more pictures may be needed to include as much breast tissue as possible.

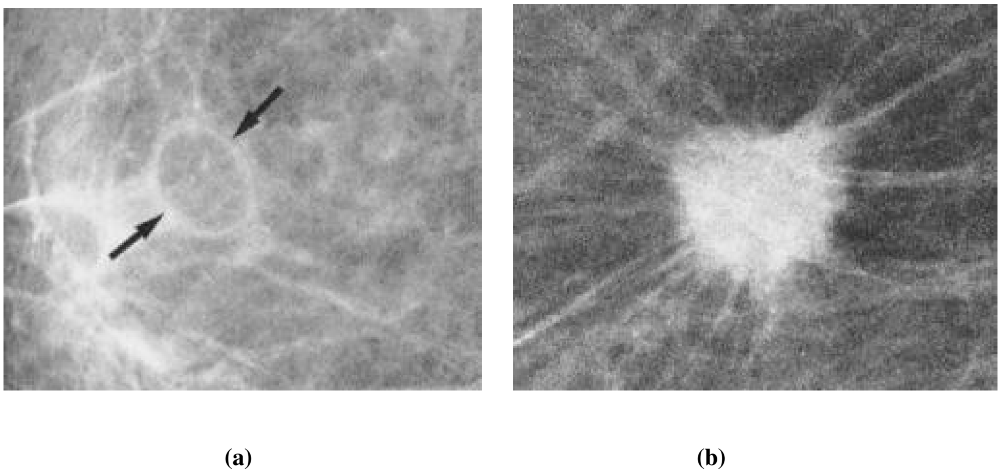


Figure 7: A mammographic image of a circumscribed lesion is in (a). The ring structure is one of the key features that can be picked out of a mammogram. A mammographic image of a spiculated lesion is in (b). The bright center or core is one feature of these lesions, as well as the radiating lines which are called spiculations.

Radiologists look for several types of features, one of which is calcifications. Calcifications are tiny mineral deposits within the breast tissue, which look like small white spots on the films. They may or may not be caused by cancer. There are two classifications of calcifications, macrocalcifications and microcalcifications. Macrocalcifications are larger calcium deposits that are most likely changes in the breasts caused by aging of the breast arteries, old injuries, or inflammation. These deposits are related to noncancerous conditions and do not require a biopsy. Macrocalcifications are more serious, and are found in about half the women over fifty, and in one of ten women under fifty [5]. Microcalcifications are tiny specks of calcium in the breast and are shown in Figure 6. They may appear alone or in clusters, and clusters are more concerning.

Microcalcifications do not always mean that cancer is present. The shape and layout of microcalcifications help the radiologist judge how likely it is that cancer is present.

A mass, which may occur with or without calcifications, is another important feature seen on mammograms. A circumscribed and a spiculated mass are shown in Figure 7. There are many non-cancerous structures in the breast that can obscure masses and have similar textures to cancers. As with calcifications, a mass can be caused by benign breast conditions or by breast cancer. Masses can be caused by many things, including cysts (non-cancerous, fluid-filled sacks) and non-cancerous solid tumors (such as fibroadenomas), but they could be cancer and usually should be biopsied if they are not cysts. A cyst cannot be diagnosed by physical exam alone, nor can it be diagnosed by a mammogram alone. To confirm that a mass is really a cyst, either breast ultrasound or removal of fluid with a needle (aspiration) is needed. If a mass is not a simple cyst (that is, if it is at least partly solid), then more imaging tests may be necessary. Some masses can be observed with periodic mammograms to look for changes in size or shape, while others may need a biopsy. The size, shape, and margins (outline or edges) of the mass help the radiologist to determine whether cancer may be present. Shapes or types of masses include round, lobulated, oval, irregular, architectural distortions, tubular, lymph nodes, asymmetric breast tissue, and focal asymmetric density, and most of these shapes are shown in Appendix A. Margins include spiculated, microlobulated, circumscribed, ill defined, and obscured. Many of these margins have been included in Appendix A. Prior mammograms may help show that a mass has not changed for many years, which would mean that the mass is likely a benign condition and a biopsy would not be needed. Having prior mammograms available to the radiologist is very important for diagnosis.

Breasts vary in density, which affects the appearance of the breast in mammograms. The American College of Radiology (ACR) Breast Imaging Reporting and Data System (BIRADS) characterizes these as ranging from 1-4, with 4 being the most dense. A dense breast presents more non-cancerous structures on a mammogram that can obscure a mass.

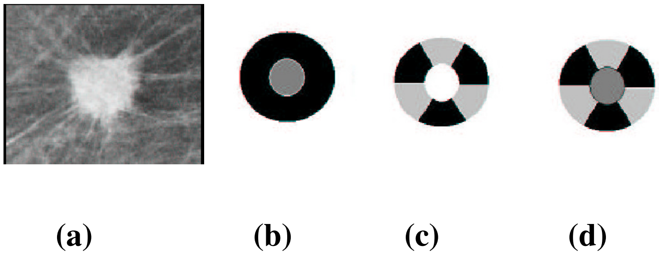


Figure 8: (a) Mammographic image of a spiculated lesion. (b) AFUM filter. (c) Cosine Gabor filter. (d) Combined filter.

The majority of work on feature analysis of mammograms has been through CAD efforts, focusing on determining the contextual similarity to cancer and finding abnormalities in a local area of a single image [53, 95]. The primary methods used range from filters to wavelets to learning methods. In this context, filters are equivalent to shapes that are searched for in an image. Wavelets are the result of applying a transform to the image, and learning methods try to apply prior knowledge to combine a set of low-level image features like pixel intensities into an accurate classification. Problems arise in using filter methods [53] because of the range of sizes and morphologies for breast cancer, as well as the difficulty in differentiating cancerous from non-cancerous structures. The size range problem has been addressed by using multi-scale models [95]. Multiple types of filters must be used to handle the variation in the morphology of various cancers. Similar issues affect wavelet methods, although their use has led to reported good results [70] with the size range issue being improved through the use of a wavelet pyramid [79]. Learning techniques have included support vector machines [25] and neural networks [70].

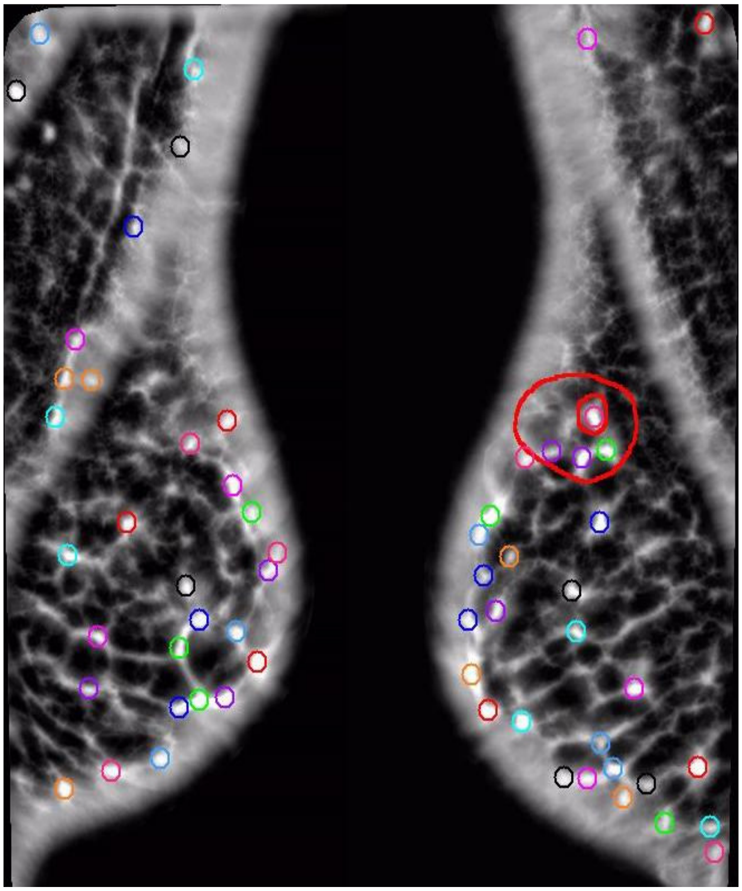


Figure 9. The distribution of the AFUM features on a mammogram. The small circles are the feature positions and strengths, while the larger shape is a hand-drawn annotation by a radiologist of the cancer. Note that the feature does find a cancer, but there are many false positives.

Our analysis starts with CAD prompts to find the contextually similar suspicious points that could be cancers in the mammograms. The CAD technique highlights the areas of the image that have bright cores, a characteristic of spiculated lesions shown in Figure 8a. The filter calculates the percent of the pixels in the outer ring that are less bright than the least bright of the pixels in the inner disk to produce a suspiciousness value, and an example is given in Figure 8b. This suspiciousness value represents the degree to which the surrounding region of a point radially decreases in intensity, and is done over several sizes. This results in focusing on the bright central core of the cancer and ignoring the radiating lines of spiculation. A second filter can be used to detect the radiating lines of spiculation, as shown in Figure 8c, but a combined filter shown in Figure 8d that detects both the cores and the spiculation could improve the performance, especially if the relative weighting of the measurements is learned on an appropriate data set.

The CAD suspiciousness calculation is performed at each pixel location *(x,y)* in the images. The minimum intensity I_{min} within r_1 is found, and then the fraction of pixels between r_1 and r_2 with intensities less than I_{min} is calculated. This yields the fraction under the minimum (FUM) for one set of r_1 and r_2. Keeping $r_1 - r_2 = b$ constant and averaging the FUM over a range of r_1 determines the average fraction under the minimum (AFUM) [45]. The AFUM is then considered to be a suspiciousness value, and represents the extent to which the surrounding region of a point radially decreases in intensity. The CAD prompt output is a set of these suspicious points that are above a certain threshold. Since this is done over a range of sizes, it can respond to cancers of different sizes. This focuses on the bright central core of the cancer and ignores the radiating lines of spiculation. The distribution of these features on a mammogram is shown in Figure 9.

Features with a high suspiciousness value have a higher chance of corresponding to an occurrence of cancer. The centroid of each local maxima in the filtered image is initially marked as a candidate feature site with its suspiciousness value. This collection of sites is then sorted in decreasing order of suspicion. All suspicious sites that are closer than 5mm from a more suspicious site are removed to prevent multiple reporting of the same site. This yields a set of potential feature sites that can be analyzed.

A further improvement might be possible by first transforming the data before filtering, such as applying wavelet analysis to the images before simply thresholding or applying the filter. This has been successfully attempted previously [39] with good results. However, an optimal solution would first combine all of the various filtering and transform methods which create meaningful suspicious points, and then learn an effective analysis from them. This is similar to the effective combination of weak classifiers into a single strong classifier through ensemble learning methods like boosting, which has been successfully used before in tumor classification [30]. Many of the images like mammograms come in pairs, so they form a set that should be very similar. If one of the pair contains cancer and the other does not, then that pair should be different. Thus, the mammogram image set provides both positive and negative examples to build on.

2.3 Proteomic Data

Several additional data sets are on hand for use in this thesis. These include proteomic patterns in serum on cancer types from ovarian to prostate cancer. For the ovarian cancer data, the goal is to identify proteomic patterns in serum that distinguish ovarian cancer from non-cancer. This study is significant to women who have a high risk of ovarian cancer due to family or personal history of cancer. The proteomic spectra were generated by mass spectroscopy and the data set provided here is 6-19-02, which includes

91 controls (Normal) and 162 ovarian cancers. The raw spectral data of each sample contains the relative amplitude of the intensity at each molecular mass / charge (M/Z) identity. There are total 15154 M/Z identities, making the data sets very high-dimensional. The intensity values were normalized according to the formula: NV = (V-Min)/(Max-Min), where NV is the normalized value, V the raw value, Min the minimum intensity and Max the maximum intensity. The normalization is done over all the 253 samples for all 15154 M/Z identities. After the normalization, each intensity value is within the range of 0 to 1.

3 Image Classification

An effort was made to provide image classifications in order to develop a distance function for CBIR of medical images and to explore the properties of the dataset. This initial approach utilizes filtering followed by spatial symmetry analysis using a variant of k-means clustering to determine an overall measure of similarity by combining the contextual similarity of the filtering with the spatial similarity of the analysis. This can be a useful measure for diagnosing mammograms since only an overall determination of cancer or no cancer is required. A secondary goal of our work is to determine the importance of similarity or asymmetry in the computer analysis of mammograms. Figure

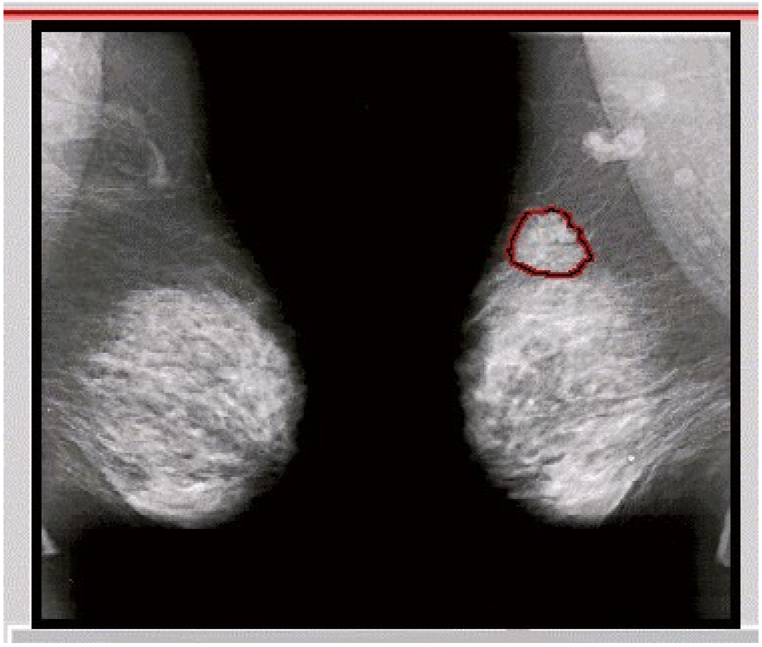


Figure 10: Mammograms of left and right breasts with cancerous area outlined. The similarity of texture between cancerous and normal tissue makes asymmetry an important tool in cancer detection.

10 shows why spatial asymmetry is important in finding cancers in mammograms since we see that the texture and appearance of cancer are both very similar to the texture and appearance of normal tissue in the breast. Our analysis starts with filtering to find the contextually similar suspicious points that could be cancers in the mammograms. The AFUM filter was used, which highlights the areas of the image that have bright cores, a characteristic of spiculated lesions, and is shown in Figure 8b. The filter results are used to rank the output and only the top thirty-two are kept. Although it may not be the optimal choice of filtering, the spatial analysis can be applied to any technique that can rank the suspiciousness of areas. The number of points returned by the filtering step is one of the variables that is learned in optimizing the analysis. Alternatively, a threshold on the suspiciousness value could have been used instead of taking the top few. However, the top few were chosen in order to try to be insensitive to image processing choices. The filter results varied significantly from image to image, which might have biased the analysis if thresholds were used.

3.1 K-Means Clustering

The K-means Algorithm [82] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. We used clustering as a basis for determining image similarity, but there were several changes that had to be made to the technique to adapt it to the application. First, instead of utilizing cluster centers as the main descriptor of the clustering, we used both linear separators in the original feature space as well as hyper-volumes to describe the clusters. Second, we adapted the clustering method to use supervised learning instead of minimizing an objective function. Third, we incorporated the clusters into several distance functions, the parameters of which were learned simultaneously with the cluster definitions to produce an image similarity classification technique.

The k-means procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. The initial position of these centroids causes different results in the final clustering. The next step is to take each point belonging to a given data set and associate it to the nearest centroid, then recalculate k new centroids as barycenters of the clusters. After finding these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. The process iterates and the k centroids change their location step by step until no more changes are done. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The design of the spatial analysis starts with a cluster determination. The procedure follows a simple way to classify a given data set through a fixed number of clusters (assume *k* clusters). This algorithm aims at minimizing an objective function $J = \sum_j^k \sum_i^n \left\| x_j^{(i)} - c_j \right\|$ where $\left\| x_j^{(i)} - c_j \right\|$ is a chosen distance measure (most often Euclidean) between the ith data point $x_j^{(i)}$ associated with the cluster j and the cluster center c_j for cluster j. J is an indicator of the distance of the n data points from their respective cluster centers.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect.

K-means is a simple algorithm that has been adapted to many problem domains. Unfortunately there is no general theoretical solution to find the optimal number of clusters for any given data set. A simple approach is to compare the results of multiple runs with different k classes and choose the best one according to a given criterion, but increasing k results in smaller error function values by definition, as well as an increasing risk of overfitting the data. This algorithm can be adjusted to work as a supervised technique for image similarity comparisons.

3.2 K-Means Variants

Our work utilizes a spatial symmetry analysis to determine an overall measure of similarity. We start with CAD prompts, which are the potentially cancerous sites output from a CAD system. We combine the contextual similarity of the CAD prompts with the spatial similarity of the analysis. This can be a useful measure for classifying mammograms since only an overall determination is required. We believe that many of the techniques described here can also be adapted for use in CAD analysis. A secondary goal of our work is to determine the importance of similarity or asymmetry in the computer analysis of mammograms.

Our analysis starts with CAD prompts to find the contextually similar suspicious points that could be cancers in the mammograms. The CAD technique highlights the areas of the image that have bright cores, a characteristic of spiculated lesions. The filter calculates the percent of the pixels in the outer ring that are less bright than the least bright of the pixels in the inner disk to produce a suspiciousness value, and an example is given in Figure 8. This suspiciousness value represents the degree to which the surrounding region of a point radially decreases in intensity, and is done over several sizes. This is focusing on the bright central core of the cancer and ignoring the radiating lines of spiculation. A second filter can be used to detect the radiating lines of spiculation, as shown in Figure 8, but a combined filter that detects both the cores and the spiculation should improve the performance, especially if the relative weighting of the measurements is learned on an appropriate data set.

For diagnosing breast cancer, the importance of correct classification of the cancerous cases is much more important than the non-cancerous cases. To reflect this, the associated weighting of the cancerous cases was varied, and we evaluate the performance of various weightings.

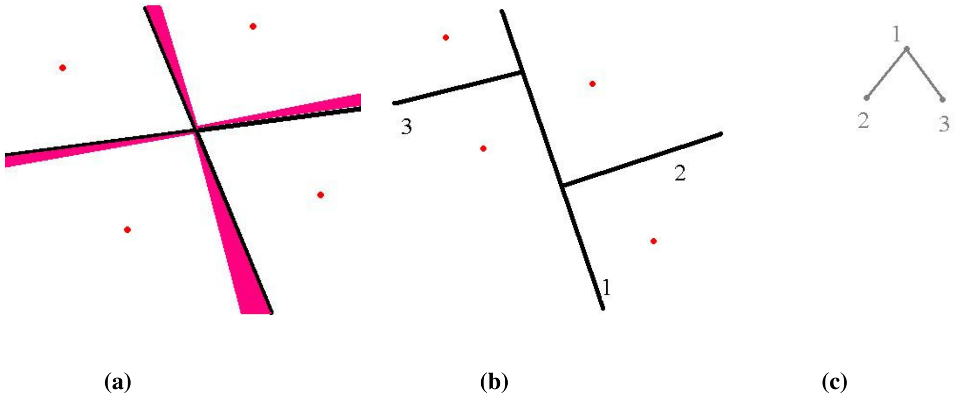


Figure 11: Clustering Techniques. The thick black lines in (a) and (b) are separators, while the small circles are the cluster centers. The shaded region in (a) is the difference in the cluster areas between using the two separators to define the clustering and using the four cluster centers. The separators in (b) are hierarchical, allowing a greater flexibility in the description of the clusters, while (c) shows the hierarchy of the separators.

Several techniques were developed to aid in the development of an improved similarity function for the classification of medical images. We used clustering as a basis for determining image similarity, but there were several changes that had to be made to the technique to adapt it to the application. First, instead of utilizing cluster centers as the main descriptor of the clustering, we used both linear separators in the original feature space as well as hyper-volumes to describe the clusters. Second, we adapted the clustering method to use supervised learning instead of minimizing an objective function. Third, we incorporated the clusters into several distance functions, the parameters of which were learned simultaneously with the cluster definitions to produce an image similarity classification technique.

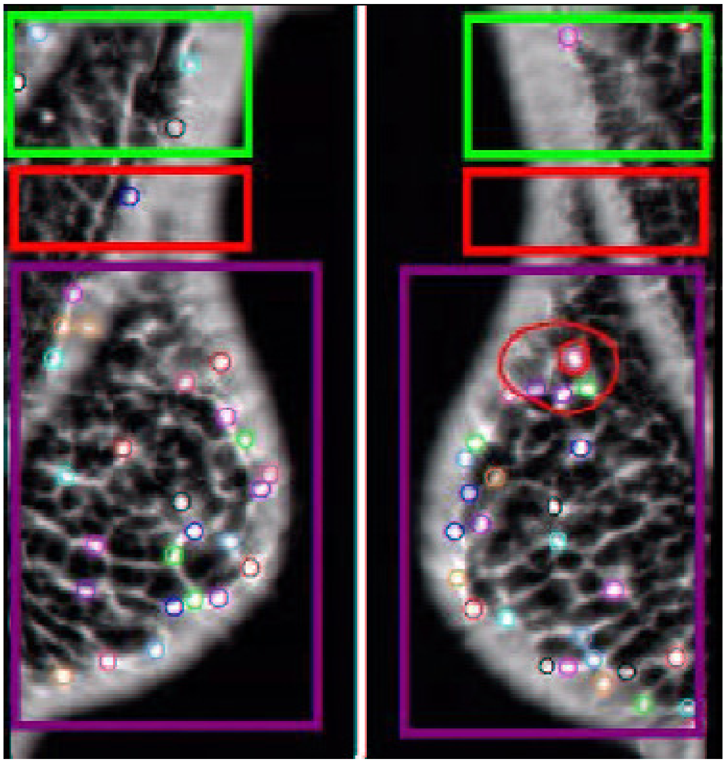


Figure 12: Example Comparison. The features used in this comparison are the small circles. The clusters in this example are the large boxy shapes containing the points. The large hand-drawn circle is the radiologist diagnosis of cancer. This case was correctly diagnosed by both the space-based and data-based techniques.

Our work utilizes a spatial symmetry distance function to determine an overall measure of similarity. The technique requires contextually significant features or probability densities which for this application are CAD prompts, the potentially cancerous sites output from a CAD system. We combine the contextual similarity of the CAD prompts with the spatial similarity of the analysis. The determination of features is discussed in Chapter 3, and an example image set is shown in Figure 1. A variety of similarity methods were explored, the most successful used a variant of clustering to determine an overall measure of similarity by combining the contextual similarity of the features with the spatial similarity of the analysis. We then use this measure for classification since only an overall determination is required.

3.2.1 Separators and Hyper-Volumes

The adaptation of clustering to use separators and hyper-volumes instead of cluster centers was motivated by a desire to minimize the number of parameters required in order to maximize the generalizability of the technique from the training data to the actual test data and thus to real applications. Creating two clusters requires two d-dimensional cluster centers, or 2d parameters like $P = (x_1, y_1, z_1, x_2, y_2, z_2)$, while using a separator plane requires a maximum of d parameters like $P = (a, b, c)$ and can be described in as few as one parameter in special cases like $P = (a)$.

Four clusters can be described using as few as two separator planes as shown in Figure 11a, greatly reducing the number of parameters required to describe the clustering. However, eliminating parameters does change the final clustering so that it does not exactly match the clustering that would have been created with traditional cluster centers. This is the tradeoff between the number of parameters and the flexibility of the technique for breaking up the feature space. The use of overlapping separators minimizes the parameters, but the use of hierarchical separators enables greater flexibility in the definition of the clusters as is shown in Figure 11b.

The use of separators or cluster centers both have the disadvantage of being space-filling so that no part of the feature space can be eliminated from the analysis at the cluster level as well as not allowing overlapping of clusters. However, using hyper-volumes instead of separators does allow both overlapping of clusters and eliminating space at the cost of including additional parameters. A simple hyper-volume is the hyper-sphere which requires d+1 parameters for a cluster center point and a radius. An example of hyper-volumes is shown in Figure 12.

These alternate definitions of clustering focus on increasing the flexibility of the clustering or on decreasing the required number of parameters. An alternate clustering is shown in Figure 13 where the cluster is designed to avoid a noisy area on the images. The focus on decreasing the number of parameters is required for to improve the generalizability of the technique when using supervised learning, which is another adaptation we did to the clustering method that we describe next.

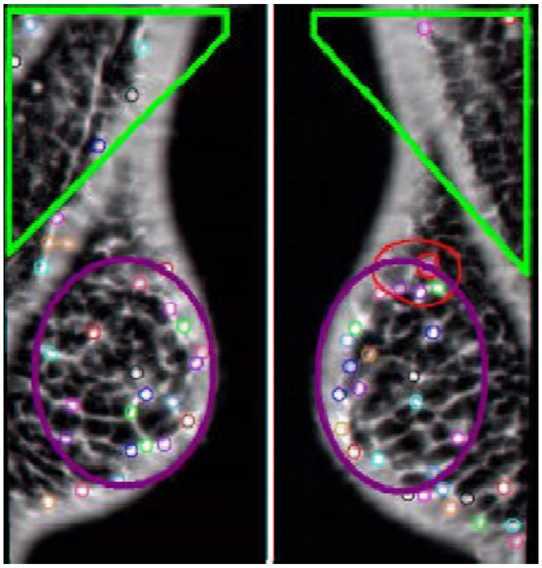


Figure 13. The volumes in this case are non-space-filling and attempt to avoid a noisy area at the chest wall.

3.2.2 Supervised Clustering

The second adaptation of the clustering method was the use of supervised learning to maximize the performance on a training set instead of minimizing an objective function. The error function that we minimize is $E = \sum_j W_j \left\| \tau(g_j, k_j; P) - c_j \right\|$ where W_j is the normalized weight of that particular case, g_j and k_j are the unregistered three-dimensional input features (sorted by one particular feature value for convenience), and $\tau(g_j,k_j;P)$ is the classification function, P are the parameters of the classification, and c_j is the correct classification of the image set j. Note that this technique is being used on image sets, but can be used to compare arbitrary images. The parameters P are learned in order to reduce the error function and includes the parameters of the clustering. Varying the weights of the cancerous and non-cancerous cases allows tuning the performance to achieve fewer false negatives at the expense of higher false positives. The learning was done using exhaustive search in order to guarantee that the result was not caught in a local minimum.

Though the learning was finally done using exhaustive search, we did experiment with hierarchical learning. This is where the first separator is learned, and then the subsequent separators are learned while only changing the parent separator by some fixed percentage and not affecting the grandparent. This is shown in Figure 11b and 11c. We also experimented with true hierarchical learning, where the parent is not allowed to vary, but this was found to be ineffective. This has the effect of reducing the number of degrees of freedom to learn by breaking the learning up into multiple levels. The learning of one level is reduced to learning the two child separators and the minimized range of the parent separator, instead of learning the entire set of separators. The inclusion of more separators is self-limiting if the separators are allowed to line up with the parent, thus not breaking up the space and indicating that the hierarchy should end at the parent for that volume of feature space. The application did not require a large number of levels in the hierarchy, allowing the use of exhaustive search to verify the results of the hierarchical learning.

3.2.3 Image Comparison Distance Functions Using Clusters

The analysis for image comparison that we used performs a comparison of clusters of features in order to maintain both a contextual and spatial comparison while avoiding an exact registration. We experimented with two different models where the clusters are defined using separators and hyper-volumes. We also experimented with a model that compares small clusters of features between images. The hyper-volume image comparison can be seen in Figure 12, where the points are assigned to clusters that are defined by large volumes of feature space and have a set spatial relationship between each other. The feature-space hyper-volumes have a pre-set registration with the corresponding volumes in the other image. For simplicity, the volumes are assumed to be non-overlapping and space-filling, but this is not required. Additionally, the volumes are assumed to contain the same hyper-volumes in the images of the left and right breasts out of symmetry. This reduces the number of parameters and increases the ability of the model to be generalized to a larger data set, based on the assumption that there are no important anatomical differences between the left and right breasts and that breast cancer is equally as likely to be in the left or right breast.

In a hyper-volume image comparison, a hyper-volume is assigned all of the suspicious points in the space *dA* that the hyper-volume spans. The parameters of the hyper-volumes are learned through parametric learning, and any model can be used to characterize the hyper-volumes in feature space. Exact registration of the suspicious points is avoided by using the volumes for the comparisons as they are registered with the corresponding volume in the other image.

The feature space is broken up into volumes *dA* as shown in Figure 12. The agglomerated distance *D* shown in Equation 1 is defined for the comparison of the two point feature sets, and the absolute value of the differences compared against an optimized threshold. Since the features are point features, they are represented using the delta function δ and there is no weighting function.

1)
$$D = \sum_{dA} \left| \left(\iiint_{dA} df \sum_{i} \delta(f - \bar{a}_i) - \delta(f - \bar{b}_i) \right) \right|$$

The point sets for the images are represented as $\{a_i\}$ and $\{b_i\}$ for images *a* and *b* respectively. The summation over *dA* is done over all of the clusters which are represented by their hyper-volume *dA*. The integration is done over the actual hyper-volume *dA* of the cluster. The summation over *i* is done over all of the features. The multi-dimensional integral over feature space provides the agglomeration aspect of the distance metric. The hyper-volumes *dA* provide the agglomeration and are learned along with a threshold in order to optimize the performance of the distance measure at classification. This allows the distance metric to be easily adapted to different image types and imaging techniques, as well as providing a method for incorporating feedback into the distance metric. This distance metric compares the distributions of spatially distributed point sets, and is sensitive to variations in the distribution for image comparison. This is useful for applications such as determining the presence of cancer. There are several other variations to this distance metric that have been explored.

A variation on this distance metric is shown in Equation 2 that learns a threshold for each cluster *dA*, which has the advantage of being able to emphasize the importance of some areas in the feature space over others. This can be used to distinguish noisy areas where many spurious suspicious points are found from important areas where even small variations are indicative of a lack of similarity. This technique of learning important areas in images can be thought of as an image discovery technique.

$$2) \qquad D_{dA} = \iiint\limits_{dA} df \sum_{i} \delta(f - \overline{a}_i) - \delta(f - \overline{b}_i)$$

A more generalized form of the similarity distance metric is given by equation 3, where the delta function is not the required function and the number of features in each image is not required to be the same. A natural choice for the function *g* is the probability density function; however, the function *g* can be determined to try to optimize the retrieval on the particular application.

3) $$D = \sum_{dA} |\left(\iiint_{dA} df \sum_{i} g(f - \overline{a}_i) - \sum_{j} g(f - \overline{b}_j)\right)|$$

We tested several variations on the image comparison ideas. The simplest model utilizes only four parameters: three parameters for a separator and one parameter for a threshold and used Equation 1. For diagnosing breast cancer, the importance of correct classification of the cancerous cases is much more important than the non-cancerous cases. To reflect this, the associated weighting of the cancerous cases was varied, and we evaluated the performance of various weightings. The second model used the same four parameters and Equation 1, but weighted the learning to give greater weight to the performance on the cancerous cases over the performance on the non-cancerous cases. The third model used the parameters of the first, but also included an additional parameter that permits selection, so that cases that do not have a minimum number of features in each cluster are not analyzed. The fourth model used seven parameters: six for two separators and one for a threshold, used Equation 1, and is shown in Figure 12. The fifth model used seven parameters as well: six for two separators and one threshold, but used Equation 2 with the same threshold for each cluster comparison. These models were motivated by the observation that the cancer would change the distribution of the suspicious points, leading to an indication of cancer. An improvement to the method would be to adaptively determine the optimal number of volumes through a split-and-merge type methodology [61].

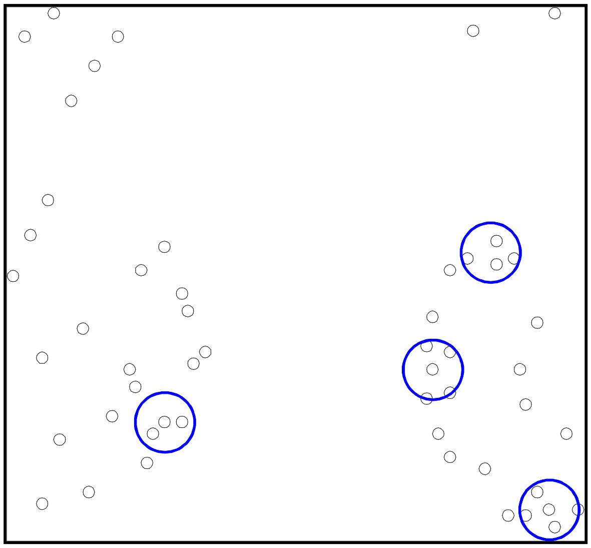


Figure 14: Small Cluster Image Comparison. The suspicious points are the small circles, with the points on the left coming from the image of the left breast and the points on the right coming from the image of the right breast. The volumes are the larger circles. This method searches for small clumps of suspicious points and then assigns a volume there, comparing the number of volumes in the two images.

The third model that we tried does not set the number of clusters arbitrarily, but instead learns the number of clusters from the data and learns the best parameterization of the clusters. These image comparisons search for small clumps of suspicious points and then assign a cluster there, as shown in Figure 14. The maximum distance between feature points and the minimum features needed to define a volume are learned on a training set. The clusters were also defined to be centered on a suspicious point because we believed that small clumps of suspicious points tended to form around the central cancer. This assumption may be incorrect, and freeing the cluster centers from that constraint may improve the performance. Exact registration is avoided again by registering the clusters instead of the image or the suspicious points. Comparing the number of clusters in the right image versus the number of clusters in the left image provides a first cut at registering the clusters since a difference in the numbers of clusters implies that some volumes cannot be registered. Improving the cluster registration may improve the performance of the method. This image comparison was motivated by the data, where we observed a small volume of suspicious points at a cancer sites.

Many approaches were attempted on this dataset. One unsuccessful approach compared the variances of the distribution of suspicious points, while another used a Naive Bayes analysis, and these are compared along with wavelet methods and commercial techniques.

3.3 K-Means Variants Evaluation

The image comparisons were applied to the mediolateral oblique (MLO) mammogram views of both the left and right breast of patients that were diagnosed with cancer and patients that were diagnosed as normal, or free from cancer. The analysis was performed over test and training data sets, with cases that were roughly split between normal mammograms and mammograms with malignant spiculated lesions from the Digital Database for Screening Mammography [54]. The focus was on one type of breast cancer which creates spiculated lesions in the breasts. Spiculated lesions are defined as breast cancers with central areas that are usually irregular and with ill-defined borders.

Their sizes vary from a few millimeters to several centimeters in diameter and they are very difficult cancers to detect [79].

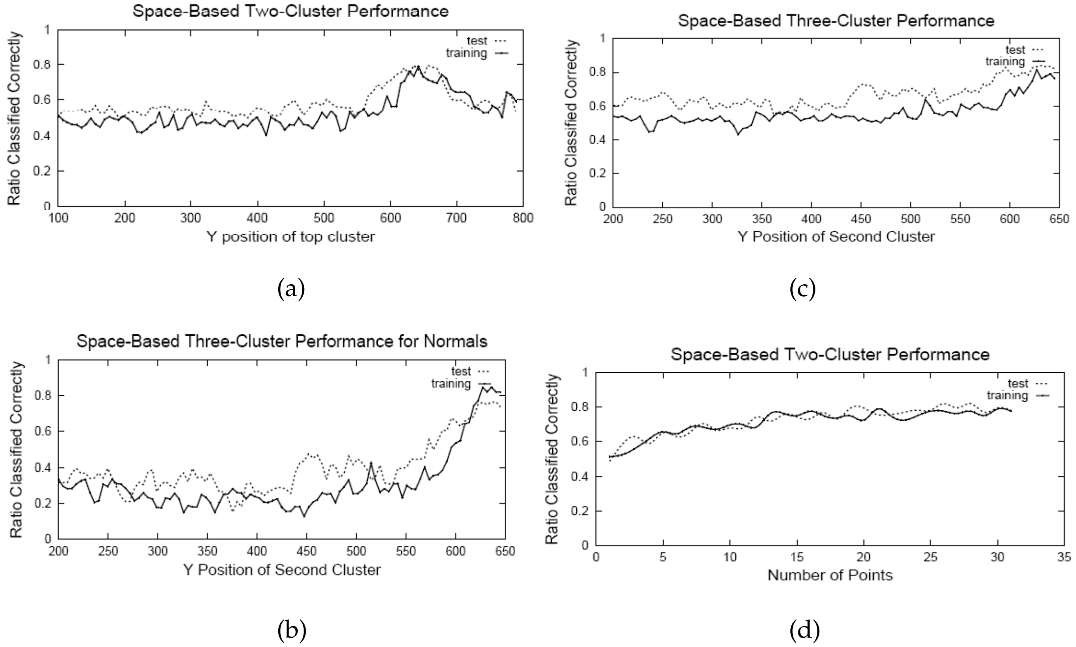


Figure 15: Comparison Data. The maxima in learning the two-cluster method with respect to one of the parameters, the y value of the second cluster are in (a) and the method is shown to generalize well from training to test data. The same information for the three-cluster method is shown in (b). The performance relative to the number of suspicious points used in the two-cluster technique is in (c). The performance of the three-cluster method on normals, or non-cancerous cases, is shown in (d).

The training set had 39 non-cancerous cases and 37 cancerous cases, while the test set had 38 non-cancerous cases and 40 cancerous cases. The data is roughly spread across the density of the breasts and the subtlety of the cancer. The breast density and subtlety were specified by an expert radiologist. The subtlety of the cancer shows how difficult it is to determine that there is cancer. The training data set was used to determine optimal parameters the volumes dA. The inputs are the extracted CAD features for each image in the screening mammogram set, as shown in Figure 8. The output is a classification as either cancerous or non-cancerous. We used exhaustive search because we could, and require only a single stage. These cases indicated that a difference in the clusters of one or more suspicious points indicated cancer in both the two and three cluster experiments.

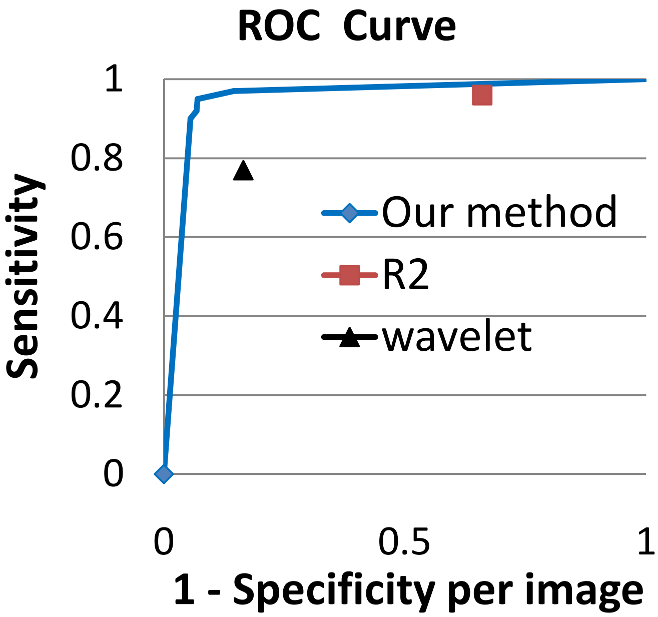


Figure 16. ROC curve demonstrating the effectiveness of this distance metric at diagnosing mammograms.

The most successful approach that we have constructed so far defined the cluster volumes dA with the parameter set P = (x1 || y1 || z1, x2 || y2 || z2, t, n) where the number of features used in the analysis is n and the threshold for the distance function is t. The first parameter x1 || y1 || z1 chooses the best dimension and best position to break up the feature space into volumes, as does the second parameter x2 || y2 || z2. This analysis appears in Figure 12 and is called the "three-cluster" approach. This used an equal weighting on the error function. Another successful approach used the parameter set P = (x1 || y1 || z1, t, n) but was heavily weighted towards correctly classifying the cancerous cases, and this will be called the "two-cluster weighted towards cancer" analysis. Yet another successful approach attempted to use automatic selection to classify only the cases that would be would be analyzed well. This approach used the parameter set P = (x1 || y1 || z1, t, n, s) where s is a required minimum occupancy of each cluster. This approach is called the "two-cluster with selection" approach. One unsuccessful approach compared the variances of the distribution of suspicious points, while another used a Naïve Bayes analysis.

3.4 K-Means Variant Results

Our results are good on all cases of the test set, correctly classifying 80% for the two-cluster as shown in Figure 15a, and 85% of the time for the three-cluster as shown in Figure 15b. The data-defined cluster model results as shown in Figure 15c were not as good, but have the potential for improvement. The results are summarized in Table 1. However, it is much more important to correctly classify the cancerous cases, and by heavily weighting the importance of the cancerous cases, we correctly classified 97% of the cancerous cases with the two-cluster model.

Neither the subtlety nor the density of the cancer had an effect on the results. The data sets density and subtlety are shown in Appendix B for both the training set and test sets. It would be possible to create a data set that is perfectly balanced in both density and subtlety if an infinite amount of analyzed images with the particular cancer were available. However, because we have only a limited number of cancerous images, there is some possibility that the imperfect distribution could affect the results of the analysis.

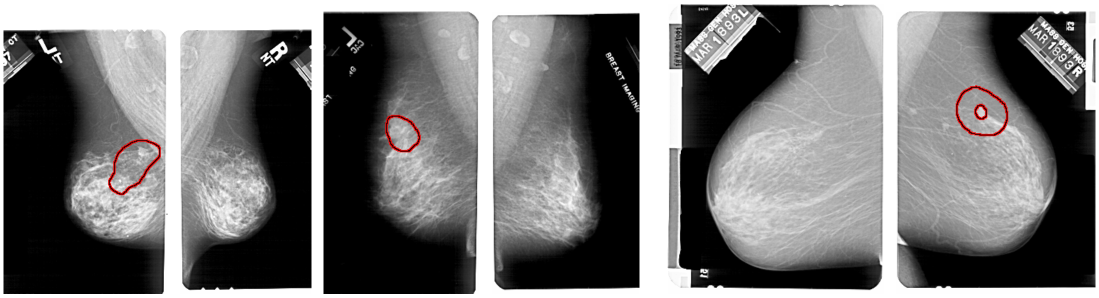
The comparison with a commercial system shows that the results are surprisingly good. Our method showed an improvement of 26% on the non-cancerous cases while matching the performance on cancerous cases with the R2 ImageChecker system [11]. The inclusion of additional factors other than asymmetry in the method should improve the results. However, the data sets used are different, as the R2 ImageChecker data contains all cancer types and our method has only the difficult to detect spiculated lesions. The R2 ImageChecker data set also had a much higher proportion of non-cancerous mammograms to cancerous cases. Our performance is shown in Figure 16.

One of the parameters that was learned was the optimal number of suspicious points to use in the analysis, and the results were always at or near the top of the range that we used, varying from 29 to 32 points depending on the model and weightings as shown in Figure 15d. This was surprising because the cancer was usually in the top sixteen if not the top eight points. However, the suspicious points do tend to cluster around a cancer, so including more suspicious points may create a greater distortion of the underlying distribution than fewer points. The learning algorithm does not get the number of points directly, only the cluster differences, so the inclusion of more points should not skew this analysis.

An interesting result from the three-cluster analysis showed that these methods could discover areas in images that are important for the classification, and this is demonstrated in Figures 15b and 15c. The analysis found a region of interest for diagnosing a mammogram as non-cancerous. These techniques can be used as a method for probing feature space for important areas.

Table 1. Results Table. The accuracy of the techniques. The * indicates different but similar data sets.

Method	**Cancerous**	**Non-Cancerous**
Three-Cluster Equation 1	90%	79%
Two-Cluster Weighted Toward Cancer Equation 1	97%	42%
Two Cluster Equation 1	87%	71%
R2 Image Checker * [9]	96%	33%
Wavelet * [13]	77%	77%
Naïve Bayes	51%	49%
Three-Cluster Equation 2	95%	73%
Variance Analysis	60%	60%
Two-Cluster With Selection Equation 1	92%	73%
Small-Cluster Analysis	51%	56%



(a) **(b)** **(c)**

Figure 17: The left and right MLO views of three cases that were misdiagnosed. The cancerous areas are outlined in red. There are significant variations in the size and morphology of spiculated lesions. Note that cases (b) and (c) both have significant differences in the size and shape of the breasts from left to right.

Our methods make use of a spatial analysis of the suspicious points, and its success is an encouraging sign for the investigation and utilization of more complicated non-local analysis techniques in medical imaging and analysis.

Analysis of the misdiagnosed cases in Figure 17 demonstrates a potential flaw in the method. When there is too much structure in one area that draws the relatively simple features that we are using into it on just a small number of cases, the method can misclassify them. A potential improvement is to incorporate a second level of classifiers that would analyze the missed diagnoses.

3.4 K-Means Variants Conclusions

Our results are strong on all cases of the test set, correctly classifying with 85% accuracy and our technique outperforms both the best academic and commercial approaches, suggesting that this is an important technique in the classification of mammograms. We have also shown that using the image comparisons to determine the classification is insensitive to the parameters of the volumes.

We created and compared multiple models, demonstrating that three area volumes worked slightly better than two, and showed that the data-defined method was not as effective. However, the data-defined method was sensitive to the presence of false positives near the breast boundary, and removing volumes at the breast boundary could improve the effectiveness of this approach. We also defined a new distance measure for the comparison of point sets and demonstrate its effectiveness in this application. The coupling of this distance measure with the parametric learning of clusters led to a highly effective classification technique.

The clusters also discovered an area of interest in mammogram comparisons which improved the diagnosis of mammograms that did not have cancer. More clusters might improve the technique, or, more importantly, they might lead to the discovery of more areas of interest. We suggest several ways that might improve on the methods that we used to compare mammograms. One method is to convert a mammogram into a connected graph structure of suspicious points and to utilize known graph comparison methods for the measure.

4 Image Database

Building effective content-based image retrieval (CBIR) systems involves the combination of image creation, storage, security, transmission, analysis, evaluation feature extraction, and feature combination in order to store and retrieve medical images effectively. This requires the involvement of a large community of experts across several fields. We have created a CBIR system called Archimedes which integrates the community together without requiring disclosure of sensitive details. Archimedes' system design enables researchers to upload their feature sets and quickly compare the effectiveness of their methods against other stored feature sets. Additionally, research into the techniques used by radiologists is possible in Archimedes through double-blind radiologist comparisons based on their annotations and feature markups. This research archive contains the essential technologies of secure transmission and storage, textual and feature searches, spatial searches, annotation searching, filtering of result sets, feature creation, and bulk loading of features, while creating a repository and testbed for the community. In the medical imaging field, CBIR techniques and clinical decision support techniques are called case-based reasoning [73] or evidence-based medicine [21, 24]. The CBIR process involves feature extraction from the images, which is described in Chapter 2, the processing of the features through a similarity function, which is described in Chapter 3, and the retrieval of images from the database, as well as feedback.

The number of digital medical images is rapidly rising, prompting the need for improved storage and retrieval systems. Image archives and imaging systems are an important economic and clinical factor in the hospital environment [109]. The management and the indexing of these large image repositories is becoming increasingly complex. Most retrievals in these systems are based on the patient identification information or image modality [74] as it is defined in the DICOM standard [92], but it is hoped that inclusion of other features can improve the effectiveness of this type of system. Archimedes includes retrieval based on features and combinations of features, as well as on patient identification information, doctor's notations, and image modality in order to develop effective CBIR. Archimedes also includes filtering of the result set in order to further refine and improve the search.

Clinical decision support techniques such as case-based reasoning [73] or evidence-based medicine [21, 24] rely on effective CBIR development. Image and visual feature-based searches will help find similar images, but textual searches are always going to be an important part of any medical CBIR system, especially through searches on patient information or characteristics. That is why searching on patient information and other text is already supported in the Archimedes system.

The integration of CBIR methods into Picture Archiving and Communication Systems (PACS) has been proposed several times. PACS are the main software components used to store and access the large amount of visual data in medical departments. Often, several layer architectures exist for quick short-term access and slow long-term storage [75], but this is becoming increasingly unnecessary as technologies have improved. The Archimedes system was designed as a web-based system for both the development and evaluation of CBIR, and provides a platform to evaluate the usefulness and effectiveness of incorporating CBIR changes into PACS.

Several frameworks for distributed image management solutions have been developed such as I2Cnet [87, 88]. Image retrieval based on visual features is often proposed but unfortunately little is said about the visual features used or the performance obtained. A real medical application of CBIR methods and the integration of these tools into medical practice has required a large group in very close cooperation for a long period of time. CBIR systems that have followed this model are the Assert system for the classification of high resolution CTs of the lung [4, 102] and the IRMA system for the classification of images into anatomical areas, modalities and viewpoints [71]. The Archimedes system bypasses this difficulty with a web-based community of researchers who can contribute features, images, results sets, diagnoses, and other expertise in an open research environment. This thesis demonstrates a technology to decentralize this process by including a large web-based collaboration of partners, each achieving individual goals while contributing to the overall goal of an improved CBIR system.

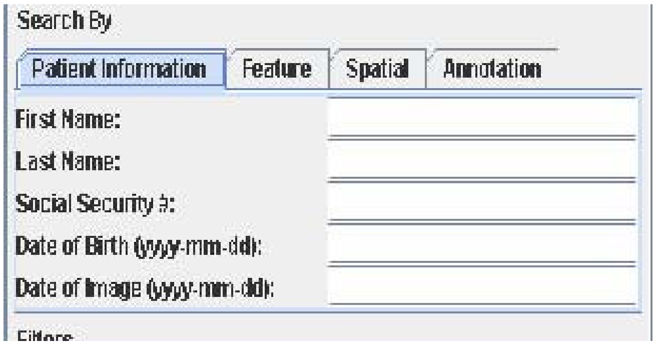
Comparing CBIR systems is often challenging because commercial companies are often unable or unwilling to share their techniques. Archimedes allows commercial companies to contribute their features or results sets without disclosing their techniques, enabling unfettered communication. The system also enables the rapid creation, storage, and download of specialized data sets for comparisons. One example of an interesting data set that Archimedes can create and store would be mammograms of high density breasts for which MRI images are also available. Comparison of CBIR results is simplified by the storage of multiple results sets for images, and the ability to quantify the results sets.

There are several different types of research that go into developing an effective CBIR system for medical images. The primary research is done by radiologists, who perform the medical scans as well as provide diagnoses. Archimedes can help radiologists by organizing their images, capturing their patient notes and digitizing their image annotations, speed up the analysis of experiments, and enable quick comparisons of different radiologist techniques such as comparing double-reading to single reading of mammograms. Web publication of research into radiologist techniques can be simplified using Archimedes' built in annonimizing and web publishing. The next stage of research is feature extraction and analysis. An example of this is the measurement of spiculation [100] as a feature used to aid in the detection of spiculated lesions. Archimedes can help with research into feature extraction by providing annotated images, comparison features, and comparison results sets, as well as data analysis and feature combination. Once the images have features and diagnoses associated with them, all of the pieces are available for research into CBIR techniques.

4.1 Data Collection and Analysis

Archimedes is designed to be able to run and analyze double-blind studies of radiologist techniques. Archimedes can be configured so that multiple radiologists can annotate the same image under different conditions without viewing the biopsy-based ground truth or the other annotations. Their input can then be viewed by a user which can only view the doctor information after it has been automatically annonimized to maintain the integrity of the study. The data can be analyzed in several ways using the capabilities of Archimedes. A particular data set can be isolated using the dataset name in the image annotations. This data set can be filtered to show only particular types of cases, such as malignant cancers, normals, or benign cancers. A particular (annonimized) doctor's diagnosis can further refine the results, showing the percentage of accurate diagnoses on different classes of images. The position of the actual cancer can be compared to the biopsied "truth" position using a spatial search, finding all of the cases where the doctor's diagnosis is within a specified distance of the "truth" position. The analysis can then be finished in a matter of hours. The results can then be stored in Archimedes, and the study can be published immediately by allowing Archimedes to enable guest user access.

Archimedes was originally designed to do double-blind research studies, but it also has an image analysis and patient records management tool that can be used by the medical research community. Radiologists can store and organize medical images such as x-rays, mammograms, CAT scans, MRIs, and any other image that is stored in a DICOM format. Radiologists can rapidly retrieve images and patient records, and can also find patients with similar images, conditions, or annotations to compare treatment successes. The software archives the addition of markups and notations to images, as well as associating text and patient info with images.



Search By

Patient Information | Feature | Spatial | Annotation

First Name:

Last Name:

Social Security #:

Date of Birth (yyyy-mm-dd):

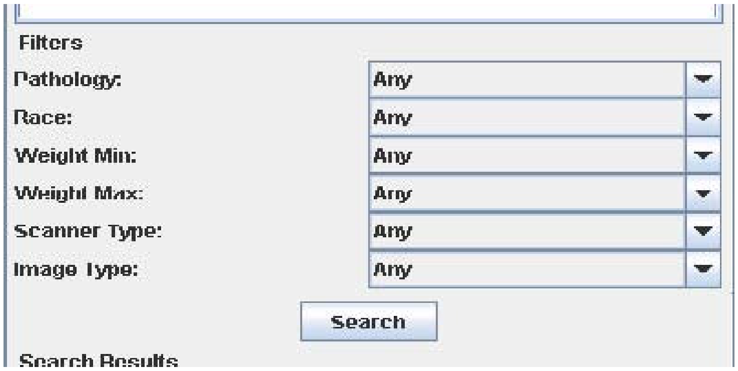
Date of Image (yyyy-mm-dd):

Filters

Figure 18. Archimedes Patient Information Search Panel. Searches over

patient information can be done using the patient's name, date of birth, social security number, or the date an image was taken. Additional search tabs are available as well.

For managing patient records and images, the primary tool is searching by patient information and text. For search by text in Archimedes, the medical professional is allowed to enter patient information (i.e. first name, last name, date of birth, etc.) into Archimedes. Once information is entered, they can use filters to further refine their search results. The searching options for patient information are shown in Figure 18, while the filtering options are shown in Figure 19, and the updated version is in Figure 20. Searching by feature allows the doctors to specify feature parameters that they wish to see in the results. This enables medical professionals to quickly find similar cases. The next type of search is an extension to searching by feature, specifying multiple features with defined spatial relationships between them. The most useful of these types of search would specify a distance between features, for example to find areas that have the features of both spiculations and bright central cores indicative of spiculated lesions in mammograms. Archimedes also allows search over comments other doctors previously made about patients or images. This works like a primitive Yahoo search over the text of the medical annotations.



Pathology: Any

Race: Any

Weight Min: Any

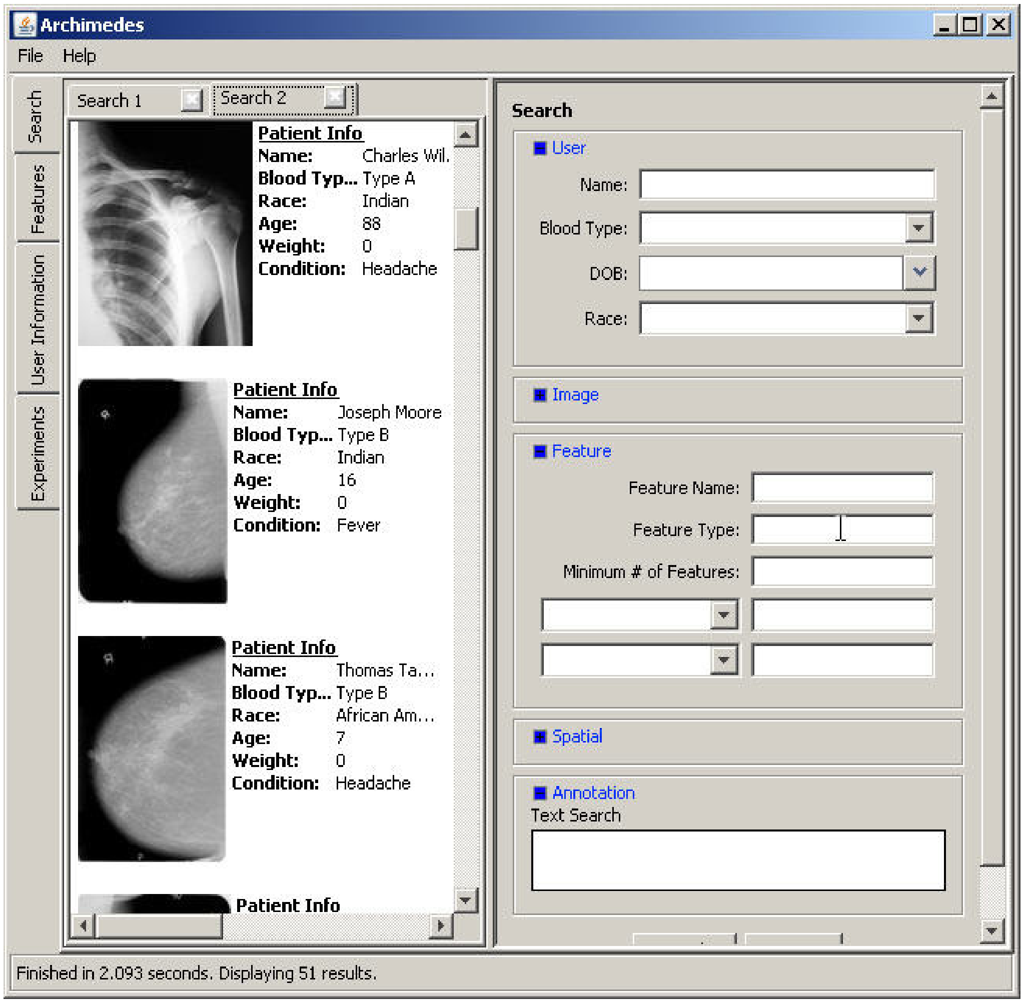
Weight Max: Any

Scanner Type: Any

Image Type: Any

Figure 19. Archimedes Filter Panel. Results from searches can be filtered based on pathology, image and scanner type, weight range, and race. This was useful in analyzing performance data across multiple parameters.

Though the basic searching on patient information and annotations is included with minimal effort, the more advanced feature and spatial searching requires extra input. There are two options: capturing doctor input or getting permission to make images available for researchers to mark up the images. For example, getting the features like spiculation into the medical database may require making those images available to the researchers who specialize in measuring spiculation. Making the images available to researchers involves setting the permissions for a individual image or groups of images to public or semiprivate, where images are available to the public at large or to a set group. The signed consent forms can also be stored in Archimedes as images. Capturing the doctor input required a viewing and input capture tool.



Patient Info
Name: Charles Wil.
Blood Typ... Type A
Race: Indian
Age: 88
Weight: 0
Condition: Headache

Patient Info
Name: Joseph Moore
Blood Typ... Type B
Race: Indian
Age: 16
Weight: 0
Condition: Fever

Patient Info
Name: Thomas Ta...
Blood Typ... Type B
Race: African Am...
Age: 7
Weight: 0
Condition: Headache

Patient Info

Finished in 2.093 seconds. Displaying 51 results.

Figure 20. Updated Archimedes Search Panel. The search results now are returned with thumbnail images and info. The search panels can be closed to maximize space.

Medical professionals can view and manipulate images on the Archimedes system and tab through set of images. There is a zooming interface in order to focus in on interesting parts of the images. Point features can be inserted and described as overlays to the images. Text annotations can be entered. Multiple overlays can be captured for each image, allowing double reading of images. This type of detailed image information is essential for the design and evaluation of CBIR systems as well as computer-aided detection and diagnosis systems. The radiologists' diagnoses are captured and the data can be accessed remotely, thereby allowing tele-medicine applications to be run on the Archimedes platform.

The Archimedes system helps store, annotate, and retrieve data and images for the radiologists, who are the primary data collection agents in the development of a medical image CBIR system. Improving their work and giving them incentives to share their images and diagnoses is one of the key steps in creating a collaborative environment to create CBIR.

4.2 Feature Capabilities

One of the main challenges in feature extraction is finding a large enough set of images of the same exact type of cancer in order to focus in on its particular characteristics. But there are a few large databases that do provide these images, for example with lung cancer images [28]. However, in order compare the effectiveness of one feature versus another on the same images, the comparison research has to be replicated. Archimedes eliminates this problem by allowing researchers to store their features in Archimedes for comparison along with the images from which they were extracted. The input, storage, and sharing of features is one of the design choices that make Archimedes unique. The open sharing of features makes comparisons possible without the need for inaccurate replication of older work, as well as enables research into feature combination both faster and more effective. Features can be combined using spatial search and then fused into a new feature type. For example, a spiculation feature can be combined with a bright central core feature for detecting spiculated lesions. The spatial search can also be used to quantify the effectiveness of the feature at predicting the position of the cancer by comparing the feature position with the biopsied "truth" cancer position.

Evaluating the effectiveness of features can be done using Archimedes extensive query capability. Finding cases when a "test" feature is near a "truth" feature can be done using the spatial query capability, as well as those "truth" features that are not near a "test" feature and vice versa. The spatial query is tunable with a variable input distance for greater flexibility. The image categories can be adjusted with the filters to isolate cases that are malignant, or other medically relevant characteristics.

Archimedes can also store and share features that are not associated with a particular position. The flexible design of the feature storage and upload make it capable of handling most features. A planned improvement is the handling of feature areas and feature volumes. Currently point features and area features are handled, but the spatial comparisons are not yet finished.

Features can be manually input through the Archimedes zooming interface, or loaded in bulk though an XML schema, with the following small example:

```
<Patient>
<Image>
<Doctor>
<Feature>
</Feature>
<Feature>
```

```
</Feature>
<Annotation>
</Annotation>
</Doctor>
</Image>
</Patient>
```

Archimedes generates a skeleton schema for a user's selected group of images in order to facilitate upload and match the annonimized patient ID with the correct image. Features can contain pixel positions using <Xpos> and <Ypos>, but it is not required. Features can contain a number of values associated with them, and these values are uploaded with <ValueXName>, <ValueXType>, and <ValueX> for the Xth value. These values can be used in limiting feature searches.

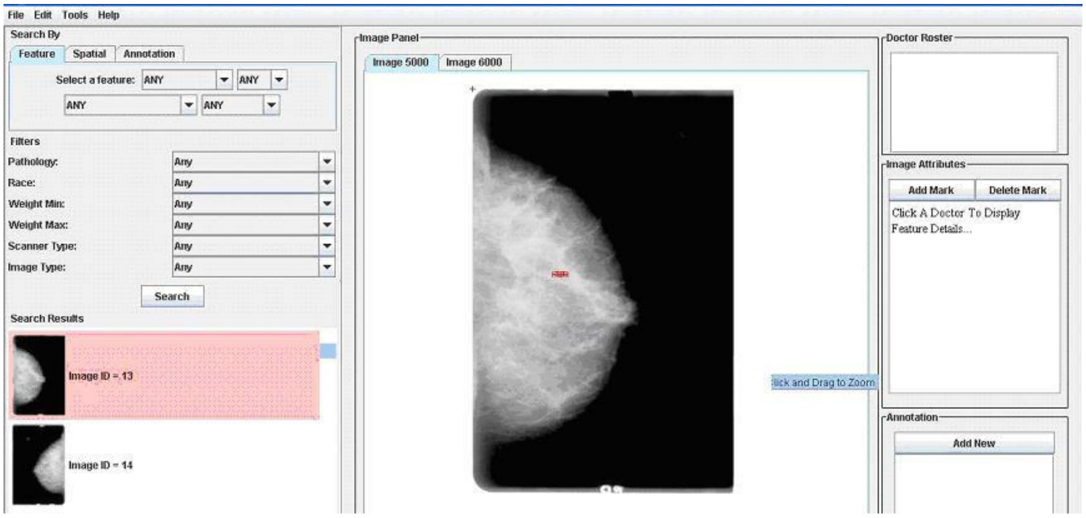
Techniques that combine features could also be stored in Archimedes, and stored as features as well. Currently, the spatial search can be used to combine features, but more complex approaches have to be done offline and uploaded as features. One planned upgrade is a learning package built into Archimedes that would simplify the development of classifications and analysis of medical images. This would allow radiologists to use Archimedes to explore relationships and use the learning package to optimize the approach.

The combination of advanced querying and feature sharing enables rapid analysis of features and combinations of features for CBIR and the comparison of computed features to "truth" features defined by a radiologist. By providing a platform for the analysis and comparison of features, Archimedes encourages the collaboration between researchers designing features as well as researchers building CBIR.

There are different types of features for different types of artifacts to be associates with images. The application has the ability to support user-added features of properly defined types. By default, Archimedes allows users to mark images with:

1. A single point, usually placed in the center of the artifact.
2. A set of one or more lines used to define a polygonal shape, used to outline or surround part of or the whole artifact.
3. A freehand markup tool that allows custom shapes to be hand drawn onto the image.
4. A rectangle tool that allows a rectangle or square to be drawn on the image.
5. An oval tool that allows an oval or circle to be drawn on the image.
6. An angle tool that allows an angle of any size to be drawn on the image, as the product of two lines meeting at one point.

These capabilities allow the functionality of Archimedes to capture the diagnosis of radiologists through whichever drawing method is preferred by them. Making the database flexible and intuitive will encourage radiologists to make use of the capabilities, and thus capture their diagnosis for reference and for research purposes.



Select a feature: ANY ANY
ANY ANY

Pathology: Any
Race: Any
Weight Min: Any
Weight Max: Any
Scanner Type: Any
Image Type: Any

Image ID = 13
Image ID = 14

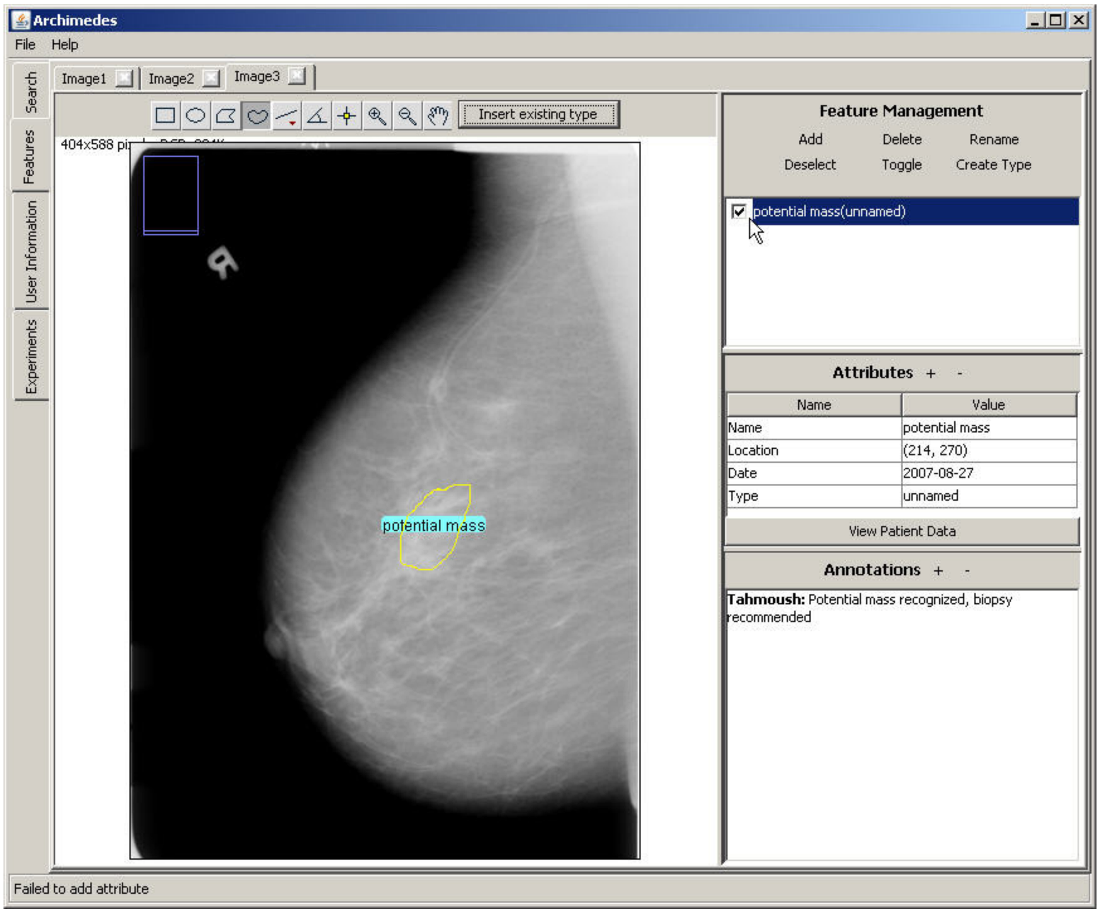
Click A Doctor To Display Feature Details...

Figure 21. The initial Archimedes user interface. The search panel and results set are on the left, the zooming interface is in the middle, and adding features and annotations are on the right. This is the semi-private view where no patient information is viewable or searchable.

4.3 CBIR

The images, categorizations, and diagnoses provided by radiologists combined with features enable the exploration of CBIR in medical images through Archimedes. Typical CBIR approaches combine features into a feature vector and use a variety of techniques to determine the most accurate similarity measure. The categorizations of images, like evaluations of breast density or cancer type or malignancy, can be used to evaluate and to verify the effectiveness of CBIR at returning similar images.

As with features, it can be difficult to compare CBIR techniques without recreating the research of others. Archimedes allows the storage of result sets to simplify the comparison of different CBIR approaches. Currently, the results sets are stored with the query image, but as images are added to the database over time the results set should change. The date needs to be added in order to prevent the comparison of newer images in competing CBIR approaches.



☑ potential mass(unnamed)

Name	Value
Name	potential mass
Location	(214, 270)
Date	2007-08-27
Type	unnamed

Tahmoush: Potential mass recognized, biopsy recommended

Failed to add attribute

Figure 22. The updated Archimedes image user interface. The search panel and results set are accessible from the tabs on the left, the zooming interface is on the left, and additional features and annotations are on the right.

Currently Archimedes only stores CBIR results, but future work would allow CBIR techniques to be stored and utilized within Archimedes as well. CBIR techniques that can be stored as a matrix operation on a feature vector can also be stored and used as an index.

4.4 Database Design

The design of Archimedes had to take into account the sensitive nature of the data as well as the multitude of regulations coming to govern this field. The design focused on satisfying HIPAA regulations in the US while maintaining the ability to adapt to other regulations.

There are four main sections to the design of a distributed database system. The first layer is the client including the Graphical User Interface (GUI). The second and third layers are the server and network protocol. The final layer is the underlying database selection and design. The issues driving the design are the security of the system and the capabilities needed to operate effectively.

The client section of the design was difficult because of the need to integrate an image tool with a data and search tool. The initial GUI is shown in Figure 21, while an updated GUI is shown in Figure 22. Instead of devoting most of the space to images as is a standard practice in image applications, Archimedes devoted a smaller space but augmented that space with the ability to zoom in on interesting parts of the image as well as grab the image and move it around within the available space. The images can also be downloaded into other viewing systems, and image sets are viewed through a tabbing system. Moving the images around appeared to be intuitive, and is similar to the technique used in GoogleMaps. However, the zooming was designed to be smooth and not stepped, with a mouse interface that was difficult to master. A planned upgrade will be to simplify the mouse interface for zooming in and out on images. The extra space was used to display auxiliary data like doctor's notes and the data searching interface and search results.

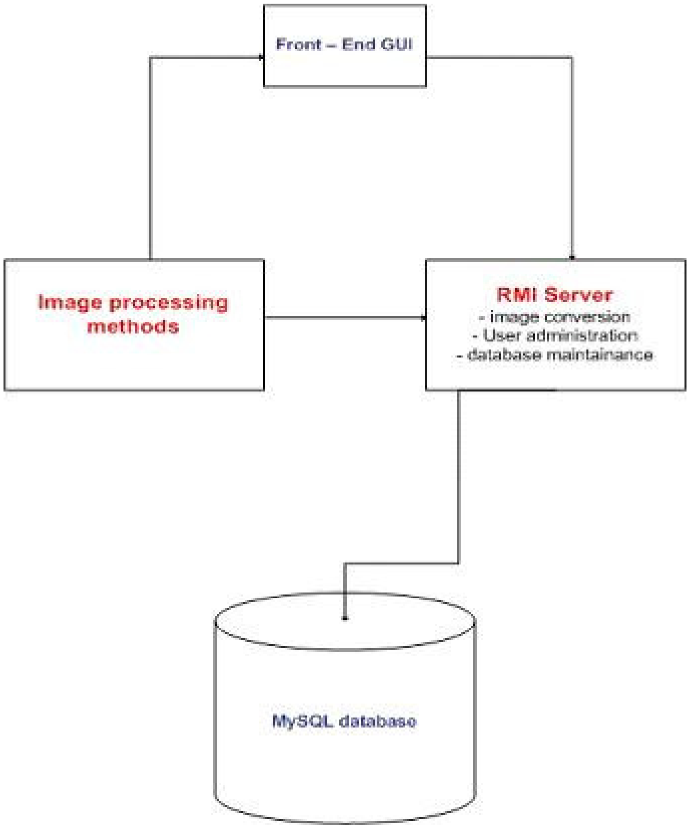


Figure 23. Archimedes High-Level Design. A web interface GUI is connected to a RMI server and a database. Note that the Image Processing methods are separate from the server.

The client design for searching and managing patient records and images focuses on searching by patient information, image features, and text. For search by patient information in Archimedes, the medical professional is allowed to enter patient information (i.e. first name, last name, date of birth, etc.) into Archimedes. Once information is entered, they can use filters to further refine their search results. Searching for images with specific features or spatial combinations of features allows doctors to further specify the results. Archimedes also allows search over text comments doctors previously made about patients or images. This works like a primitive Yahoo search over the text of the medical annotations. This free form of text search, combined with patient information, was surprisingly popular and simple to use, possibly because it mimicked a well-known searching application.

The system must maintain a high level of security due to privacy issues associated with maintaining sensitive patient medical information. The application is web-based for simplified deployment and tele-medicine uses, but this makes security more of an issue. Information transmitted from the server to the front-end is encrypted via the AES encryption scheme. All modifications during system use are monitored and logged by the system, and the viewing of the logs is limited to administrators. Images are transferred from the server to the client, where the client allows manipulations and exploration of the images. The transmission of images over an encrypted connection caused significant problems for the usability of the system. One key to improving the performance on images is to strip out all HIPAA-regulated information from the images and transmit it separately over an encrypted connection. Note that DICOM is not used to communicate between the client and the server, but is used when images are entered or downloaded into other applications. Because the choice of server environment was flexible, Java was chosen as the development language.

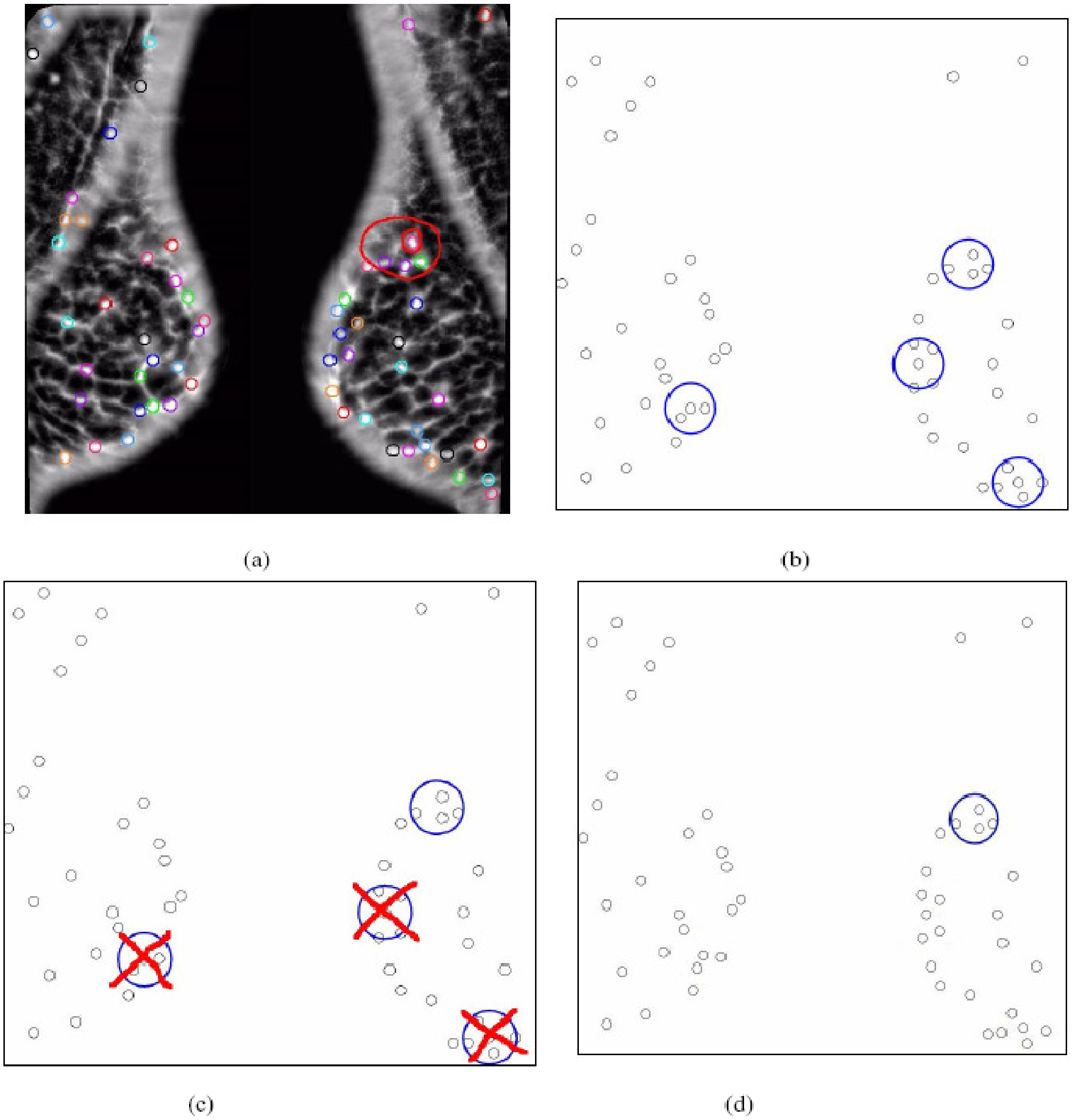


Figure 24. An example mammogram image pair that might be stored in Archimedes is in (a). Archimedes spatial search can be used to find clusters of features, and this is shown in (b). Storing these clusters as features, we can use a spatial query again to find all of the clusters that are a certain distance away from the breast boundary, which eliminates three noise clusters as shown in (c) and leaves only the actual cancer as shown in (d). This is one of the ways that Archimedes can be used to combine features.

There are many possible databases that could have been selected for use in Archimedes, including products from Oracle and Microsoft. However, the spatial search requirements of Archimedes dictated the database choice. Our prototype used MySQL for simplicity as is shown in Figure 23, but the final design uses the PostgreSQL open source SQL compliant relational database. PostgreSQL runs on all major operating systems including Linux, UNIX, BeOS, and Windows, which makes it highly portable and therefore extensible in the scope of our project. PostgreSQL allows all of the features of an advanced database, including transactions, tablespaces, and foreign keys. Using PostGIS, an extension for the PostgreSQL database, adds support for geographic objects and spatially enables the database. PostGIS complies with the Simple Features Specification for SQL and is an Open Source project as well. Both programs have been extensively tested and are considered secure, stable products. Using our Database API we are also able to support Oracle databases with the Spatial Extension. This makes the project more extensible for the future.

The design paid careful attention to the access to images because of privacy issues. Access to images can be either tightly controlled and private, public, or semi-private, while access to patient information is always tightly controlled and private. Administration is simplified through the use of groups, where semi-private images have groups of trusted medical professionals associated with them to help provide analysis. This is helpful for administration, where the hospital doctors or a subset are defined as the image default group setting.

The incorporation of DICOM 3 capability is necessary in any medical image system. However, XML support was also included mainly because it was simple. It became the preferred method for uploading data that was not already in DICOM format. The XML schema included tags such as <Patient>, <Image>, <Doctor>,<Feature>, <XPos>, <YPos>, and <Annotation>.

Image processing packages can be incorporated into Archimedes through the ability to upload not only text and images, but also features in images, their positions, and associated annotations. The separation of image processing packages and the database application was chosen to maintain the flexibility of the system and the ability to incorporate multiple different packages and is shown in Figure 23. The feature characteristics need to be flexible, and can be defined at upload.

Archimedes is a three-tiered application including backend server, server logic unit, and web front end user interface. The server can run on any machine using a Unix, Linux, or Windows operating system that can support Java. An example of how it can be used is in Figure 24.

5 Improvement to CAD

Having developed the similarity function for medical images described in Chapter 3 and an image database described in Chapter 4, we used the database to help improve the computer-aided detection of breast cancer. Computer-aided detection (CAD) of mammograms could be used to avoid missed diagnoses, and has been shown to increase the number of cancers detected by more than nineteen percent [41]. Improving the effectiveness of CAD could improve the detection of breast cancer, and could improve the survival rate by detecting the cancer earlier.

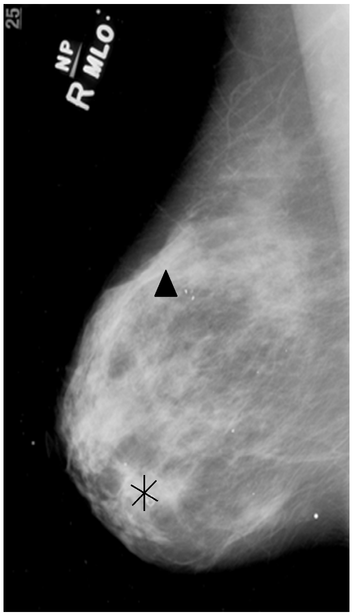


Figure 25. Typical CAD markups. The triangle marks a cluster of calcifications, and the star marks a potential mass.

The typical CAD system takes in a mammogram set and displays it for the radiologist. The system also provides markers on potential cancerous sites as found by the system. An example of these markers is shown in Figure 25. The determination of these markers and the evaluation of their effectiveness in helping radiologists are the main thrust of CAD research.

The hope for CAD is that the cancers missed by the radiologist are marked by the computer and brought to the attention of the radiologist. Most computer-aided detection (CAD) systems are tested on images which contain cancer on the assumption that images without cancer would produce the same number of false positives. However, a pre-screening system is designed to remove the normal cases from consideration, and so the inclusion of a pre-screening system into CAD dramatically reduces the number of false positives reported by the CAD system. We define three methods for the inclusion of pre-screening into CAD.

5.1 Incorporation of Asymmetry into CAD

There are three basic methods for including pre-screening into CAD analysis. The first is the strict method, where the pre-screening removes the non-cancerous cases entirely from the consideration of the CAD software. The second is probabilistic, where the probability of the case being cancerous or non-cancerous is determined by the pre-screening system and then incorporated into the CAD analysis. We also describe an improvement on our technique that we call an optimal approach, where a learning approach is used to try to determine the optimal factors for the inclusion of the pre-screening results into the CAD analysis. These methods will be defined and compared below.

The strict method is the simplest to define. Images that are screened as normal are removed from consideration by the CAD analysis. Since there are no false positives drawn from these cases, the number of false positives per image decreases. This is the most effective technique at reducing the number of false positives, but it is also the most dangerous as mistakes by the pre-screening system cannot be rectified by the CAD system.

The probabilistic method relies on the statistics of the pre-screening method to adjust the output of the CAD system. To incorporate prescreeing into a CAD system, we made use of Bayes Theorem, P(CancerSite | Pre-screen) = {P(Pre-screen | CancerSite) P(CancerSite) / P(Pre-screen)}. The sites where pre-screening indicates cancer are thus given an increased probability of being cancerous, while sites where pre-screening does not indicate cancer are given a reduced probability of being cancerous. Since the pre-screening measurement is applied to on entire case, all of the sites in those cases are affected similarly.

The optimal approach is a variant of the probabilistic approach, but instead of deriving the change from the underlying probabilities, the change is learned on a training set of cases. In theory, this approach can optimize the incorporation of pre-screening into CAD, but can be difficult in practice. In this case, P(CancerSite | Pre-screen) = A(Pre-screen) P(CancerSite), where A(Pre-screen) is the learned adjustment factor. This approach has more flexibility than the probabilistic approach, but is mush harder to implement. The choice of what to optimize is also a concern. There are two main options, optimizing the area under the ROC curve or optimizing the accuracy of the CAD results in a certain range of specificity. Both approaches were attempted and will be discussed.

5.2 Evaluation

The analysis was performed with the same cases that were used for the analysis in Chapter 3. The training data set was used to determine the parameter A(Pre-screen) for the optimal approach. The other approaches were tested against the same test set in order to be unbiased.

The results were good at low numbers of false positives in all three techniques, and it is at high and medium numbers of false positives where techniques distinguish themselves. Using the probabilistic approach to incorporate pre-screening into CAD is shown to work well at low numbers of false positives per image and can improve the performance by over 70%, but at high levels of false positives per image, this technique has minimal effect. This is expected since using Bayes Theorem merely reduces the probability of the false positives and does not eliminate them.

The results of the strict approach are identical to the results of the probabilistic approach at low levels of false positives, but diverge at higher levels of false positives. Since this approach eliminates the false positives instead of just diminishing them, the results at high levels of false positives per image are worse than the probabilistic approach because true positives are eliminated. However, in medium levels of false positives, the performance is significantly better than the probabilistic approach.

The optimal approach was tuned to determine the best performance at both low levels of false positives and the overall area under the ROC curve. The performance under both converged to the strict approach; however, this may be due to the pre-screening technique that was chosen.

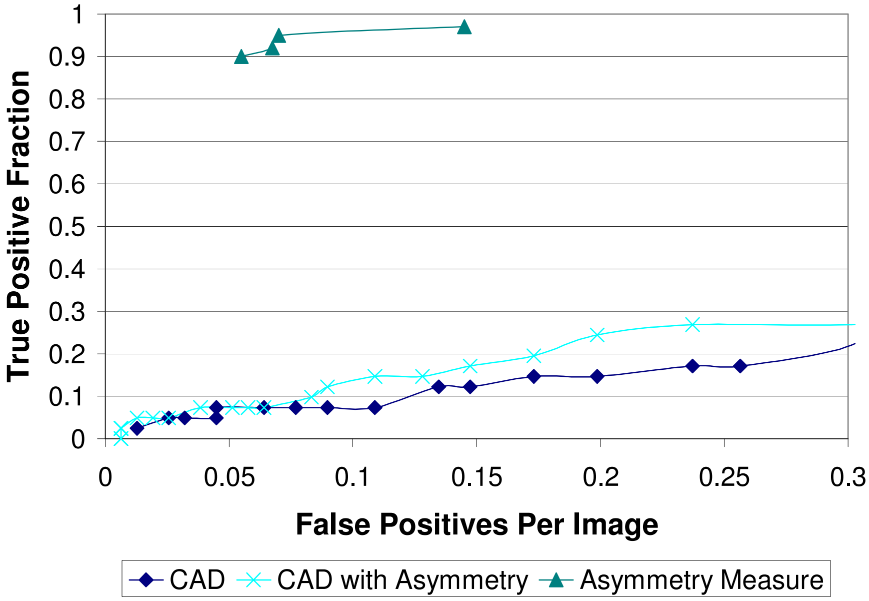


Figure 26: ROC curve comparing the CAD system before and after the inclusion of the Three-Cluster approach to measuring asymmetry. The inclusion of asymmetry improves the CAD system by up to 77%. The asymmetry measure has a very low level of false positives per image because it does not try to determine the position of the cancer, it merely determines the presence of cancer.

The overall performance is still strongly dependent on the effectiveness of the CAD system. The accuracy of the pre-screening is essential in order to prevent true positives from having their probabilities diminished, and the specificity is important for improving the effectiveness of the CAD system.

The incorporation of the classification results back into the original CAD system does significantly improve the original CAD system, as shown in Figure 26. The results of incorporating our classification into CAD were good, increasing the accuracy by up to 71% at a set level of false positives per image. The improvement is most apparent at low levels of false positives. Incorporating asymmetry into CAD can improve the effectiveness at low levels of false positives per image. We incorporated it as an afterthought, while it would be more effective as a feature used at the beginning of the CAD prompt calculation process. However, we did determine that asymmetry is a powerful technique by itself or incorporated into CAD. This indicates that further research into techniques that can compare images and thus measure asymmetry in mammograms may significantly improve the effectiveness of CAD algorithms.

6 Structured Classification and Retrieval

In order to create an effective classification technique for bioinformatics data, methods are needed to efficiently retrieve data based on similarity to a given exemplar or set of exemplars. This type of query is referred to as similarity retrieval. Of these queries, the nearest neighbor query is particularly important, and it is the one that is emphasized in this chapter. An apparently straightforward solution to finding the nearest neighbor is to compute a Voronoi diagram for the data points (i.e., a partition of the space into regions where all points in the region are closer to the region's associated data point than to any other data point), and then locate the Voronoi region corresponding to the query poInt. The problem with this solution is that the combinatorial complexity of the search process in high dimensions, expressed in terms of the number of objects, is prohibitive thereby making it virtually impossible to store the Voronoi diagram which renders its applicability moot.

Most methods utilize the information in the data and adjust the process to choose the best dimensions, but do not choose the best dimensions for each individual query point in order to improve the performance. In this chapter we explore the effectiveness of adjusting the retrieval process in response to the query process in response to the query point. Making use of the dimensions where the query point is near to a boundary instead of near the middle of the range provides a higher probability of pruning with that dimension. This method is significantly improved when distance functions with a higher order are used because the large contributions of a few dimensions are more relevant in that case. We also try to guarantee to not be worse than sequential search.

Nearest neighbor retrieval is a basic method used for classification. However, because of the curse of dimensionality, the difference in the distance to one class or the other becomes minimal and the accuracy suffers, prompting the use of methods like support vector machines (SVM) [68]. In this paper we compare nearest and farthest neighbor classifications that have been modified with our high-dimension techniques with SVM classifications to determine whether the curse of dimensionality has been reduced.

Nearest neighbor techniques often use the Minkowski metrics like the the L_2-norm to measure similarity between data points. However, the L_2-norm is not necessarily relevant to many emerging applications involving high-dimensional data [1]. Often these are used after dimension-reduction techniques like SVD. We experiment with a new reduced-dimension distance function that is designed to rapidly determine the maximum lower bound on the high-dimensional distance.

In high-dimensional nearest neighbor there are both indexed methods like the GESS method [31] and grid structures [69], and unindexed approaches. The method in this paper is an unindexed approach.

Several new approaches are discussed in this chapter, including choosing the dimensions to analyze based on the dimensions that are relevant to both the data and the query point, called dimensional choice, and a new distance function that measures the maximum lower bound on the high-dimensional distance, called the UL-Distance. The combination of improved dimensional ranking and a distance function that uses fewer dimensions is shown to be an effective combination.

6.1 UL-Distance

Normal nearest neighbor approaches break up the feature space well by creating a Vornoi space, but are susceptible to bad data points. A typical distance metric that is used is a Minkowski metric of order U as in Equation 4.

4)

$$UL - Dist = \sqrt[U]{\sum_{i=0}^{L} \max_i \left(| x_i - q_i |\right)^U}$$

where d is the number of dimensions in the feature space. Here $x_i - q_i$ indicates the difference in the data point x and the point to be classified q in the ith dimension. The Euclidean distance function uses a value of U=2, while the Manhattan distance function uses U=1. We define a new non-metric distance function called the UL-Distance which is defined to be a maximum lower limit on a high-dimensional Minkowski distance metric in Equation 5.

5)

$$Dist = \sqrt[U]{\sum_{i=0}^{d} \left(| x_i - q_i |\right)^U}$$

The max function here picks out the ith maximum from the set. The number of dimensions used in the distance is expressly limited to L, which will speed up retrieval, but the contributions are from the dimensions that will maximize the distance in order to maintain as much accuracy as possible. The factor L is the number of dimensions used in the distance, and the U is the order of the distance function. Though this technique provides an approximation to the high-dimensional distance that can be used for pruning, it can also be an effective distance function on its own. Note that it is not a distance metric because the triangle inequality is not guaranteed to hold for different query points. This distance function is sensitive to the few dimensions that are different, instead of being overwhelmed by the number of dimensions that are similar. A second non-metric distance function called the LL-Distance is nearly identical to the UL-Distance except that it calculates the minimum lower limit and is Equation 6.

6)

$$LL - Dist = \sqrt[U]{\sum_{i=0}^{L} \min_i \left(| x_i - q_i | \right)^U}$$

In order to provide an example of why using a distance function that uses all of the dimensions but only calculates with a few could be significantly faster than a Euclidean distance, we use a Chessboard distance and calculate the farthest neighbor as shown in Figure 27. Though this may not be the most useful calculation, it is the simplest example. When doing a farthest-neighbor search, using the diameter of the data can be an effective technique. Since the Chessboard distance metric requires finding the point with the maximum difference in only one dimension, storing the points that are on the diameter of the data set allows the lookup of the farthest neighbor in each dimension. The point (or points if multiple points with the same value are stored) on the diameter of each dimension is compared with the query point, the dimension with the maximum is determined, and the farthest neighbor is looked up. This would allow the calculation of the farthest neighbor under this distance

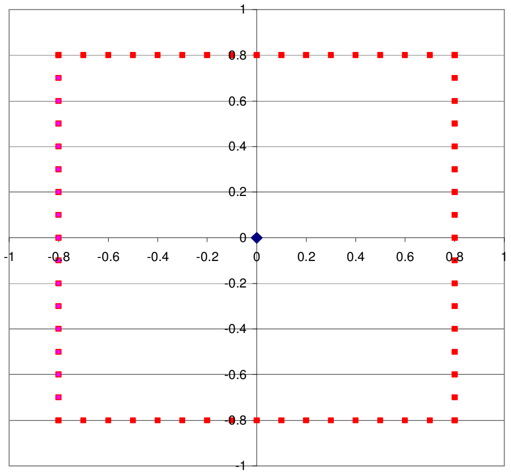


Figure 27: Worst-Case 2D Search. The data in this example are the small squares, while the query point is the diamond at (0,0). The points on the diameter are stored in the data structure, or a single point can be used if space is an issue. All of the data points are equidistant from the query point, requiring all four surfaces to be accessed to find all farthest neighbors in the chessboard distance function.

metric in O(d) time with O(d) storage. This is the same amount of work necessary to calculate one distance, and is a factor of n better than sequential search. This example demonstrates that the dimensionality of the distance function is very important for the performance of retrievals.

Since the number of dimensions used by the UL-Distance function is limited to L, this distance function is effectively a dimension reduction technique that operates on all of the dimensions. An additional technique is required to make the retrieval faster.

These distance functions do fulfill the properties of positive definiteness, where D(a,b) >= 0 for all a and b, symmetry, where D(a,b) = D(b,a), and identity, where D(a,a)

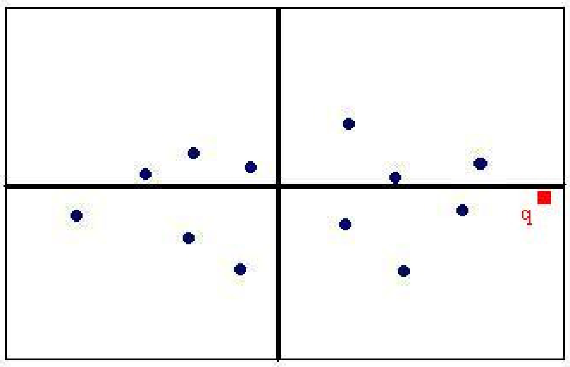


Figure 28: An example 2D data set where choosing to search using the x-dimension is preferred. The red square q is the query point, and the blue circles are the data. In this case the x-dimension is very significant for determining the nearest neighbor, while the contributions from the y-dimension are not as significant and the y-dimension could be neglected. Dimensional Choice would let us choose the x-dimension and ignore the y-dimension unless it is needed.

= 0. However, it is not guaranteed to fulfill the triangle inequality because different dimensions are used.

An alternative approach to limiting the number of dimensions is to only take dimensions that contribute more than a certain threshold. However, that makes comparisons with other techniques difficult. Note that these distance functions are not normalized so that comparing the distances with different values of L can be misleading, which makes the threshold approach more difficult. Normalization is feasible for particular values of U.

6.2 Using Dimensional Choice

In low-dimensional search, the choice of which dimension to incorporate into the search first is not that important, but examples where is is are shown in Figures 28 and 29. However, when there are thousands of dimensions in the data set, the choice is much more important. Choosing the best dimension to start the search does require additional work to determine the best dimension, with work of O(d) to O(d log d) to sort the dimensions, as well as knowledge of the diameters of the data set. However, this is only a small amount of work compared to a complete high-dimensional search, which for sequential search is O(nd) where n is the number of data points and d is the number of dimensions. Additionally, if one is using SVD as is often recommended when using high-dimensional data, the initial transformation or projection into the SVD coordinates dominates the work required to implement dimensional choice.

The underlying data structure must be extremely flexible in order to utilize dimensional choice, which is why it is not used in low-dimensional cases. The idea behind this technique was mentioned by Nene and Nayar [85], where they suggest a projection method that could order the analysis of the dimensions in order to minimize the total work. However, they were working with only sixteen dimensions, so we analyze the full effect of this technique on their projection method. Their technique determines the points that are within a distance of ε from the query point by accessing the data in each dimension and winnowing down the potential nearest neighbors. The distance ε that should be used is determined to be rather large when there are sparse data points and a large number of dimensions.

Dimensional choice can be used to first estimate the dimensions that have the largest potential to winnow down the number of potential nearest neighbors without actually

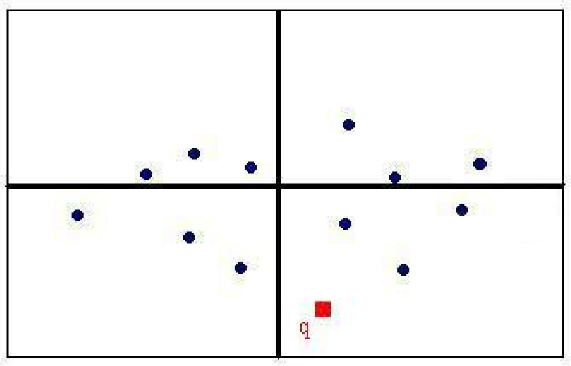


Figure 29: An example 2D data set where choosing to search using the y-dimension is valuable in determining the nearest neighbor to the query point q, even though the data set would indicate that a x-dimension is preferred. The red square q is the query point, and the blue circles are the data. In both the x and y dimensions, the contributions to the total distance can be significant. The difference between this case and the case in Figure 28 is the position of the query position.

analyzing those dimensions. This choice is distribution dependent and could be calculated as such. Note that when the query point is at the edge of the data set, the space that has to be searched is only ε instead of 2ε (since the range is from x-ε to x+ε and in this case half of the space will be empty). So in the case of the uniform distribution (where this technique works the worst), there are dimensions that can as much as double the effective winnowing. In the case of Gaussian distributed data, the effect is even better because the winnowing is done at the tails of the distribution. Additionally, the important dimensions can be determined a priori, so that many dimensions need never be analyzed.

In order to realize the effect of the improved winnowing, an additional adjustment should be included to the Nene and Nayar approach. Their approach continues through all of the dimensions regardless of the number of points remaining in the hyper-cube. A stopping condition should be included so that analyzing the dimensions stops when there are a set number of points left in the hyper-cube. Using dimensional choice reduces the work by at least a factor of two for a uniform distribution in high dimensions, and a significantly better factor for a Gaussian distribution.

PCA analysis utilizes a limited number of the eigenvectors V with the largest eigenvalues λ of the diagonalized covariance matrix D to limit the dimensions. However, this neglects the importance of the query point itself. The difficulty with this is demonstrated by comparing Figures 36 and 37, where the query point determines whether the dimension can be neglected. Dimensional choice can be built as an extension of PCA in the following way. While PCA selects the eigenvectors with the largest variance λ, the query point can be included by selecting the dimensions with the largest value of the difference from the mean (q_i - μ_i) and the largest variance λ. We use a combination factor *C* to balance these two factors to give us a priority value P

7)

where i is the appropriate dimension, *q* is the query point, μ is the mean. Selecting the dimensions based on P-Value instead of λ gives the dimension prioritization a sensitivity to the query poInt.

$$P - Value = \lambda_i + C(q_i - \mu_i)$$

Combining the UL-Distance and dimensional choice methods for nearest and farthest neighbor searches can provide significant improvement in speed. In order to determine the farthest neighbor in the UL-Distance, the dimensions of the query point are compared with the mean and variance of that dimension. Then the dimension which has the highest possible contribution is analyzed first to get a distance. The remaining dimensions are checked until the current distance difference cannot be exceeded because the potential contributions from the remaining dimensions are too small. An additional level of approximation can be included by estimating an earlier stopping poInt. This method operates in O(d log d + na)) where a is the number of dimensions that had to be analyzed at the worst case and is dominated by the initial sort, but can be reduced to O(d + n) ~ O(d) if only a partial sort is done initially. This compares favorably with the O(nd) of sequential scans.

In the case of the Euclidean or Manhattan distance metrics, the gains from adapting to the query point are not as profound as under the UL-distance because all of the dimensions have to be analyzed. However, we have demonstrated that the dimension of the distance function is what drives the difficulty in retrieval. This motivates the creation of new distance metrics like the UL-Distance that emulate Minkowski distance metrics but use a lower dimensionality.

Using Dimensional Choice differs from using PCA in several important ways. First, PCA uses the same dimensions for every classification distance, while Dimensional Choice is adaptive and uses a different set of dimensions for each classification distance depending on the dimensions that are important for that particular point as well as those that are important for the data overall. Dimensional Choice works better with distance functions that are inherently lower dimensional like the UL-Distance and the Chessboard distance functions because the combination of a limited number of dimensions and an effective choice of those dimensions complement each other.

6.3 Search using Similar Neighbor

Many nearest neighbor applications require the exact nearest neighbor. However, when looking for similarity, often the approximate nearest neighbor is sufficient. Judging whether approximate nearest neighbor is good enough requires an understanding of the underlying structure of similarity that is embedded into the space. The amount of approximation allowed depends on the tolerance of the system for mis-classification of points as similar. A better approach is finding a similar neighbor instead of the nearest neighbor. This avoids the discussion of how much approximation is tolerable by going directly to the question of similarity.

An example of success in similar neighbor would be finding a point that is not the nearest neighbor, but is similar to the query poInt. An example of failure would be finding any point that is not similar, even if it is the nearest neighbor. This measure of success is less strict in terms of actual distances to the objects that are retrieved but more strict in terms of the similarity of the objects to the query.

6.4 Experiments

The main questions for these techniques are what the speed improvement is, and what the change in accuracy is. In order to determine the change in accuracy, we look at a nearest neighbor application in recognizing prostate cancer. Here the loss in accuracy is judged by whether the classification loses accuracy, sensitivity, or specificity, instead of determining whether the particular nearest neighbor is exactly the same. This looser definition of accuracy is more of a functional definition as large high-dimensional data sets are increasingly used for classification. The accuracy will be compared with other nearest neighbor and SVM approaches.

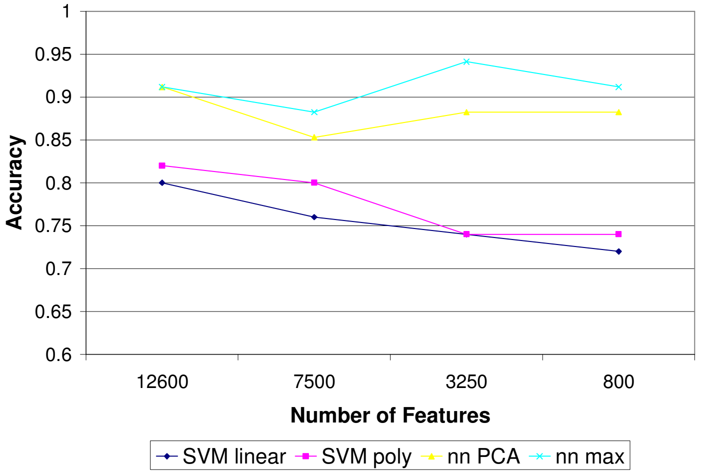


Figure 30: The accuracy of the methods versus the number of features (or dimensions). The nearest neighbor methods performed surprisingly well against the SVM. The nn PCA method is the nearest neighbor with PCA dimensions included, while the nn max uses the UL-Distance. The general flatness of the nearest neighbor methods is encouraging for using nearest neighbor with dimension reduction methods. The nn max does outperform all other methods. The general flatness of the nearest neighbor methods is encouraging for using nearest neighbor with dimension reduction methods.

We used a data set obtained from Clinical Proteomic Program Databank. The experimental data is a set of prostate cancer samples. The experiment analyzed serum proteomic mass spectra generated by SELDI-TOF to discriminate the sera of men with histopathologic diagnosis of prostate cancer (serum prostate-specific antigen [PSA] ≥ 4 ng/mL) from those men without prostate cancer (serum PSA < 1 ng/mL). In this data set, there are 63 normal (non-cancer) samples, and 69 cancer samples.

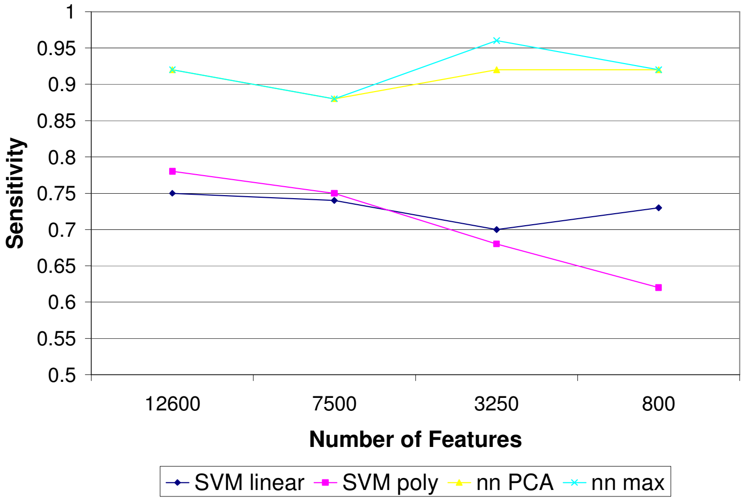


Figure 31: The sensitivity of the methods versus the number of features. The nn PCA method is the nearest neighbor with PCA dimensions included, while the nn max uses the UL-Distance. The similarity of the sensitivity of the two methods suggests that it is the specificity and not the sensitivity that makes the nn max the better technique. Both outperform SVM.

A SVM was used to compare the accuracy loss for the serum proteomic pattern analysis. A SVM is a blend of linear modeling and instance-based learning. A SVM selects a small number of critical boundary samples, called support vectors, from each category and builds a linear discriminate function that separates them as widely as possible. A kernel is used to automatically inject the training samples into a higher-dimensional space, and to learn a separator in that space [68]. In linearly separable cases, SVM constructs a hyper-plane, which separates the two different categories of feature vectors with a maximum margin, i.e., the distance between the separating hyper-plane and the nearest training vector. The training instances that lie closest to the hyper-plane are support vectors [68]. Linear and polynomial kernels were used. The feature selection method was MIT correlation, which is also known as signal-to-noise statistic [46].

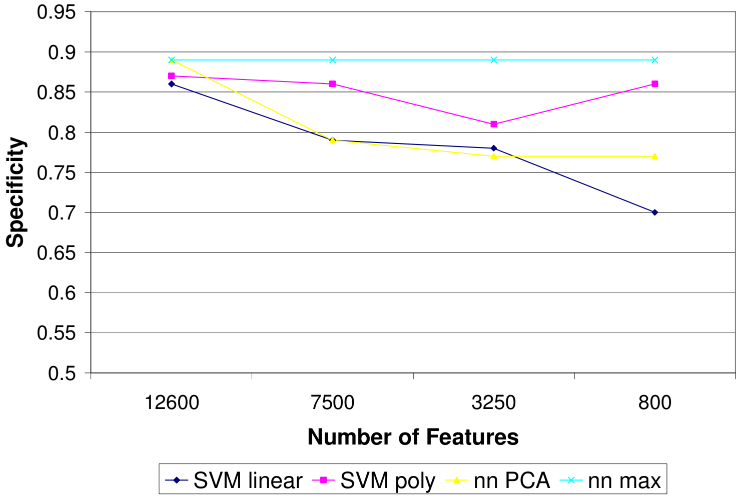


Figure 32: The specificity of the methods versus the number of features. The nn PCA method is the nearest neighbor with PCA dimensions included, while the nn max uses the UL-Distance. The flatness of the nn max method demonstrates that the dimension reduction does not adversely affect the specificity.

The speed and accuracy improvement was measured on the same computer with the competing algorithms of the PCA nearest neighbor with a Euclidean distance metric versus the UL-Distance of order 2 with Dimensional Choice. The accuracy was also compared with two SVM approaches. Because of the limited supply of data, we used one sample as the test case and the remainder as the training cases and did this for each case. The drawback to this approach is that the result of each individual test is not independent of the results of the other tests.

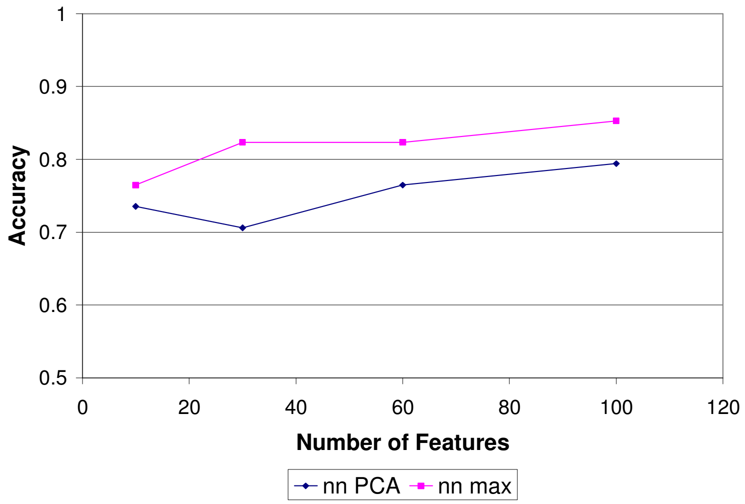


Figure 33: The accuracy of the nearest neighbor methods versus the number of features at a small number of features. The nn PCA method is the nearest neighbor with PCA dimensions included, while the nn max uses the UL-Distance. The loss in accuracy at an extremely small number of features is significant, but not terrible. The performance of the UL-Distance does improve the performance by 3-12% over the PCA features technique.

6.5 Results

The results were interesting overall as the nearest neighbor classification performed better overall than both of the SVM techniques. This is shown in Figures 38, 39, and 40. The use of the UL-Distance significantly increased the accuracy of the nearest-neighbors technique at low levels of features, as is shown in Figure 33, but performed at a similar level to the PCA choice of dimensions at high levels of features, as is shown in Figures 30, 31, and 32.

The accuracy of the classification is maintained with the reduction of the number of features from 12600 to 800 with the UL-Distance, while the PCA choice of dimensions shows slight degradation as is shown in Figure 30. This reduction in the dimensionality of the data by almost 70% without a loss of accuracy is encouraging. However, the accuracy is degraded below 100 dimensions as shown in Figure 33, but only by 6% in order to achieve a dimension reduction of 99%. Note that using the UL-Distance instead of the PCA technique improved the accuracy by up to 12% at low numbers of dimensions, as is shown in Figure 33. Of course, these results are dependent on the data set used.

The specificity of the classification is surprisingly stable with dimension reduction under the UL-Distance, as is shown in Figure 32. The other methods did not fare as well. The sensitivity of the classification with the UL-Distance and nearest neighbor was surprisingly good, as is shown in Figure 31.

Farthest neighbor classification did not perform well and is not shown. However, the farthest neighbor and nearest neighbor classifications did not tend to misclassify the same data points, which implies that the combination of the two might produce a better overall classifier.

6.6 Conclusion

This work has demonstrated significant dimension reduction, up to 70% reduction in the number of dimensions in the data set with no loss in accuracy or over 99% reduction with only a 6% loss in accuracy. The method can actually perform better with fewer dimensions than the nearest neighbor with all of the dimensions. The data set may be part of the reason, though it is a typical prostate cancer data set.

We have developed a new distance function called the UL-Distance that can be effectively used to replace Euclidean or other Minkowski metrics for high-dimensional nearest neighbor operations. This performed at up to 12% better than alternate approaches.

Combining this new distance function with a technique of Dimensional Choice where the best dimensions to analyze are guessed using information about the underlying data and the query itself in order to minimize the amount of work required to perform the nearest neighbor search with the UL-Distance achieved significant savings in work. The time to perform a nearest neighbor search is reduced by a factor of five with no loss of accuracy, but can be improved up to a factor of ten at some loss of accuracy.

We demonstrate that the curse of dimensionality is not based on the dimension of the data itself, but primarily upon the effective dimension of the distance function. The effective dimension of the UL-Distance is set to a factor of L even though it can act on any of the possible d dimensions. We also note that the higher the order U of the UL-Distance function, the better the approximation performs since the small factors that would be included from neglected dimensions are effectively reduced when using a higher order distance function.

We note that this work is preliminary and does require more extensive analysis. However, the combination of a more effective ranking of dimensions using dimensional choice and a dimension-limiting distance function appear to be an effective combination when using high-dimensional data.

7 Conclusion

This thesis touched on many of the problems facing the classification and retrieval of cancer images and data. We developed a method for differencing and classifying images, which we then incorporated it into CAD. We developed a database for the collection and analysis of cancer images and data. We also analyzed better approaches to retrieve and analyze high-dimensional cancer data.

Our results are strong on all cases of the test set for classifying breast cancer images, correctly classifying with 85% accuracy and our technique outperforms both the best academic and commercial approaches, suggesting that this is an important technique in the classification of mammograms. We have also shown that using the image comparisons to determine the classification is insensitive to the parameters of the approach.

We created and compared multiple models, demonstrating improved results over both academic and commercial approaches. We also defined a new distance measure for the comparison of point sets and demonstrate its effectiveness in this application. The coupling of this distance measure with the parametric learning of clusters led to a highly effective classification technique.

The clusters also discovered an area of interest in mammogram comparisons which improved the diagnosis of mammograms that did not have cancer. More clusters might improve the technique, or, more importantly, they might lead to the discovery of more areas of interest. We suggested several ways that might improve on the methods that we used to compare mammograms.

The incorporation of the classification results back into the original CAD system does significantly improve the original CAD system. The results of incorporating our classification into CAD were good, increasing the accuracy by up to 71% at a set level of false positives per image. The improvement is most apparent at low levels of false positives. Incorporating asymmetry into CAD can improve the effectiveness at low levels of false positives per image. We also determined that asymmetry is a powerful technique by itself or incorporated into CAD. This indicates that further research into techniques that can compare images and thus measure asymmetry in mammograms may significantly improve the effectiveness of CAD algorithms.

We have created a secure web-enabled HIPPA-compliant database for the storage, retrieval, manipulation, and annotation of medical images and medical records for the development and evaluation of CBIR methods. The most unique quality is the ability to input, store, and share multiple feature sets and result sets for each image, thereby allowing greater flexibility for CBIR and allowing web collaboration in the development of CBIR. Each expert needed for the development of CBIR gains advantages in their individual work by collaborating in Archimedes, while improving the project overall. The advanced querying and feature storage capabilities provide rapid analysis and comparisons radiologist techniques, medical image features and CBIR techniques.

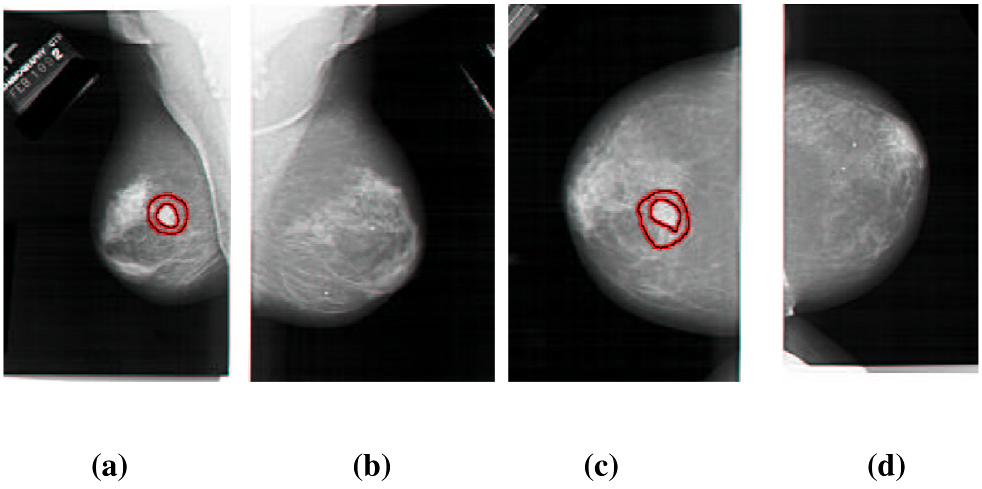
The work on retrieval of cancer information has demonstrated significant dimension reduction, up to 70% reduction in the number of dimensions in the data set with no loss in accuracy or over 99% reduction with only a 6% loss in accuracy. The method can actually perform better with fewer dimensions than the nearest neighbor with all of the dimensions. The data set may be part of the reason, though it is a typical prostate cancer data set.

We developed a new distance function called the UL-Distance that can be effectively used to replace Euclidean or other Minkowski metrics for high-dimensional nearest neighbor operations. This performed at up to 12% better than alternate approaches. Combining this new distance function with a technique of Dimensional Choice where the best dimensions to analyze are guessed using information about the underlying data and the query itself in order to minimize the amount of work required to perform the nearest neighbor search with the UL-Distance achieved significant savings in work.

We demonstrate that the curse of dimensionality is not based on the dimension of the data itself, but primarily upon the effective dimension of the distance function. The effective dimension of the UL-Distance is set to a factor of L even though it can act on any of the possible d dimensions. We also note that the higher the order U of the UL-Distance function, the better the approximation performs since the small factors that would be included from neglected dimensions are effectively reduced when using a higher order distance function.

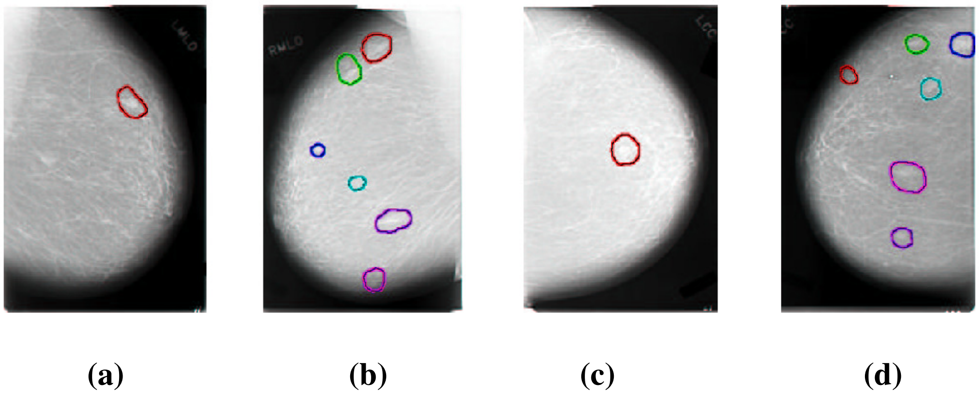
We note the combination of a more effective ranking of dimensions using dimensional choice and a dimension-limiting distance function appear to be an effective combination when using high-dimensional data.

Appendix A Mammogram Images



(a) (b) (c) (d)

Figure 34: The typical set of four images that make up a mammogram, the side view of the left breast in (a), the side view of the right breast in (b), the top view of the left breast in (c), the top view of the right breast in (d). The cancerous areas are outlined in red. This image set was correctly classified by the method described in Chapter 3.



(a) (b) (c) (d)

Figure 35: Round masses with circumscribed margins. The side view of the left breast is in (a), the side view of the right breast is in (b), the top view of the left breast is in (c), the top view of the right breast is in (d). The cancerous areas are outlined in red.

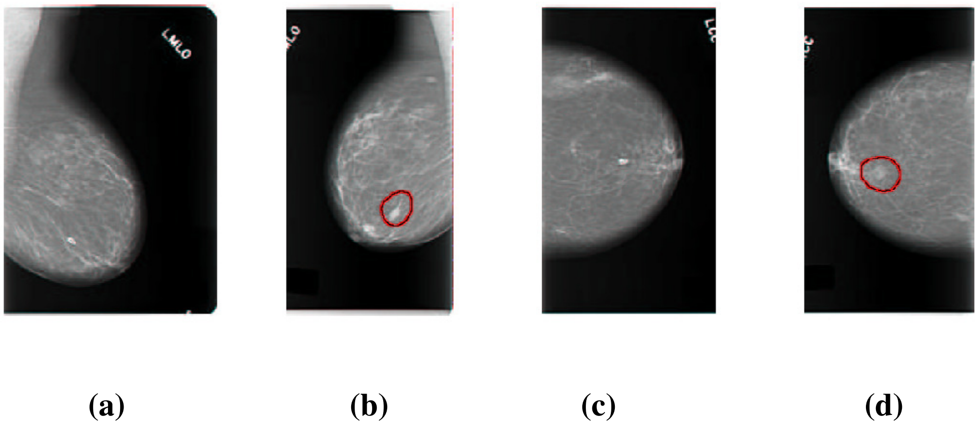


Figure 36: A round mass with microlobulated margins. The side view of the left breast is in (a), the side view of the right breast is in (b), the top view of the left breast is in (c), the top view of the right breast is in (d). The cancerous areas are outlined in red.

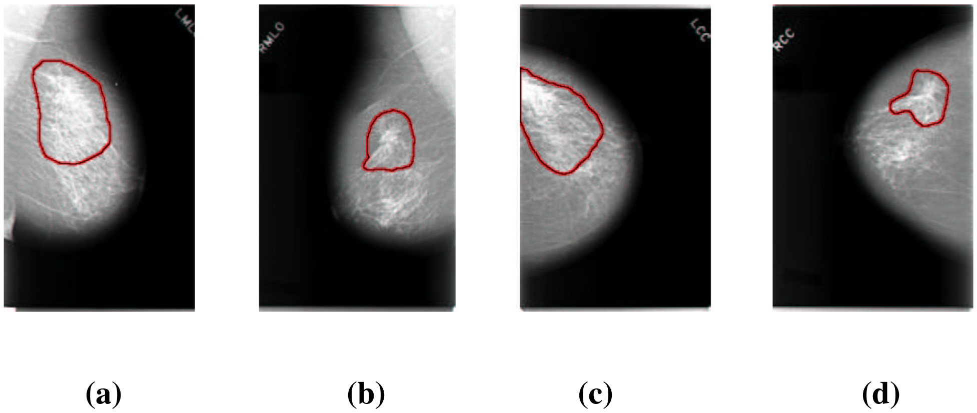
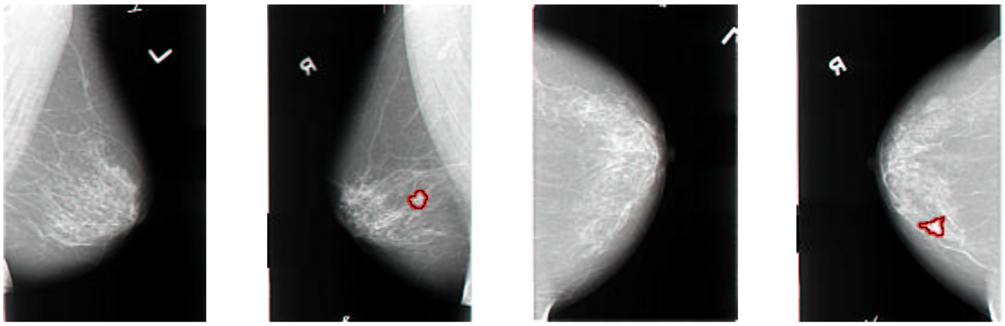
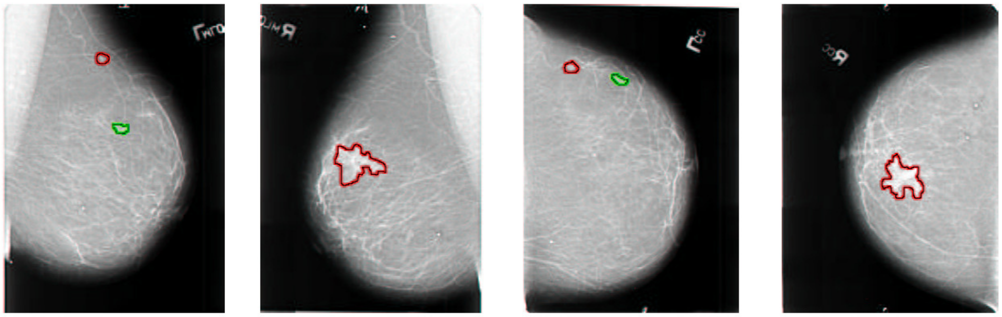


Figure 37: An architectural distortion with spiculated margins. The side view of the left breast in (a), the side view of the right breast in (b), the top view of the left breast in (c), the top view of the right breast in (d). The cancerous areas are outlined in red.



(a) **(b)** **(c)** **(d)**

Figure 38: An architectural distortion with microlobulated margins. The side view of the left breast is in (a), the side view of the right breast is in (b), the top view of the left breast is in (c), the top view of the right breast is in (d). The cancerous areas are outlined in red.



(a) **(b)** **(c)** **(d)**

Figure 39: A cancer in a lymph node. The side view of the left breast is in (a), the side view of the right breast is in (b), the top view of the left breast is in (c), the top view of the right breast is in (d). The cancerous areas are outlined in red.

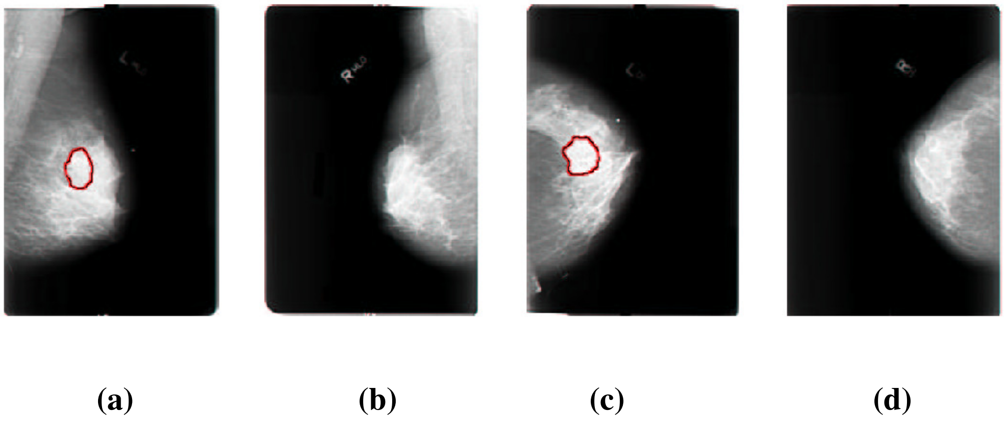

(a) (b) (c) (d)

Figure 40: A focal asymmetric density. The side view of the left breast in (a), the side view of the right breast in (b), the top view of the left breast in (c), the top view of the right breast in (d). The cancerous areas are outlined in red.

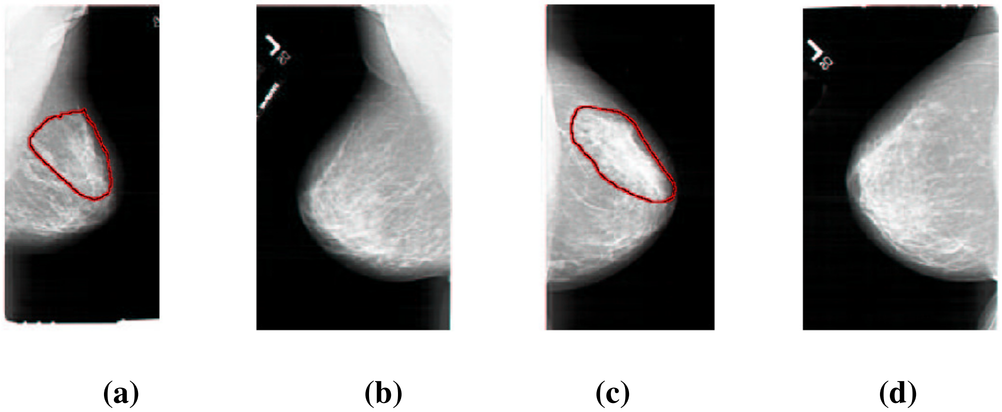
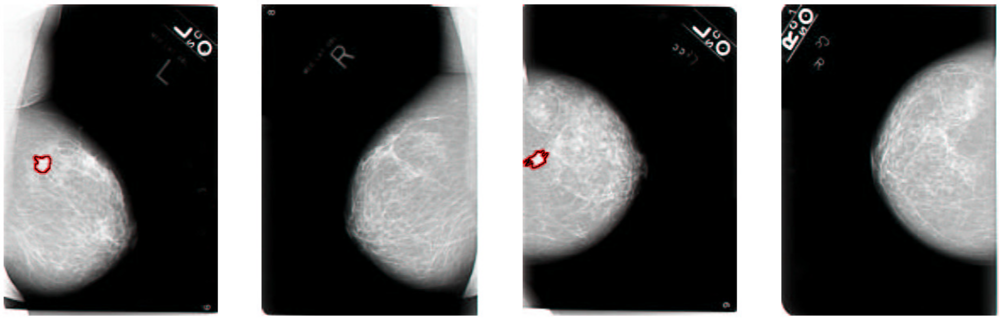
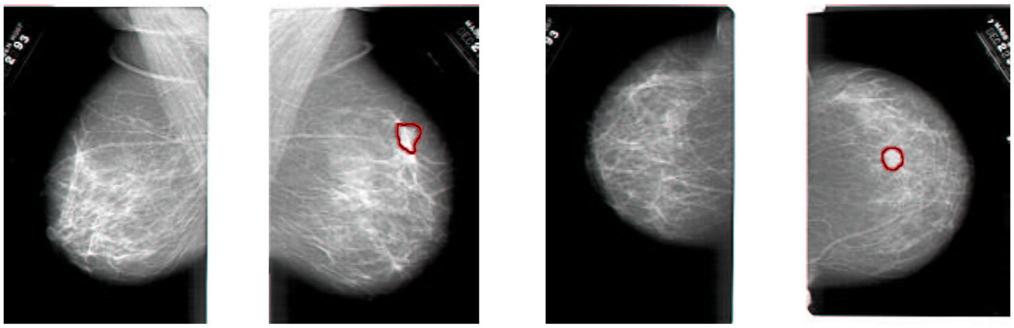

(a) (b) (c) (d)

Figure 41: A cancerous asymmetric breast tissue. The side view of the left breast is in (a), the side view of the right breast is in (b), the top view of the left breast is in (c), the top view of the right breast is in (d). The cancerous areas are outlined in red.



(a) (b) (c) (d)

Figure 42: A lobulated mass with spiculated margins. The side view of the left breast in (a), the side view of the right breast in (b), the top view of the left breast in (c), the top view of the right breast in (d). The cancerous areas are outlined in red.



(a) (b) (c) (d)

Figure 43: A round mass with an obscured margin. The side view of the left breast in (a), the side view of the right breast in (b), the top view of the left breast in (c), the top view of the right breast in (d). The cancerous areas are outlined in red.

Appendix B Data

Training Data

Name	Age	Density	Assessment	Subtlety	Digitizer
A_1134_1	69	2	4	4	DIGITIZER HOWTEK 43.5
A_1156_1	65	3	5	5	DIGITIZER HOWTEK 43.5
A_1159_1	69	1	5	5	DIGITIZER HOWTEK 43.5
A_1160_1	53	2	4	1	DIGITIZER HOWTEK 43.5
A_1163_1	70	4	5	3	DIGITIZER HOWTEK 43.5
A_1166_1	61	2	5	5	DIGITIZER HOWTEK 43.5
A_1174_1	80	2	5	4	DIGITIZER HOWTEK 43.5
A_1203_1	51	4	4	2	DIGITIZER HOWTEK 43.5
A_1212_1	30	4	5	5	DIGITIZER HOWTEK 43.5
A_1217_1	65	3	4	2	DIGITIZER HOWTEK 43.5
A_1222_1	40	3	5	5	DIGITIZER HOWTEK 43.5
A_1224_1	57	3	4	3	DIGITIZER HOWTEK 43.5
A_1229_1	65	2	5	4	DIGITIZER HOWTEK 43.5
A_1236_1	58	4	5	5	DIGITIZER HOWTEK 43.5
A_1252_1	67	3	5	4	DIGITIZER HOWTEK 43.5
A_1262_1	58	2	5	5	DIGITIZER HOWTEK 43.5
A_1403_1	57	3	5	5	DIGITIZER HOWTEK 43.5
A_1417_1	66	3	5	5	DIGITIZER HOWTEK 43.5
A_1467_1	40	3	5	1	DIGITIZER HOWTEK 43.5
A_1486_1	67	4	5	2	DIGITIZER HOWTEK 43.5
A_1520_1	52	4	3	4	DIGITIZER HOWTEK 43.5
A_1587_1	71	3	5	2	DIGITIZER HOWTEK 43.5
A_1589_1	69	4	5	4	DIGITIZER HOWTEK 43.5
A_1592_1	40	4	4	1	DIGITIZER HOWTEK 43.5
A_1620_1	67	2	5	3	DIGITIZER HOWTEK 43.5
A_1622_1	63	3	4	2	DIGITIZER HOWTEK 43.5
A_1642_1	71	2	5	5	DIGITIZER HOWTEK 43.5
A_1671_1	78	2	4	2	DIGITIZER HOWTEK 43.5
A_1693_1	44	4	5	4	DIGITIZER HOWTEK 43.5
A_1700_1	77	3	5	1	DIGITIZER HOWTEK 43.5
A_1701_1	71	1	5	3	DIGITIZER HOWTEK 43.5
A_1720_1	71	4	4	2	DIGITIZER HOWTEK 43.5
A_1726_1	61	3	4	3	DIGITIZER HOWTEK 43.5
A_1790_1	46	4	4	1	DIGITIZER HOWTEK 43.5
A_1896_1	87	3	5	5	DIGITIZER HOWTEK 43.5

A_1899_1	43	3	5	4	DIGITIZER HOWTEK 43.5
A_1908_1	48	2	5	4	DIGITIZER HOWTEK 43.5
D_4500_1	63	3	5	4	DIGITIZER HOWTEK 43.5
D_4501_1	56	2	5	4	DIGITIZER HOWTEK 43.5
D_4502_1	41	1	5	4	DIGITIZER HOWTEK 43.5
D_4503_1	52	2	5	4	DIGITIZER HOWTEK 43.5
D_4505_1	58	2	5	4	DIGITIZER HOWTEK 43.5
D_4506_1	36	4	5	4	DIGITIZER HOWTEK 43.5
D_4508_1	37	3	5	4	DIGITIZER HOWTEK 43.5
D_4510_1	42	2	5	4	DIGITIZER HOWTEK 43.5
D_4511_1	50	2	5	4	DIGITIZER HOWTEK 43.5
D_4512_1	43	3	5	4	DIGITIZER HOWTEK 43.5
D_4513_1	37	3	5	4	DIGITIZER HOWTEK 43.5
D_4514_1	43	3	5	4	DIGITIZER HOWTEK 43.5
D_4515_1	38	2	5	4	DIGITIZER HOWTEK 43.5
D_4516_1	42	3	5	4	DIGITIZER HOWTEK 43.5
D_4517_1	42	3	5	4	DIGITIZER HOWTEK 43.5
D_4518_1	44	3	5	4	DIGITIZER HOWTEK 43.5
D_4519_1	55	1	5	4	DIGITIZER HOWTEK 43.5
D_4520_1	37	4	5	4	DIGITIZER HOWTEK 43.5
D_4521_1	39	2	5	4	DIGITIZER HOWTEK 43.5
D_4522_1	41	4	5	4	DIGITIZER HOWTEK 43.5
D_4523_1	62	2	5	4	DIGITIZER HOWTEK 43.5
D_4524_1	48	4	5	4	DIGITIZER HOWTEK 43.5
D_4525_1	61	1	5	4	DIGITIZER HOWTEK 43.5
D_4526_1	52	1	5	4	DIGITIZER HOWTEK 43.5
D_4527_1	51	3	5	4	DIGITIZER HOWTEK 43.5
D_4528_1	50	4	5	4	DIGITIZER HOWTEK 43.5
D_4529_1	58	1	5	4	DIGITIZER HOWTEK 43.5
D_4530_1	58	4	5	4	DIGITIZER HOWTEK 43.5
D_4532_1	52	2	5	4	DIGITIZER HOWTEK 43.5
D_4533_1	47	2	5	4	DIGITIZER HOWTEK 43.5
D_4534_1	36	1	5	4	DIGITIZER HOWTEK 43.5
D_4536_1	47	1	5	4	DIGITIZER HOWTEK 43.5
D_4537_1	51	3	5	4	DIGITIZER HOWTEK 43.5
D_4538_1	41	3	5	4	DIGITIZER HOWTEK 43.5
D_4539_1	45	3	5	4	DIGITIZER HOWTEK 43.5
D_4540_1	46	4	5	4	DIGITIZER HOWTEK 43.5
D_4541_1	49	2	5	4	DIGITIZER HOWTEK 43.5
D_4542_1	57	1	5	4	DIGITIZER HOWTEK 43.5
D_4543_1	37	3	5	4	DIGITIZER HOWTEK 43.5

Test Data

Name	Age	Density	Assessment	Subtlety	Digitizer
A_1112_1	88	3	5	5	DIGITIZER HOWTEK 43.5
A_1114_1	81	4	4	3	DIGITIZER HOWTEK 43.5
A_1122_1	48	2	5	4	DIGITIZER HOWTEK 43.5
A_1127_1	58	4	5	1	DIGITIZER HOWTEK 43.5
A_1140_1	81	2	5	4	DIGITIZER HOWTEK 43.5
A_1147_1	77	3	5	3	DIGITIZER HOWTEK 43.5
A_1149_1	68	2	4	2	DIGITIZER HOWTEK 43.5
A_1155_1	48	4	5	5	DIGITIZER HOWTEK 43.5
A_1168_1	87	4	4	2	DIGITIZER HOWTEK 43.5
A_1169_1	37	1	5	2	DIGITIZER HOWTEK 43.5
A_1171_1	73	1	5	5	DIGITIZER HOWTEK 43.5
A_1207_1	77	2	5	1	DIGITIZER HOWTEK 43.5
A_1211_1	67	3	5	5	DIGITIZER HOWTEK 43.5
A_1228_1	73	3	5	5	DIGITIZER HOWTEK 43.5
A_1233_1	46	4	4	4	DIGITIZER HOWTEK 43.5
A_1234_1	65	3	5	4	DIGITIZER HOWTEK 43.5
A_1237_1	46	4	5	5	DIGITIZER HOWTEK 43.5
A_1247_1	52	2	5	5	DIGITIZER HOWTEK 43.5
A_1258_1	43	3	5	4	DIGITIZER HOWTEK 43.5
A_1401_1	70	4	5	2	DIGITIZER HOWTEK 43.5
A_1416_1	55	4	5	4	DIGITIZER HOWTEK 43.5
A_1468_1	71	1	5	5	DIGITIZER HOWTEK 43.5
A_1485_1	51	3	5	2	DIGITIZER HOWTEK 43.5
A_1504_1	69	4	5	3	DIGITIZER HOWTEK 43.5
A_1510_1	58	4	5	5	DIGITIZER HOWTEK 43.5
A_1573_1	73	2	5	3	DIGITIZER HOWTEK 43.5
A_1577_1	85	3	5	1	DIGITIZER HOWTEK 43.5
A_1618_1	62	3	5	3	DIGITIZER HOWTEK 43.5
A_1628_1	75	2	5	5	DIGITIZER HOWTEK 43.5
A_1658_1	83	3	5	4	DIGITIZER HOWTEK 43.5
A_1669_1	52	1	5	4	DIGITIZER HOWTEK 43.5
A_1673_1	48	4	5	4	DIGITIZER HOWTEK 43.5
A_1674_1	61	4	5	1	DIGITIZER HOWTEK 43.5
A_1804_1	48	4	4	1	DIGITIZER HOWTEK 43.5
A_1821_1	75	3	5	5	DIGITIZER HOWTEK 43.5

A_1827_1	78	3	4	3	DIGITIZER HOWTEK 43.5
A_1892_1	69	3	5	2	DIGITIZER HOWTEK 43.5
A_1906_1	48	2	5	4	DIGITIZER HOWTEK 43.5
A_1985_1	41	4	5	2	DIGITIZER HOWTEK 43.5
A_1999_1	81	2	5	3	DIGITIZER HOWTEK 43.5
D_4544_1	43	2	5	3	DIGITIZER HOWTEK 43.5
D_4545_1	54	3	5	3	DIGITIZER HOWTEK 43.5
D_4546_1	63	2	5	3	DIGITIZER HOWTEK 43.5
D_4547_1	57	2	5	3	DIGITIZER HOWTEK 43.5
D_4551_1	44	1	5	3	DIGITIZER HOWTEK 43.5
D_4552_1	52	3	5	3	DIGITIZER HOWTEK 43.5
D_4553_1	46	3	5	3	DIGITIZER HOWTEK 43.5
D_4555_1	40	3	5	3	DIGITIZER HOWTEK 43.5
D_4557_1	47	4	5	3	DIGITIZER HOWTEK 43.5
D_4558_1	61	2	5	3	DIGITIZER HOWTEK 43.5
D_4559_1	55	3	5	3	DIGITIZER HOWTEK 43.5
D_4560_1	44	3	5	3	DIGITIZER HOWTEK 43.5
D_4561_1	50	3	5	3	DIGITIZER HOWTEK 43.5
D_4562_1	42	2	5	3	DIGITIZER HOWTEK 43.5
D_4563_1	47	2	5	3	DIGITIZER HOWTEK 43.5
D_4565_1	36	3	5	3	DIGITIZER HOWTEK 43.5
D_4566_1	35	4	5	3	DIGITIZER HOWTEK 43.5
D_4567_1	52	2	5	3	DIGITIZER HOWTEK 43.5
D_4570_1	38	2	5	3	DIGITIZER HOWTEK 43.5
D_4571_1	56	3	5	3	DIGITIZER HOWTEK 43.5
D_4572_1	50	1	5	3	DIGITIZER HOWTEK 43.5
D_4574_1	56	2	5	3	DIGITIZER HOWTEK 43.5
D_4575_1	39	2	5	3	DIGITIZER HOWTEK 43.5
D_4576_1	58	1	5	3	DIGITIZER HOWTEK 43.5
D_4577_1	40	2	5	3	DIGITIZER HOWTEK 43.5
D_4582_1	38	4	5	3	DIGITIZER HOWTEK 43.5
D_4583_1	52	2	5	3	DIGITIZER HOWTEK 43.5
D_4584_1	45	3	5	3	DIGITIZER HOWTEK 43.5
D_4585_1	36	4	5	3	DIGITIZER HOWTEK 43.5
D_4587_1	41	3	5	3	DIGITIZER HOWTEK 43.5
D_4588_1	47	2	5	3	DIGITIZER HOWTEK 43.5
D_4589_1	53	2	5	3	DIGITIZER HOWTEK 43.5
D_4590_1	44	3	5	3	DIGITIZER HOWTEK 43.5
D_4591_1	40	4	5	3	DIGITIZER HOWTEK 43.5
D_4592_1	51	2	5	3	DIGITIZER HOWTEK 43.5
D_4593_1	62	2	5	3	DIGITIZER HOWTEK 43.5

D_4594_1	38	3	5	3	DIGITIZER HOWTEK 43.5
D_4595_1	47	3	5	3	DIGITIZER HOWTEK 43.5

References

1. C. C. Aggarwal. Towards systematic design of distance functions for data mining applications. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 9–18, Washington, D.C., August 2003.

2. R. Agrawal, C. Faloutsos, A. Swami, Efficient similarity search in sequence databases, In *Proceedings of the Fourth International Conference on Foundations of Data Organization and Algorithms* (FODO'93), Evanston, IL, in: Lecture Notes in Computer Science, Volume 730, Springer, 1993, pp. 69-84.

3. R. Agrawal, K.I. Lin, H.S. Sawhney, K. Shim, Fast similarity search in the presence of noise, scaling, and translation in time-series databases, In *Proceedings of the 21st International Conference on Very Large Databases* (VLDB'95), Morgan Kaufmann, September 1995, pp. 490-501, Zurich, Switzerland.

4. A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, A. Marchiori, Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment, *Radiology* 228(1), July 2003, pages 265-270.

5. American Cancer Society. *Breast Cancer Facts and Figures 1999-2000*. American Cancer Society, Inc., Atlanta, GA, 1999.

6. J.A. Anderson, 1995, *An Introduction to Neural Networks*, The MIT Press.

7. M. Artae, M. Jogan, and A. Leonardis, "Incremental PCA for on-line visual learning and recognition," In *Proceedings 16th International Conference on Pattern Recognition*, pp. 781-784, volume 3, 2002.

8. S. Arya and T. Malamatos. Linear-size approximate voronoi diagrams. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 147–155, Las Vegas, NV, January 2002.

9. S. Arya, T. Malamatos, and D. M. Mount. Space-efficient approximate voronoi diagrams. In *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing*, pages 721–730, Montr´eal, Que´ebec, Canada, January 2002.

10. S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y.Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM*, 45(6):891–923, November 1998.

11. S. Astley and F.J. Gilbert. Computer-aided detection in mammography. *Clinical Radiology*, 59(5):390–9, May 2004.

12. S. Astley, T. Mistry, C.R.M. Boggis, V. F. Hillier. Should we use humans or a machine to pre-screen mammograms? In *Proceedings of the Sixth International Workshop on Digital Mammography*, Bremen, Germany, pages 476– 480, June 2002.

13. R. E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

14. R. Benetis, C. S. Jensen, G. Karciauskas, and S. Saltenis. Nearest neighbor and reverse nearest neighbor queries for moving objects. In *Proceedings of the International*

Database Engineeing and Applications Symposium, M. A. Nascimento M. T. O¨ zsu and O. R. Za¨iane, eds., pages 44–53, Edmonton, Canada, July 2002.

15. S. Berchtold, C. Bohm, and H.-P. Kriegel. Improving the query performance of high-dimensional index structures by bulk-load operations. In *Advances in Database Technology — EDBT'98, Proceedings of the 1st International Conference on Extending Database Technology*, H.-J Schek, F. Saltor, I. Ramos, and G. Alonso, eds., pages 216–230, Valencia, Spain, March 1998.

16. S. Berchtold, C. Bohm, H.-P. Kriegel, J. Sander, and H. V. Jagadish. Independent quantization: An index compression technique for high-dimensional data spaces. In *Proceedings of the 16th IEEE International Conference on Data Engineering*, pages 577–588, San Diego, CA, February 2000.

17. S. Berchtold, H.P. Kriegel, S3: similarity search in cad database systems, In *Proceedings ACM SIGMOD International Conference on Management of Data*, Tucson, AZ, May 1997, pp. 564-567.

18. K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *Proceedings of the 7th International Conference on Database Theory (ICDT'99)*, C. Beeri and P. Buneman, eds., pages 217–235, Berlin, Germany, January 1999. Also Springer-Verlag Lecture Notes in Computer Science 1540.

19. A.L. Blum and P. Langley, Selection of relevant features and examples in machine learning, *Artificial Intelligence*, volume 97, pp. 245-271, 1997.

20. C. Bohm, B. Braunmuller, F. Krebs, and H.-P. Kriegel. Epsilon grid order: an algorithm for the similarity join on massive high-dimensional data. In *Proceedings of the ACM SIGMOD Conference*, pages 379–390, Santa Barbara, CA, May 2001.

21. J.P. Boissel, M. Cucherat, E. Amsallem, P. Nony, M. Fardeheb, W. Manzi, M.C. Haugh, Getting evidence to prescribers and patients or how to make EBM a reality, *Studies in Health Technology and Informatics,* volume 95, 2003, pages 554-559.

22. T. Bozkaya and M. Ozsoyoglu. Distance-based indexing for high-dimensional metric spaces. In *Proceedings of the ACM SIGMOD Conference*, J. Peckham, ed., pages 357–368, Tucson, AZ, May 1997.

23. S. Brin. Near neighbor search in large metric spaces. In *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB)*, U. Dayal, P. M. D. Gray, and S. Nishio, eds., pages 574–584, Zurich, Switzerland, September 1995.

24. A.A.T. Bui, R.K. Taira, J.D.N. Dionision, D.R. Aberle, S. El-Saden, H. Kangarloo, Evidence-based radiology: requirements for electronic access, *Academic Radiology,* volume 9, number 6, June 2002, pages 662-669.

25. R. Campanini, A. Bazzani, A. Bevilacqua, D. Bollini, D. Dongiovanni, E. Iampieri, N. Lanconelli, A. Riccardi, M. Roffilli, and R. Tazzoli. A novel approach to mass detection in digital mammography based on support vector machines. In *Proceedings of the 6th International Workshop on Digital Mammography*, Bremen, Germany, June 2002.

26. P. Ciaccia and M. Patella. PAC nearest neighbor queries: approximate and controlled search in high dimensional and metric spaces. In *Proceedings of the 16th IEEE*

International Conference on Data Engineering, pages 244–255, San Diego, CA, February 2000.

27. P. Ciaccia, M. Patella, and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB)*, M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld, eds., pages 426–435, Athens, Greece, August 1997.

28. L.P. Clarke, B.Y. Croft, E. Staab, H. Baker, D.C. Sullivan, National Cancer Institute initiative: Lung image database resource for imaging research, *Academic Radiology,* volume 8, number 5, May 2001, pages 447-50.

29. K. Daphne and M. Sahami, Toward optimal feature selection, In *Proceedings 13th International Conference on Machine Learning*, Morgan Kaufmann, pages 284-292, February 1996.

30. M. Dettling and P. Buhlmann. Boosting for tumor classification with gene expression data. *Bioinformatics*, volume 19, number 9, pages 1061–1069, June 2003.

31. J.P. Dittrich and B. Seeger. GESS: a scalable similarity-join algorithm for mining large data sets in high dimensional spaces. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 47–56, San Francisco, California, August 2001.

32. R.O. Duda, P.E. Hart, and D.G Stork, *Pattern Classification*, second ed. John Wiley, 2001, USA.

33. I. El-Naqa, Y. Yang, N.P. Galatsanos, R.M. Nishikawa, and M.N. Wernick. A similarity learning approach to content based image retrieval: application to digital mammography. *IEEE Transactions on Medical Imaging*, 23(10):1233–1244, October 2004.

34. C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, W. Equitz, Efficient and effective querying by image content, *Journal of Intelligent Information Systems*, Volume 3, July 1994, pp. 231-262.

35. C. Faloutsos and K.-I. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the ACM SIGMOD Conference*, pages 163–174, San Jose, CA, May 1995.

36. C. Faloutsos, M. Ranganathan, Y. Manolopoulos, Fast subsequence matching in time-series databases, In *Proceedings ACM SIGMOD International Conference on Management of Data*, Minneapolis, MN, 1994, pp. 419-429.

37. W. Fan, M.D. Gordon, and P. Pathak, Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison, *Decision Support Systems*, volume 40, issue 2, pp. 213-233, August 2005.

38. H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, and A. El Abbadi. Vector approximation based indexing for non-uniform high dimensional data sets. In *Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM)*, pages 202–209, McLean, VA, November 2000.

39. R.J. Ferrari, R.M. Rangayyan, J.E.L. Desautels, A.F. Frere. Analysis of asymmetry in mammograms via directional filtering with Gabor wavelets. *IEEE Transactions on Medical Imaging*, 20(9):953-64 Sept 2001.

40. BD Fornage, JG Lorigan, and E. Andry. Fibroadenoma of the breast: Us appearance. *Radiology*, 172(3):671–675, Sept 1989.

41. T.W. Freer and M.J. Ulissey. Screening mammography with computer-aided detection: Prospective Study of 12,860 Patients in a Community Breast Center. *Radiology*, 220:781–786, April 2001.

42. J.E. Gary, R. Mehrotra, Similar shape retrieval using a structural feature index, *Information Systems*, Volume 18, No. 7, October 1993, pp. 525-537.

43. N. Gershenfeld. *The Nature of Mathematical Modeling*. Cambridge University Press, Cambridge, United Kingdom, 1999.

44. A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB)*, M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik, and M. L. Brodie, eds., pages 518–529, Edinburgh, Scotland, September 1999.

45. J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, Washington, DC, volume 1, pages 487–493, October 2003.

46. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S Lander,

Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286, pages 531-537, Oct 1999.

47. G. H. Golub and C. F. van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

48. I. Gondra and D.R. Heisterkamp. Learning in region-based image retrieval with generalized support vector machines. In *Proceedings of the Computer Vision and Pattern Recognition*, page 149, June 2004.

49. V. Gudivada and V. Raghavan. Design and evaluation of algorithms for image retrieval by spatial similarity. *ACM Transactions on Information Systems*, 13(2):115–144, April 1995.

50. A. Guimond and G. Subsol. Automatic MRI database exploration and applications. *Pattern Recognition and Artificial Intelligence*, 11(8):1345–1365, 1997.

51. R.W. Hamming. Error-detecting and error-correcting codes. *Bell System Technical Journal*, 29(2):147–160, April 1950.

52. S. Har-Peled. A practical approach for computing the diameter of a point set. In *Proceedings of the 17th Annual Symposium on Computational Geometry*, pages 177–186, Medford, MA, June 2001.

53. M.D. Heath and K.W. Bowyer. Mass detection by relative image intensity. In *The Proceedings of the 5th International Conference on Digital Mammography*, Toronto, Canada, Medical Physics Publishing, Madison, WI, June 2000.

54. M.D. Heath, K.W. Bowyer, D. Kopans et al. Current status of the digital database for screening mammography. In *Digital Mammography*, Kluwer Academic Publishers, pages 457–60, 1998.

55. A. Hinneburg, C. C. Aggarwal, and D. A. Keim. What is the nearest neighbor in high dimensional spaces. In *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB)*, A. El Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, and K.-Y. Whang, eds., pages 506–515, Cairo, Egypt, September 2000.

56. K. Hiraoka, K. Hidai, M. Hamahira, H. Mizoguchi, T. Mishima, and S. Yoshizawa, Successive learning of linear discriminant analysis: sanger-type algorithm, In *Proceedings 14th International Conference Pattern Recognition*, pp. 2664-2667, volume 2, 2000.

57. G. R. Hjaltason and H. Samet. Incremental distance join algorithms for spatial databases. In *Proceeding of the ACM SIGMOD Conference*, L. Hass and A. Tiwary, eds., pages 237–248, Seattle, WA, June 1998.

58. G. R. Hjaltason and H. Samet. Index-driven similarity search in metric spaces. *ACM Transactions on Database Systems*, 28(4):517–580, December 2003.

59. G. R. Hjaltason and H. Samet. Properties of embedding methods for similarity searching in metric spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):530–549, May 2003. Also University of Maryland Computer Science TR-4102.

60. R. Hoch, Using IR techniques for text classification in document analysis, In *Proceedings 17th Annual International ACM SIGIR Conference Research and Development in Information Retrieval*, pp. 31-40, 1994.

61. S.L. Horowitz and T. Pavlidis. Picture segmentation by a tree traversal algorithm. *Journal of the ACM*, 23(2):368–388, April 1976.

62. M. E. Houle. SASH: a spatial approximation hierarchy for similarity search. *Technical Report RT0517*, IBM Tokyo Research Lab Technical Report, Tokyo, Japan, March 2003.

63. M. E. Houle and J. Sakuma. Fast approximate similarity search in extremely high-dimensional data sets. In *Proceedings of the 21st IEEE International Conference on Data Engineering*, pages 619–630, Tokyo, Japan, April 2005.

64. G. Hristescu and M. Farach-Colton. Cluster-preserving embedding of proteins. Technical report, Rutgers University, Piscataway, NJ, 1999.

65. D. P. Huttenlocher, D. A. Klanderman, and W. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, September 1993.

66. P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on the Theory of Computing*, pages 604–613, Dallas, May 1998.

67. H.V. Jagadish, A Retrieval Technique for Similar Shapes, In *Proceedings ACM SIGMOD International Conference on Management of Data*, Denver, CO, 1991, pp. 208-217.

68. T. Joachims, “Making large-scale support vector machine learning practical”, In the *Proceedings of the International Conference on Machine Learning* (ICML’99), 200-209 (1999).

69. D. Kalashnikov and S. Prabhakar. Similarity joins for low- and high- dimensional data. In *Proceedings of the 8th International Conference on Database Systems for Advanced Applications* (DASFAA’03), pages 7–16, Kyoto, Japan, March 2003.

70. B.L. Kalman, S.C. Kwasny, and W.R. Reinus. Diagnostic screening of digital mammograms using wavelets and neural networksto extract structure. *Technical Report Technical Report 98-20*, Washington University, 1998.

71. D. Keysers, J. Dahmen, H. Ney, B. B. Wein, T. M. Lehmann, A statistical framework for model-based image retrieval in medical applications, *Journal of Electronic Imaging* 12 (1) (2003) 59-68.

72. F. Korn and S. Muthukrishnan. Influence sets based on reverse nearest neighbor queries. In *Proceedings of the ACM SIGMOD Conference*, W. Chen, J. Naughton, and P. A. Bernstein, eds., pages 201–212, Dallas, May 2000.

73. C. LeBozec, M.C. Jaulent, E. Zapletal, P. Degoulet, Unified modeling language and design of a case-based retrieval system in medical imaging, in *Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA*), Nashville, TN, USA, 1998, 887-891.

74. T. M. Lehmann, M. O. Guld, C. Thies, B. Fischer, M. Keysers, D. Kohnen, H. Schubert, B.B. Wein, Content-based image retrieval in medical applications for picture

archiving and communication systems, in *Medical Imaging*, Volume 5033 of SPIE Proceedings, San Diego, California, USA, 2003, 109-117.

75. H.U. Lemke, PACS developments in Europe, *Computerized Medical Imaging and Graphics* 27 (2002) 111-120.

76. V. A. Levenshtein. Binary codes capable of correcting deletions, insertion, and reversals. *Cybernetics and Control Theory*, 10(8):707–710, 1966.

77. J. Linda, W. Burhenne, S.A. Wood, C.J. D'Orsi, S.A. Feig, D.B. Kopans, K.F. O'Shaughnessy, E.A. Sickles, L. Tabar, C.J. Vyborny, and R.A. Castellino. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology*, 215(554–562), 2000.

78. D.D. Lewis, Feature selection and feature extraction for text categorization, in *Proceedings of the Workshop on Speech and Natural Language*, pp. 212-217, 1992.

79. S. Lui, C.F. Babbs, and E.J. Delp. Multiresolution detection of spiculated lesions in digital mammograms. *IEEE Transactions on Image Processing*, 6:874–884, June 2001.

80. H. Li, T. Jiang, and K. Zhang, Efficient and robust feature extraction by maximum margin criterion, in *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp. 97-104, 2004.

81. Y. Li, L. Xu, J. Morphett, and R. Jacobs, An integrated algorithm of incremental and robust PCA, in *Proceedings of the International Conference on Image Processing*, pp. 245-248, 2003.

82. J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281–297, 1967.

83. P. Miller and S. Astley. Detection of breast asymmetry using anatomical features. In *Proceedings of the International Society for Optical Engineering Conference on Biomedical Image Processing and Biomedical Visualization*, 1905:433-442, 1993.

84. G. Navarro. Searching in metric spaces by spatial approximation. *VLDB Journal*, 11(1):28–46, 2002.

85. S.A. Nene and S.K. Nayar. A simple algorithm for nearest-neighbor search in high dimensions. *Pattern Analysis and Machine Intelligence*, 19(9):989–1003, September 1997.

86. E. Oja, Subspace methods of pattern recognition, *Pattern Recognition and Image Processing Series*, volume 6, 1983.

87. S.C. Orphanoudakis, C.E. Chronaki, S. Kostomanolakis, I2Cnet: A system for the indexing, storage, and retrieval of medical images by content, *Medical Informatics* 19 (2) (1994) 109-122.

88. S.C. Orphanoudakis, C.E. Chronaki, D. Vamvaka, I2Cnet: Content based similarity search in geographically distributed repositories of medical images, *Computerized Medical Imaging and Graphics* 20 (4) (1996) 93-207.

89. W. W. Peterson. Addressing for random access storage. *IBM Journal of Research and Development*, 1(2):130–146, April 1957.

90. M.P. Popli. Pictorial essay: sonographic differentiation of solid breast lesions. *Indian Journal of Radiological Imaging*, 12(2):275–279, 2002.

91. G. Rahbar, A.C. Sie, G.C. Hansen, J.S. Prince, M.L. Melany, H.E. Reynolds, V.P. Jackson, J.W. Sayre, and L.W. Bassett. Benign versus malignant solid breast masses: US differentiation. *Radiology*, 213:889–894, 1999.

92. B. Revet, *DICOM Cook Book for Implementations in Modalities*, Philips Medical Systems, Eindhoven, Netherlands, 1997.

93. Y. Rubner, C. Tomasi, and L. Guibas. Adaptive color image embedding for database navigation. In *Proceedings of the Asian Conference on Computer Vision*, January 1998.

94. Y. Rubner and C. Tomasi. Texture metrics. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 4601-4607, October 1998.

95. P. Sajda, C. Spense, and L. Parra. Capturing contextual dependencies in medical imagery using hierarchical multi-scale models. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pages 165–168, 2002.

96. M. Sallam and K.W. Bowyer. Registering time-sequences of mammograms using a two-dimensional unwarping technique. In *Digital Mammography*, K.Doi, M.L. Giger, R.M. Nishikawa, and R.A. Schmidt, eds., Amsterdam, the Netherlands, Elsevier, 1996, pages 291–296.

97. H. Samet. *Foundations of Multidimensional and Metric Data Structures*. Morgan-Kaufmann, San Francisco, CA, 2006.

98. M.P. Sampat and A.C. Bovik**,** Detection of spiculated lesions in mammograms, in *Proceedings of the 25th Annual International IEEE Conference on Engineering in Medicine and Biology Society*, (1) (2003) 810-813.

99. R. F. Santos Filho, A. J. M. Traina, C. Traina Jr., and C. Faloutsos. Similarity search without tears: the OMNI family of all-purpose access methods. In *Proceedings of the 17th IEEE International Conference on Data Engineering*, pages 623–630, Heidelberg, Germany, April 2001.

100. T. B. Sebastian and B. B. Kimia. Metric-based shape retrieval in large databases. In *Proceedings of the 16th International Conference on Pattern Recognition*, R. Kasturi, D. Laurendau, and C. Suen, eds., volume 3, pages 291–296, Quebec City, Canada, August 2002.

101. H. Shin, B. Moon, and S. Lee. Adaptive multi-stage distance join processing. In *Proceedings of the ACM SIGMOD Conference*, W. Chen, J. Naughton, and P. A. Bernstein, eds., pages 343–354, Dallas, May 2000.

102. C.R. Shyu, C.E. Brodley, A.C. Kak, A. Kosaka, A.M. Aisen, L.S. Broderick, ASSERT: A physician-in-the-loop content-based retrieval system for HRCT image databases, *Computer Vision and Image Understanding* 75 (1/2) (1999) 111-132.

103. A. Soffer and H. Samet. Pictorial queries by image similarity. In *Proceedings of the 13th International Conference onPattern Recognition*, volume 3, pages 114–119, Vienna, Austria, August 1996.

104. I. Stanoi, D. Agrawal, and A. El Abbadi. Reverse nearest neighbor queries for dynamic databases. In *Proceedings ACM SIGMOD Workshop on Research Issues in*

Data Mining and Knowledge Discovery, D. Gunopulos and R. Rastogi, eds., pages 44–53, Dallas, May 2000.

105. I. Stanoi, M. Riedewald, D. Agrawal, and A. El Abbadi. Discovery of influence sets in frequently updated databases. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB)*, P. M. G. Apers, P. Atzeni, S. Ceri, S. Paraboschi, K. Ramamohanarao, and R. T. Snodgrass, eds., pages 99–108, Rome, Italy, September 2001.

106. A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. *Radiology*, 196:123–134, 1995.

107. H. A. Swett and P. L. Miller. Icon: a computer-based approach to differential diagnosis in radiology. *Radiology*, 163:555–558, 1987.

108. J. K. Uhlmann. Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters*, 40(4):175–179, November 1991.

109. M. W. Vannier, E. V. Staab, L. C. Clarke, Medical image archives - present and future, in *Proceedings of the International Conference on Computer-Assisted Radiology and Surgery*, Paris, France, 2000, 565-570.

110. N. Vujovic and D. Brzakovic. Establishing the correspondence between control points in pairs of Mammographic images. *IEEE Transactions on Image Processing*, 6(10):1388–1399, October 1997.

111. A.R. Webb, *Statistical Pattern Recognition*, second edition, John Wiley, 2002.

112. R. Weber, H. J. Schek, and S. Blott. A quantitative analysis and performance study for similarity search methods in high-dimensional spaces. In *Proceedings of the*

24th International Conference on Very Large Data Bases (VLDB), A. Gupta, O. Shmueli, and J. Widom, eds., pages 194–205, New York, August 1998.

113. J. Weng, Y. Zhang, and W.-S. Hwang, Candid covariance-free incremental principal component analysis, in *IEEE Transactions Pattern Analysis and Machine Intelligence*, volume 25, pp. 1034-1040, 2003.

114. M. A. Wirth, C. Choi, and A. Jennings. A nonrigid-body approach to matching mammograms. In *Proceedings of the 7th International Conference on Image Processing and its Applications*, volume 2, pages 484–8, Manchester, UK, 1999.

115. M. A. Wirth, J. Narhan, and D. Gray. Nonrigid mammogram registration using mutual information. In *Proceedings of SPIE Medical Imaging: Image Processing*, San Diego, CA, 4684:562–573, February 2002.

116. J. Yan, B.Y. Zhang, S. C. Yan, Z. Chen, W. G. Fan, Q. Yang, W. Y. Ma, and Q. S. Cheng, IMMC: incremental maximum, marginal criterion, In *Proceedings of the 10th ACM SIGKDD International Conference Knowledge Discovery and Data Mining*, pp. 725-730, 2004.

117. J. Yan, N. Liu, B. Y. Zhang, S. C. Yan, Q. S. Cheng, W. G. Fan, Z. Chen, W. S. Xi, and W. Y. Ma, OCFS: orthogonal centroid feature selection, In *Proceedings 28th Annual International ACM SIGIR Conference Research and Development in Information Retrieval*, 2005.

118. C. Yang and K. I. Lin. An index structure for efficient reverse nearest neighbor queries. In *Proceedings of the 17th IEEE International Conference on Data Engineering*, pages 485–492, Heidelberg, Germany, April 2001.

119. C. Yang and K. I. Lin. An index structure for improving nearest closest pairs and related join queries in spatial databases. In *Proceedings of the International Database Engineeing and Applications Symposium*, M. A. Nascimento M. T. O¨ zsu and O. R. Za¨iane, eds., pages 140–149, Edmonton, Canada, July 2002.

120. Y. Yang and J. O. Pedersen, A comparative study on feature selection in text categorization, In *Proceedings of the 14th International Conference Machine Learning*, pp. 412-420, 1997.

121. Jianhua Yao and R. Taylor. Assessing accuracy factors in deformable 2D/3D medical image registration using a statistical pelvis model. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2:1329–1334, October 2003.

122. P. N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 311–321, Austin, TX, January 1993.

123. F.F. Yin, M.L. Giger, K. Doi, C.E. Metz, C.J. Vyborny, and R.A. Schmidt. Computerized detection of masses in digital mammograms: analysis of bilateral subtraction images. *Medical Physics*, 18:955–63.