Mining Medical Images

R. Bharat Rao, Glenn Fung, Balaji Krishnapuram, Jinbo Bi, Murat Dundar, Vikas Raykar, Shipeng Yu, Sriram Krishnan, Xiang Zhou, Arun Krishnan, Marcos Salganicoff, Luca Bogoni, Matthias Wolf, Anna Jerebko, Jonathan Stoeckel

Image and Knowledge Management-CAD and Knowledge Solutions (IKM-CKS)

Siemens Medical Solutions USA, Inc., 51 Valley Stream Parkway, Malvern, PA-19355

bharat.rao@siemens.com

ABSTRACT

Advances in medical technology have greatly increased information density for imaging studies. This may result from increased spatial resolution facilitating greater anatomical detail, increased contrast resolution allowing evaluation of more subtle structures than previously possible, or increased temporal image acquisition rate.

However, such technological advances, while potentially improving the diagnostic benefits of a study, may result in "data overload" while processing this information. This often manifests as increased total study time, defined as the combination of acquisition, processing and interpretation times; even more critically, the vast increase in data does not always translate to improved diagnosis/treatment selection. This paper describes a related series of clinically motivated data mining products that extract the key, actionable information from the vast amount of imaging data in order to ensure an improvement in patient care (via more accurate/early diagnosis) and a simultaneous reduction in total study time. *Several thousand units of the products described in this paper have been commercially deployed in hospitals around the world since 2004*[1].

While each application targets a specific clinical task, they share the common methodology of transforming raw imaging data, through knowledge-based data mining algorithms, into clinically relevant information. This enables users to spend less time interacting with an image volume to extract the clinical information it contains, while supporting improved diagnostic accuracy. Although image processing plays an equally critical role in these software, this paper focuses primarily on the data mining challenges involved in developing commercial products.

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Computer Aided Diagnosis, data mining, image processing.

1. INTRODUCTION

The invention of the X-ray by William Röntgen in 1895 (Stanton 1896) revolutionized medicine. Thanks to the science of *in-vivo* imaging, doctors were able to look inside a patient's body without resorting to dangerous procedures – the term "exploratory surgery" has all but vanished from our lexicon today.

The fundamental value of the X-ray remains the same today, as it was over 100 years ago – different structures (bone, cartilage, tissue, tumor, metal, etc.) can be identified based on their ability to block the X-ray/Röntgen beam. The initial uses of *in-vivo* imaging were to diagnose broken bones and locate foreign objects, such as, bullets, inside a patient's body. As imaging techniques and resolutions improved, physicians began to use these methods to locate medical abnormalities (e.g., cancer), both for planning surgery and for diagnosing the disease. The state-of-the-art of medical imaging improved to the point that it soon required its own specialty, namely, radiologists, who were skilled at interpreting these increasingly complex images.

The introduction of computers and the subsequent invention of computed tomography (CT) (Ambrose and Hounsfield 1973) in the 1970s created another paradigm – that of 3-dimensional imaging. X-ray beams were used to compute a 3-d image of the inside of the body from several 2-d X-ray images taken around a single axis of rotation. Radiologists were now not only able to detect subtle variations of structures in the body; they were now able to locate them within a fixed frame of reference. Early CT's generated images or slices orthogonal to the long axis of the body, modern scanners allow this volume of data to be reformatted in various planes or even visualized as volumetric (3D) representations of structures.

It is natural to ask whether the improved resolution of medical imaging has clinical value. Consider the use of CTs to diagnose lung cancer. Lung cancer is the most commonly diagnosed cancer worldwide, accounting for 1.2 million new cases annually. Lung cancer is an exceptionally deadly disease: 6 out of 10 people will die within one year of being diagnosed. The expected 5-year survival rate for all patients with a diagnosis of lung cancer is merely 15%. In the United States, lung cancer is the leading cause of cancer death for both men and women and costs almost $10 billion to treat annually.

However, lung cancer prognosis varies greatly depending on how early the disease is diagnosed; as with all cancers, *early detection* provides the best prognosis. At one extreme are the patients diagnosed with distant tumors (that have spread far from the lung, Stage IV patients), for whom the 5-year survival rate is just 2%. The prognosis of early stage lung cancer patients (Stage I) is more optimistic with a

[1] These products are commercially available from Siemens Medical Solutions USA, Inc.

mean 5 year survival rate of about 49%. This follows logically since early detection implies the cancer is found when it is still relatively small in size (thus, fewer cancer cells in the body) and localized (before it has spread). Therefore, many treatment options are viable (surgery, radiotherapy, chemotherapy) if it is detected early.

In order to identify and characterize minute lung nodules the resolution of the image must be improved. The recent development of multi-detector computed tomography (MDCT) scanners has made it feasible to detect lung cancer at very early stages, and the number of lung nodules routinely identified in clinical practice is steadily growing. The key factor in CT is the slice thickness (the distance between two axial cross-sectional X-rays)–smaller slice thickness means increased resolution. Today's MDCT's are capable of locating lung nodules that are 2-8mm in size; cancers found at this early stage have excellent prognosis. Despite these technologies, only 24% of lung cancer cases are diagnosed at an early stage (Jemal, et al. 2007). Many potentially clinically significant lesions remain undetected.

One contributing factor could be the explosion of MDCT imaging data: just 8 years ago, the 2-slice CT could acquire 41 axial images of the thorax in a 30-second scan (single breath hold); the state-of-the-art 64-slice dual-source CT acquires up to 3,687 axial images in 30 seconds for each patient. Figure 1 illustrates two such images for a single patient, and each image must then be carefully examined by a radiologist to identify which of the marks on the image correspond to normal structures (air passage), benign tumors, lung diseases other than cancer, and early-stage lung cancer. Despite the exponential increase in data in a few years, radiologists have roughly the same case load (or in some cases greater) than was the case 20 years ago when they examined a handful of images per patient.

1.1 Mining Medical Images

There is a growing consensus among clinical experts that the use of computer-aided detection/diagnosis (CAD) software can improve the performance of the radiologist. The proposed workflow is to use CAD as a second reader (i.e., in conjunction with the radiologist) – the radiologist first performs an interpretation of the image as usual, and then runs the CAD algorithm (typically a set of image processing algorithms followed by a classifier), and highlights structures identified by the CAD algorithm as being of interest to the radiologist. The radiologist examines these marks and concludes the interpretation. Figure 1 shows super-imposed CAD marks on the images. Clinical studies have shown that the use of CAD software not only offers the potential to improve the detection and recognition performance of a radiologist, but also to reduce mistakes related to misinterpretation (Armato-III, Giger and Mac Mahon 2001, Naidich, Ko and Stoechek 2004).

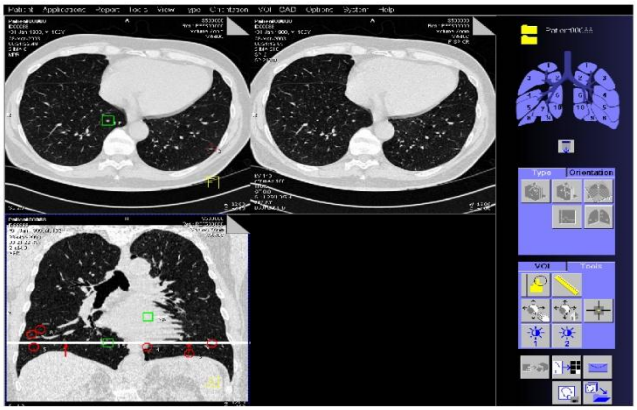

Figure 1 Suspicious regions highlighted on a Lung CT.

The principal value of CAD is determined not by its stand-alone performance, but rather by carefully measuring the incremental value of CAD in normal clinical practice, such as the number of additional lesions detected using CAD. Secondly, CAD systems must not have a negative impact on patient management (for instance, false positives which cause the radiologist to recommend unnecessary biopsies and potentially dangerous follow-ups).

This explosion in data is not confined to CT alone. The invention of the CT was rapidly followed by the development of 3-d magnetic resonance imaging (MRI). MRI uses a powerful magnetic field to align the water molecules in the body, and thus provides much greater contrast between the different soft tissues of the body than CT. Positron emission tomography (PET) and Single photon emission computed tomography (SPECT) use radioactive isotopes to provide functional imaging. Recently, medicine has been moving towards fusion of these different imaging modalities to combine functional and structural properties in a single image. As should be obvious, the ability to identify and characterize increasingly minute structures and subtle variations in bodily function in 3-d images has resulted in an enormous explosion in the amount of data that must be processed by a radiologist. It is estimated that in a few years, medical images will constitute fully 25% of all the data stored electronically.

This vast data store also provides additional opportunities for data mining. CAD algorithms have the ability to automatically extract features, quantify various lesions and bodily structures, and create features than can be subsequently mined to discover new knowledge. This new knowledge can be further fed back into medicine as CAD progresses from detecting abnormal structures, to characterizing structures (identifies structures of interest, and also indicating whether they are malignant or not). Another area of interest is the use of CAD for change detection – for instance, to automatically measure tumors from images taken at different point in times and determine if the tumor size has changed. Such methods can be used both for diagnosis (malignant tumors grow quickly) &

therapy monitoring (is the tumor shrinking with the treatment). The discussion of CAD for change detection & for characterizing structures is beyond the scope of this paper. We shall restrict our attention to practical applications that are clinically deployed today.

Medical image mining products can be useful even if they don't go all the way to computer aided detection. Sometimes it is sufficient to analyze images & quantify key features that are known to be highly diagnostic. Consider the automatic quantification of ultrasound images. So far, all the modalities we have discussed take a snapshot of the body at a particular instant. Cardiac ultrasound captures the very fast motion of the heart, as a results we have an added dimension of time to our data. We describe mining software that tracks the motion of the heart and automatically measures key clinical variables (ejection fractions) that characterize the function of the heart.

1.2 Clinical Trials

Commercial products for mining medical images need to be rigorously validated in clinical trials before they are cleared for sale by national regulatory bodies like the Food & Drug Administration (FDA). All the products described in this paper have been validated in clinical trials, and are sold internationally. Some of the CAD & image quantification products are cleared for sale in the US, and others are still in the process of obtaining regulatory approval (they are sold in other countries). More details will be provided for each product in Section 4.

The rest of the paper is organized as follows: Section 2 provides the clinical motivations for the image-mining systems described in this paper. Section 3 describes some of the original research in data-mining and machine learning that was necessary to develop these systems with a clinically acceptable level of accuracy. Section 4 summarizes some of the results obtained from clinical validation studies. Section 5 concludes the paper by summarizing the key lessons learnt while developing such high impact data mining applications.

2. CLINICAL MOTIVATION

In this section we will describe some of the most commonly diagnosed cancers with some background information and clinical motivation for CAD software.

2.1 Lung cancer

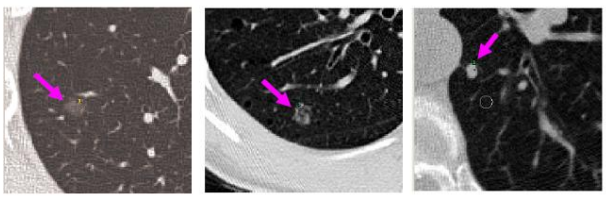

Figure 2 Different nodule morphologies in the lung (from left to right)--Ground glass opacity nodule (GGN), Part-solid nodule (PSN), and Solid nodule (SN).

Lung cancer is the most commonly diagnosed cancer worldwide, accounting for 1.2 million new cases annually. Lung cancer is an exceptionally deadly disease: 6 out of 10 people will die within one year of being diagnosed. The expected 5-year survival rate for all patients with a diagnosis of lung cancer is merely 15%, compared to 65% for colon, 89% for breast, and 99.9% for prostate cancer.

For lung cancer CAD systems are developed to identify suspicious regions called nodules (which are known to be precursors of cancer) in CT scans of the lung. Clinically a solid nodule is defined as an area of increased opacity more than 5mm in diameter which completely obscures underlying vascular marking. Translating this definition into image features and data mining algorithms is the key challenge. While it is universally acknowledged that solid nodules are precursors for lung cancer, recently there has been increased interest in detecting what are known as part-solid nodules (PSN) & ground-glass opacities (GGN). A GGN is defined as an area of a slight, homogenous increase in density, which did not obscure underlying bronchial and vascular markings. GGNs are known to be extremely hard to detect. Several studies (Suzuki, et al. 2006) have pointed out that they are an indicator of early cancer.

One important factor when designing CAD systems for mining lung images is the relative difficulty in obtaining ground truth for lung cancer. Whereas, for example, in breast cancer virtually all suspicious lesions are routinely biopsied (providing definitive histological ground truth), a lung biopsy is a dangerous procedure, with a 2% risk of serious complications (including death). It makes obtaining definitive lung cancer ground truth infeasible, particularly for patients being evaluated for early signs of lung cancer. So very often CAD systems are built using image annotations from multiple expert radiologists.

2.2 Breast cancer

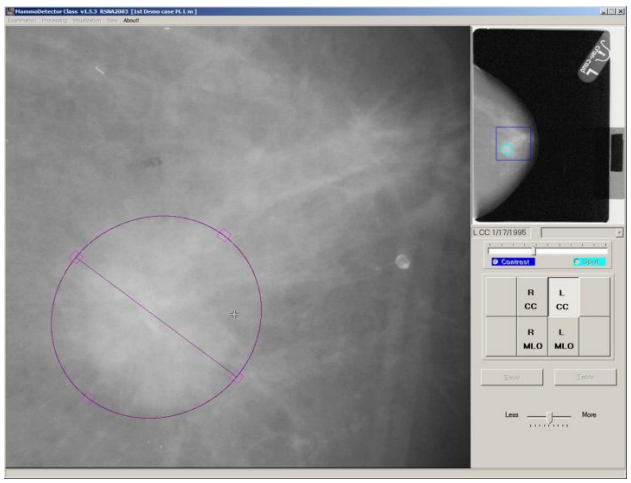

Figure 3 A typical malignant mass on a mammogram.

Breast cancer is the second most common form of cancer in women, after non-melanoma skin cancer (Group 2009). Breast cancer is the number one cause of cancer death in Hispanic women. It is the second most common cause of cancer death in white, black, Asian/Pacific Islander, and American Indian/Alaska Native women. In 2005 alone 186,467 women and 1,764 men were diagnosed with breast cancer; 41,116 women and 375 men died from the disease.

Breast cancer is an abnormal growth of the cell the normally line the ducts and the lobules. Figure 3 shows a typical abnormal growth called mass on a digital mammogram. X-ray Mammography is now accepted as a valid method for breast cancer screening, after many years in which its effectiveness was questioned. Current guidelines recommend screening mammography every year for women, beginning at age 40. CAD systems search for abnormal areas of density, mass, or calcification in a digitized mammographic image. These abnormal areas generally indicate the presence of cancer. The CAD system highlights these areas on the images, alerting the radiologist to the need for a further diagnostic imaging or a biopsy.

2.3 Colon cancer

Colorectal cancer (CRC) is the third most common cancer in both men and women. It is estimated that in 2004, nearly 147,000 cases of colon and rectal cancer will be diagnosed in the USA, and more than 56 730 people would die from colon cancer, accounting for approximately 11% of all cancer deaths. Early detection of colon cancer is the key to reducing the 5-year survival rate. In particular, since it is known that in over 90% of cases the progression stage for colon cancer is from local (polyp adenomas) to advanced stages (colorectal cancer), it is critical that major efforts be devoted to screening of colon cancer and removal of lesions (polyps) when still in a early stage of the disease.

Colorectal polyps are small colonic findings that may develop into cancer at a later stage (See Figure 4). Screening of patients and early detection of polyps via Optical Colonoscopy (OC) has proved to be efficient as the mortality rate from colon cancer is currently decreasing despite an aging population. CT Colonoscopy (CTC), also known as Virtual Colonoscopy (VC) is an increasingly popular alternative to standard OC. In VC, a volumetric CT scan of the distended colon is reviewed by the physician by looking at 2D slices and/or using a virtual fly-through in the computer-rendered colon, searching for polyps. Interest in VC is increasing due to its many advantages over OC (better patient acceptance, lower morbidity, possibility of extra-colonic findings, etc.), with only a small penalty on sensitivity if the reader is a trained radiologist.

Polyp Enhanced Viewing (PEV) systems exploit the full 3-D volume of the colon and use specific image processing & feature calculation algorithms to boost radiologists' sensitivity (Bogoni, et al. 2005) while detecting polyps.

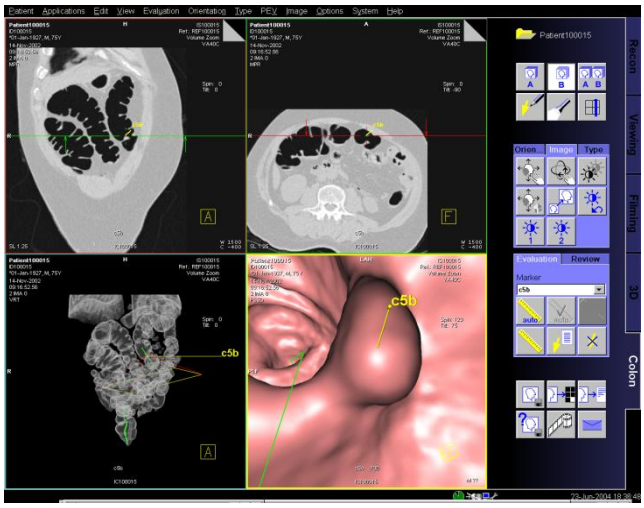

Figure 4 CT scan with enhanced visualization of a polyp in the colon.

2.4 Pulmonary Embolism

Pulmonary Embolism (PE) is a sudden blockage in a pulmonary artery caused by an embolus that is formed in one part of the body and travels to the lungs in the bloodstream through the heart. PE is the third most common cause of death in the US with at least 600,000 cases occurring annually. It causes death in about one-third of the cases, that is, approximately 200,000 deaths annually. Most of the patients who die do so within 30 to 60 minutes after symptoms start; many cases are seen in the emergency department. We developed a fast yet effective approach for computer aided detection of pulmonary embolism (PE) in CT pulmonary angiography (CTPA). Our research has been motivated by the lethal, emergent nature of PE and the limited accuracy and efficiency of manual interpretation of CTPA studies.

Treatment with anti-clotting medications is highly effective, but sometimes can lead to subsequent hemorrhage and bleeding; therefore, the anti-clotting medications should be only given to those who really need. This demands a very high specificity in PE diagnosis. Unfortunately, PE is among the most difficult conditions to diagnose because its primary symptoms are vague, non-specific, and may have a variety of other causes, making it hard to separate out the critically ill patients who suffer from PE. PE cases are missed in diagnosis more than 400,000 times in the US each year. If pulmonary embolism can be diagnosed and appropriate therapy started, the mortality can be reduced from approximately 30 percent to less than ten percent; roughly 100,000 patients die who would have survived with the proper and prompt diagnosis and treatment. A major clinical challenge, particularly in an emergency room scenario, is to quickly and correctly diagnose patients with PE and then send them on to treatment. A prompt and accurate diagnosis of PE is the key to survival.

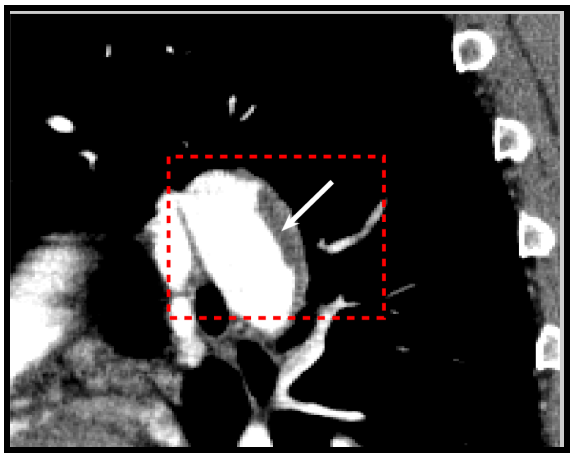

Figure 5 Highlighted Pulmonary embolism on the Lung CT

2.5 Cardiac

Cardiovascular Disease (CVD) is a global epidemic that is the leading cause of death worldwide (17 million deaths per year). In the United States, CVD accounted for 38% of all deaths in 2002 [7] and was the primary or contributing cause in 60% of all deaths. *Coronary Heart Disease* (CHD) accounts for more than half the CVD deaths (roughly 7.2 mil. deaths worldwide every year, and 1 of every 5 deaths in the US), and is the *single* largest killer in the world.

One important tool required for assessing the condition of the heart is the automatic assessment of the left ventricular ejection fraction (EF). EF is a relevant criterion for pharmacologic, defibrillator, and resynchronization therapy, therefore, being able to automatically calculate a robust EF measure is of interest to improve clinical workflow. Currently, the method widely used in clinical practice consists of a subjective visual estimation of EF, even though it is prone to significant variability.

The reliable delineation of the left ventricle (LV) for robust quantification requires years of clinical experience and expertise by echocardiographers and sonographers. Even with acceptable image quality, issues such as trabeculations of the myocardium, fast-moving valves, chordi and papillary muscles, all contribute to the challenges associated with delineation of the LV. Technical issues, such as the fact that a 2D plane is acquired on a twisting 3D object, make this problem even more difficult. Limited success has been achieved in automated quantification based on LV delineation with methods that simply look for a border between black and white structures in an image.

3. DATA MINING CHALLENGES

This section describes the intuition & key ideas motivating some of the data mining algorithms developed by our group to improve medical image processing applications. [2]

[2] While our group has written over 150 papers on this topic, for brevity we are only describing a selected subset.

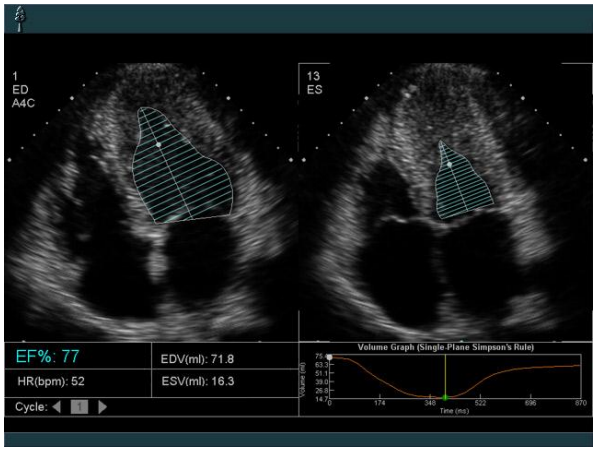

Figure 6 Automated measurement of ejection fraction

The goal is to either extract key quantitative features summarizing vast volumes of data, or to enhance the visualization of potentially malignant nodules, tumors, emboli, or lesions in medical images like CT scan, X-ray, MRI etc. Most medical image mining algorithms operate in a sequence of three stages (see Figure 7):

1. **Candidate generation:** This stage identifies suspicious unhealthy regions of interest (called candidates) from a medical image. This step is based on image processing algorithms which try to search for regions in the image which look like the particular anomaly/lesion. While this step can detect most of the anomalies (around 90-100% sensitivity), the number of false positives will be extremely high (on the order of 60-300 false positives/image).

2. **Feature extraction:** This step involves the computation of a set of descriptive morphological or texture features for each of the candidates using advanced image processing techniques.

3. **Classification:** This stage differentiates candidates that are true lesions from the rest of the candidates based on candidate feature vectors. The goal of the classifier is to reduce the number of false positives (to 2-5 false positives/patient, image) without appreciable decrease in the sensitivity.

Image quantification & enhanced visualization algorithms do not necessarily include a classifier, but they often use image processing & pattern recognition algorithms for candidate generation & feature extraction. CAD systems use all three stages described above & aid the radiologist by marking the location of likely anomalies on a medical image. The radiologist then makes a decision whether to conduct a biopsy or other follow-ups. Since biopsies are expensive and invasive, CAD systems demand as few false positives (2-5 false positives/patient, image) as possible while at the same time achieving high sensitivity (>80%).

While all the three stages are equally important, in this article we will focus on the data mining challenges & will not discuss image processing algorithms.

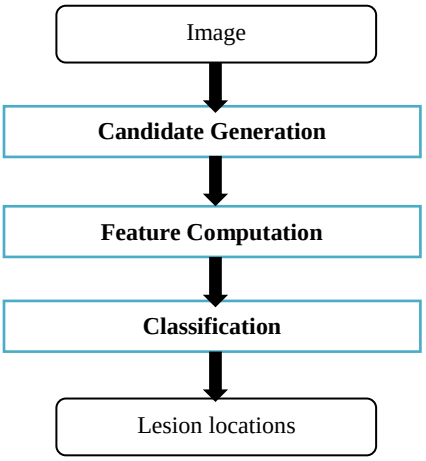

Figure 7 Typical data-flow architecture of software for mining clinical-image data

Many standard algorithms, such as support vector machines (SVM), back-propagation neural nets, kernel Fisher discriminant, have been used to detect malignant structures & to quantify key features. However, these general-purpose learning methods make implicit assumptions that are commonly violated in this application domain, resulting in sub-optimal performance.

For example, traditional learning methods almost universally assume that the training samples are independently drawn from an identical — albeit unobservable — underlying distribution (the ubiquitous i.i.d. assumption), which is almost never the case in medical image mining systems. There are high levels of correlations among the suspicious locations from the same region of an image, so the training samples are clearly not independent. Further, these standard algorithms try to maximize classification accuracy over all candidates. However, this particular accuracy measure is not very relevant for CAD. In Section 3.1 we show how the multiple-instance learning paradigm can solve both these problems. In Section 3.5 we show how to handle correlations among different candidates in the same image.

In CAD it is common that only an extremely small portion of the candidates identified in the candidate generation stage are actually associated with true malignant lesions, leading to a highly unbalanced sample distribution over the normal (negative) and abnormal (positive) classes. In Section 3.2 we show that cascaded classification schemes are extremely useful in balancing skewed data distribution as well as to reduce the runtime of the CAD system.

Unlike scenarios where we are given a set of features, in CAD the features are engineered by the researchers. When searching for descriptive features, researchers often deploy a large amount of experimental image features to describe the identified candidates, which consequently introduces a lot of irrelevant or redundant features. Sparse model estimation is often desired and beneficial (described in Section 3.3). The candidate generation step generally produces hundreds of candidates for a CT scan. Computing all the features can be very time-consuming. Hence it is imperative that the final classifier uses as few features as possible without any decrease in the sensitivity.

Medical domain knowledge often sheds a light on the essential learning process. For example a-priori we may know that there are three different kinds of abnormalities (positives). Efficiently incorporating related medical domain knowledge into the automatic learning algorithms yields better CAD systems. In Section 3.4 we show some of our solutions to incorporate domain knowledge. Section 3.6 and Section 3.7 describe some methods multi-task learning and learning with supervision from multiple experts.

3.1 Multiple-Instance Learning

In order to train a classifier, a set of CT scans/mammograms is collected from hospitals. These scans are then read by expert radiologists who mark the lesion locations; this constitutes our ground truth for learning. The candidate generation step generates a lot of potential candidates. Any candidate which is close to the radiologist mark is considered a positive example for training and the rest of the candidates are considered as negative examples. Candidates are labeled positive if they are within some pre-determined distance from a radiologist mark (see Figure 8 for an illustration); some of the positively labeled candidates may actually refer to healthy structures that just happen to be near a mark, thereby introducing labeling errors in the training data. These labeling errors can potentially sabotage the learning process by 'confusing' a classifier that is being trained with faulty labels, resulting in classifiers with poor performance. As shown in (Fung, et al. 2006) multiple-instance-learning is one of the effective ways to deal with this problem. During this labeling process, we also obtain information about which candidates point to the same underlying ground-truth lesion. While this information is typically discarded during the development of traditional classifiers, the multiple-instance learning (MIL) framework can utilize this information to extract more statistical power from the data.

In the MIL framework the training set consists of *bags*. A bag contains many instances. All the instances in a bag share the same bag-level label. A bag is labeled positive if it contains *at-least* one positive instance. A negative bag means that *all* instances in the bag are negative. The goal is to learn a classification function that can predict the labels of unseen instances and/or bags. Figure 9 illustrates that MIL can yield very different classifiers over the conventional single instance learning. The single instance classifier on the left is trying to reject as many negative candidates as possible and detect as many positives as possible. The MIL classifier on the right tries to detect at-least one candidate in a positive bag and reject as many negative candidates as possible.

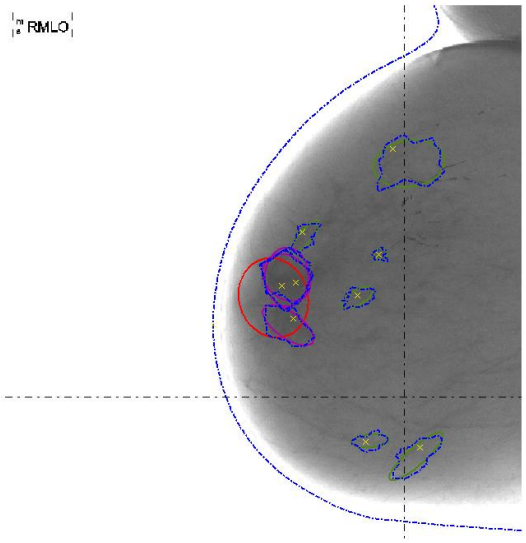

Figure 8 A mammogram of the right breast illustrating the concept of multiple candidates pointing to the same ground truth. The red ellipse is the lesion as marked by the radiologist (ground truth). The blue contours are the candidates generated by our algorithm.

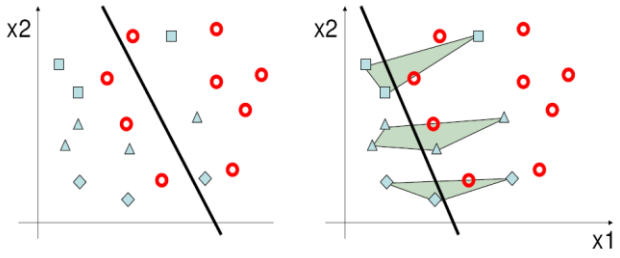

Figure 9 Illustration of single-instance learning (left) and multiple instance learning (right) for a toy problem. The red circles are negative candidates. The blue shapes are positives. There are three positive bags (square, triangle, and diamond).

There is another important reason why MIL is a natural framework for CAD. The candidate generation algorithm produces a lot of spatially close candidates. Even if one of these is highlighted to the radiologist and other adjacent or overlapping candidates are missed, the underlying lesion would still have been detected. Hence while evaluating the performance of CAD systems we use the bag level sensitivity, *i.e.*, a classifier is successful in detecting a lesion if at least one of the candidates pointing to it is predicted as a PE. MIL naturally lends itself to model our desired accuracy measure during training itself.

We have proposed several new MIL algorithms (Fung, et al. 2006, Raykar, Krishnapuram, et al. 2008, Krishnapuram, et al. 2008, Bi and Liang, Multiple Instance Learning of Pulmonary Embolism Detection with Geodesic Distance along Vascular Structure 2007, Chen, Bi and Wang 2006, Wu, Bi and Boyer 2009) specifically for the CAD domain. These involve a way of modifying the traditional classifiers for multiple-instance learning. These modifications have substantially improved our classifier's accuracy.

For example (Fung, et al. 2006) modified the SVM by forming convex hulls of the instances in each individual bag. The main idea is to search a good representative point in the convex hull and correctly classify this point in contrast to the conventional method where all the points in need to be correctly classified. This allows the classifiers to have certain degree of tolerance to noisy labels and makes use of the practical observation that not all candidates close to a nodule mark need to be identified. Mathematically, the convex-hull representation idea can be used with many loss functions and regularization operators.

In (Raykar, Krishnapuram, et al. 2008) we incorporate the definition of a positive bag to modify the link function in logistic regression. Standard logistic regression uses a sigmoid function to model the probability of the positive class. For MIL since we have the notion of a positive bag the probability that a bag contains at-least one positive instance is one minus the probability that all of them are negative. The algorithm selects features & designs the classifier jointly. Our results show that MIL based classifier selects much fewer features than conventional logistic regression and at the same time achieves better accuracy.

3.2 Cascaded classification architecture

Typical CAD training data sets are large and extremely unbalanced between positive and negative classes. In the candidate identification stage, high sensitivity (ideally close to 100 %) is essential, because any cancers missed at this stage can never be found by the CAD system, this high sensitivity at the candidate generation stage is achieved at the cost of a high false positives (less than 1% of the candidates are true lesions), making the subsequent classification problem highly unbalanced. Moreover, a CAD system has to satisfy stringent real-time performance requirements in order for physicians to use it during their diagnostic analysis.

These issues can be addressed by employing a cascade framework in the classification approach as discussed in (Bi, Periaswamy, et al. 2006). In Figure 10 a typical cascade classification scheme is shown. The key insight here is to reduce the computation time and speed-up online learning. This is achieved by designing simpler yet highly sensitive classifiers in the earlier stages of the cascade to reject as many negative candidates as possible before calling upon classifiers with more complex features to further reduce the false positive rate. A positive result from the first classifier activates the second classifier and a positive result from the second classifier activates the third classifier, and so on. A negative outcome for a candidate at any stage in the cascade leads to its immediate rejection.

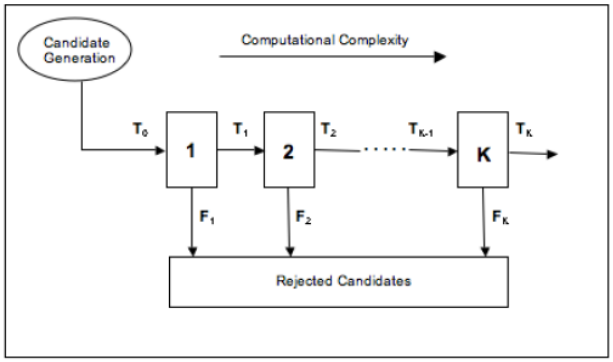

Figure 10 A general cascade framework used for online classification and training.

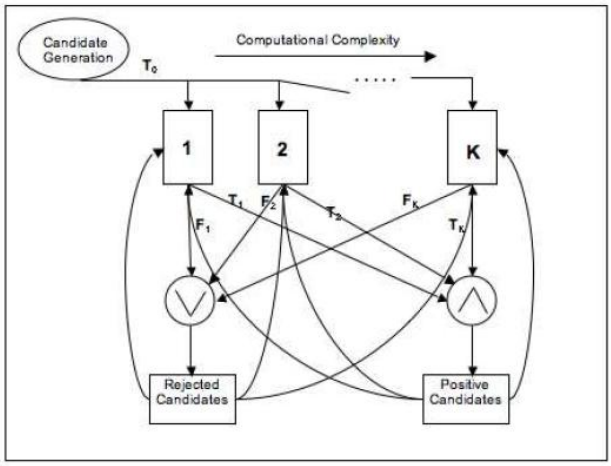

Figure 11 A novel AND-OR framework for training a cascade of classifiers.

The method in (Bi, Periaswamy, et al. 2006) investigated a cascaded classification approach that solves a sequence of linear programs, each constructing a hyperplane (linear) classifier. The linear programs are derived through piece-wise linear cost functions together with the l_1-norm regularization condition. The main idea is to incorporate the computational complexities of individual features into the feature selection process. Each linear program employs an asymmetric error measure that penalizes false negatives and false positives with different costs. An extreme case is that the penalty for a false negative is infinity, which is used in the early stage of the cascade design to alleviate the skewed class distribution and preserve high detection rates.

The approach in (Bi, Periaswamy, et al. 2006) follows the standard cascade procedure to train classifiers sequentially for each different stage, which amounts to a greedy scheme, meaning that the individual classifier is optimized only toward the specific stage given the candidates survived from the stages prior to it. The classifiers are not necessarily optimal to the overall structure where all stages are taken into account. A novel AND-OR cascade training strategy as illustrated in Figure 11 was proposed in (Dundar and Bi, Joint Optimization of Cascaded Classifiers for Computer Aided Detection 2007) to optimize all the classifiers in the cascade in parallel by minimizing the regularized risk of the entire system. By optimizing classifiers together, it implicitly provides mutual feedback to different classifiers to adjust parameter design. This strategy takes into account the fact that in a classifier cascade, a candidate is classified as positive only if all classifiers say it is positive, which amounts to an AND relation among classifiers. Nevertheless, a candidate is labeled as negative as long as one of the classifiers views it as negative, an OR relation of classifiers.

3.3 Feature selection/Learning sparse models

Feature selection has long become an important problem in statistics and machine learning and is highly desired in CAD applications. When searching for descriptive features, researchers often deploy a large amount of experimental image features to describe the identified candidates, which consequently introduces a lot of irrelevant or redundant features. It is also well-known that a reduction of features improves the classifier's generalization capability.

However, the problem of selecting an "optimal" subset of features from a large pool (in the orders of up to hundreds) of potential image features is known to be NP-hard. An early Lung CAD system utilized a greedy forward selection approach. Given a subset of features, the greedy approach consists of finding a new single feature from the feature pool that improves classification performance when considering the expanded subset of features. This procedure begins with an empty set of features and stops when classification performance does not improve significantly when any remaining feature is added. At each step, classification performance is measured based on Leave-One-Patient-Out (LOPO) cross-validation procedure (Dundar, Fung and Bogoni, et al. 2004).

Recent research has focused more on general sparsity treatments to construct sparse estimates of classifier parameters, such as in LASSO, the 1-norm SVM, and sparse Fisher's discriminant analysis. In (Dundar, Fung and Bi, et al. 2005) we proposed a sparse formulation for Fisher Linear Discriminant that scales well to large datasets; our method inherits all the desirable properties of FLD, while improving on handling large numbers of irrelevant and redundant features.

In (Raykar, Krishnapuram, et al. 2008) using a multiple instance learning setup we proposed a method to do feature selection and classifier design jointly using a Bayesian paradigm. Our results show that MIL based classifier selects much fewer features than conventional logistic regression and at the same time achieve better accuracy.

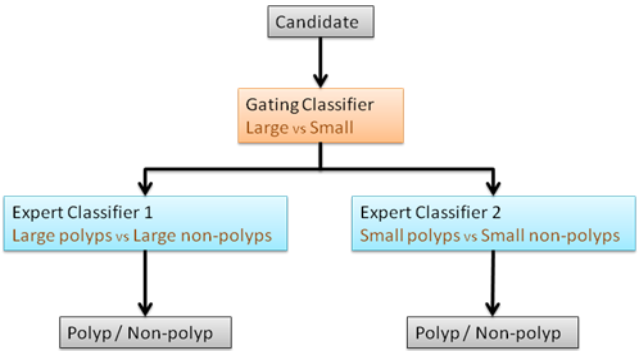

Figure 12 A typical gated classification architecture.

3.4 Gated classification architecture

Incorporating medical knowledge and prior observations can be critical to improving the performance of the CAD system. For example lesions have various characteristics in their shapes, sizes, and appearances. The simplest example is that lesions can be very big or small. Many of the image features are calculated by averaging over the voxels within segmented nodule. Features calculated on large lesions will hence be more accurate than those evaluated on a small one. Consequently, it may be more insightful to construct classifiers with separate decision boundaries respectively for large candidates and small candidates. *Gating* is a technique used to automatically learn meaningful clusters among candidates and construct classifiers, one for each cluster, to classify true candidates from false detections. (See Figure 12) This can obviously be extended to incorporate different kinds of knowledge, for instance, to exploit differences between the properties of central versus peripheral nodules, or between sessile and flat polyps.

A novel Bayesian hierarchical mixture of experts (HME) has been developed and tested in our Lung CAD system. The basic idea behind the HME is to decompose a complicated task into multiple simple and tractable subtasks. The HME model consists of several domain experts and a gating network that decides which experts are most trustworthy on any input pattern. In other words, by recursively partitioning the feature space into sub-regions, the gating network probabilistically decides which patterns fall in the domain of expertise of each expert.

In many scenarios we know what kind of false positives our system generates. We may also have labels for the different sub-classes in the negatives. In (Dundar, Wolf, et al. 2008) we presented a methodology to take advantage of the subclass information available in the negative class to achieve a more robust description of the target class. The subclass information which is neglected in conventional binary classifiers provides a better insight of the dataset and when incorporated into the learning mechanism acts as an implicit regularizer. We proposed a method to train a polyhedral classifier jointly, where each face of the polyhedron can classify each of the negative sub-class. The linear faces of the polyhedron achieve robustness whereas multiple faces provide flexibility.

3.5 Handling internal correlations

Most classification systems assume that the data used to train and test the classifier are independently drawn from an identical underlying distribution. For example, samples are classified one at a time in a support vector machine (SVM), thus the classification of a particular test sample does not depend on the features from any other test samples. Nevertheless, this assumption is commonly violated in many real-life problems where sub-groups of samples have a high degree of correlation amongst both their features and their labels. Due to spatial adjacency of the regions identified by a candidate generator, both the features and the class labels of several adjacent candidates can be highly correlated during training and testing. We proposed batch-wise classification algorithms to explicitly account for correlations (Vural, et al. 2009).

In this setting, correlations exist among both the features and the labels of candidates belonging to the same (batch) image both in the training data-set and in the unseen testing data. Furthermore, the level of correlation can be captured as a function of the pair wise-distance between candidates: the disease status (class-label) of a candidate is highly correlated with the status of other spatially proximate candidates, but the correlations decrease as the distance is increased. Most conventional CAD algorithms classify one candidate at a time, ignoring the correlations amongst the candidates in an image. Explicitly accounting for the correlation structure between the labels of the test samples, the algorithms proposed in (Vural, et al. 2009) jointly predict class assignments of spatially nearby candidates to improve the classification accuracy significantly.

3.6 Multiple-task Learning

We are often faced with a shortage of training data for learning classifiers for a task. However we may have additional data for closely related, albeit non-identical tasks. For example our data set includes images from CT scanners with two different reconstruction kernels. While training the classifier we could ignore this information and pool all the data together. However, there are some systematic differences that make the feature distributions slightly different. Alternatively, we could train a separate classifier for each kernel, but a large part of our data set is from one particular kernel and we have a smaller data set for the other. In (Raykar, Krishnapuram, et al. 2008, Bi, Xiong, et al. 2008) we use multi-task learning that tries to estimate models for several classification tasks in a joint manner. Multi-task learning can compensate for small sample size by using additional samples from related tasks, and exchanging statistical information between tasks.

3.7 Learning from multiple experts

In many CAD applications it is actually quite difficult to obtain the ground truth. The actual gold standard (whether

it is cancer or not) can be obtained from biopsies, but since it is an expensive and an invasive process, often CAD systems are built from labels assigned by multiple radiologists who identify the locations of malignant lesions. Each radiologist visually examines the medical images and provides a subjective (possibly noisy) version of the gold standard. In practice, there is a substantial amount of disagreement among the experts, and hence it is of great practical interest to determine the optimal way to learn a classifier in such a setting.

In (Raykar, Yu, et al. 2009) we propose a Bayesian framework for supervised learning in the presence of multiple annotators providing labels but no absolute gold standard. The proposed algorithm iteratively establishes a particular gold standard, measures the performance of the annotators given that gold standard, and then refines the gold standard based on the performance measures. Experimental results indicate that the proposed method is superior to the commonly used majority voting baseline.

3.8 Scalability for massive data

Often a great amount of candidates are commonly produced in the candidate generation stage to uncover any suspicious regions, which results in large massive training data. This imposes a requirement for the scalability of the learning algorithms. Typically we have observed that linear models are more computationally tractable than sophisticated non-linear methods. Boosting algorithms are also efficient to scale up with large data.

3.9 Detection of shapes

In this section we will briefly describe our proposed solution for the estimation of the ejection fraction. Accurate analysis of the myocardial wall motion of the left ventricle is crucial for the evaluation of the heart function. This task is difficult due to the fast motion of the heart muscle and respiratory interferences. It is even worse when ultrasound image sequences are used since ultrasound is the noisiest among common medical image modalities such as MRI or CT. Figure 13 illustrates the difficulties of the tracking task due to signal dropout, poor signal to noise ratio or significant appearance changes.

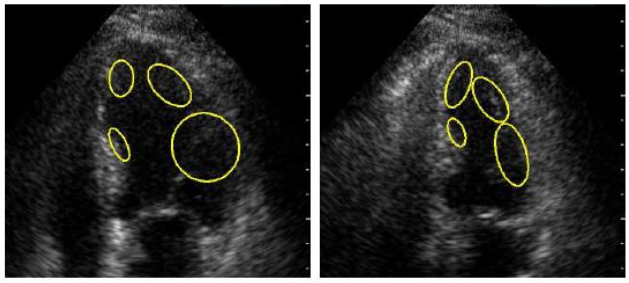

Figure 13 Echocardiography images with area of acoustic drop-out, low signal to noise ratio and significant appearance changes. Local wall motion estimation has covariances (depicted by the solid ellipses) that reflect noise.

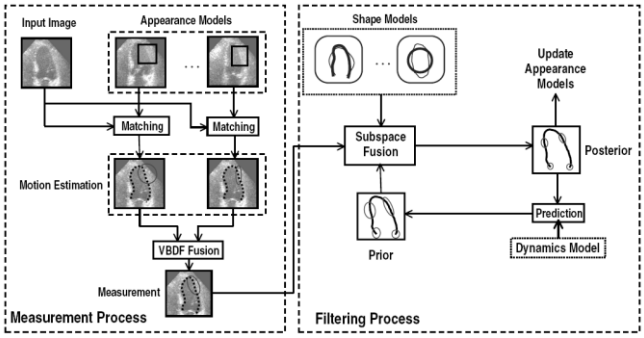

Figure 14 The block diagram of the robust tracker with the measurement and filtering processes

In (Georgescu, et al. 2004) a unified framework was introduced for fusing motion estimates from multiple appearance models and fusing a subspace shape model with the system dynamics and measurements with heteroscedastic noise. The appearance variability is modeled by maintaining several models over time. This amounts for a nonparametric representation of the probability density function that characterizes the object appearance. Tracking is performed by obtaining independently from each model a motion estimate and its uncertainty through optical flow. The diagram of the proposed robust tracking proposed is illustrated in Figure 14. The approach is robust in two aspects: in the measurement process, Variable-Bandwidth Density-based fusion is used for combining matching results from multiple appearance models and in the filtering process, fusion is performed in the shape space to combine information from measurement, prior knowledge and models while taking advantage of the heteroscedastic nature of the noise.

4. CLINICAL IMPACT

The true measure of impact for a medical image mining system is not in terms of how accurate it is, but rather how much a radiologist can benefit by using the software. For example, most CAD systems are deployed in a *second reader* mode, *i.e.*, the radiologist invokes the CAD only after he/she has read the case without any prompting from the CAD software.

A radiologist is likely to use the software only if it is clinically validated. In order to measure the impact of our software in such a scenario we have conducted several clinical studies/trials with our collaborators in different parts of the world. These studies have been conducted independently by our collaborators and the results have been disseminated at the annual meetings of the Radiological Society of North America, European Society of Radiology, and various radiology journals. In this section we will briefly describe some of the most recent studies and results, which should give a sense of the real impact of our medical image mining products. In order to keep this section concise we have only presented a sample

of the studies and omitted a large number of other studies from the literature. *All the systems studied below are commercially distributed worldwide*[3].

4.1 Lung

In a clinical validation study submitted to the FDA, we analyzed a retrospective sample of 196 cases from 4 large research hospitals. CT Scans were collected from patients referred for routine assessment of clinically or radiographically known or suspected pulmonary nodules. These cases contained a total of 1320 nodules as confirmed by a majority of a panel of 5 expert radiologists. The cases were interpreted independently by 17 community radiologists first without and then with the use of our LungCAD product. Every one of these 17 radiologists improved their detection of solid nodules ≥ 3 mm to a statistically significant extent. The average reader improvement in AUC using the nonparametric ROC technique for detecting nodules was 0.048 ($p<0.001$) with a 95% confidence interval of (0.036, 0.059). This study showed a statistically significant improvement in the area under the nonparametric ROC curve with the use of our Lung CAD software for detection of lung nodules

A subsequent clinical study (Godoy, et al. 2008) was done with 54 chest CT scans by a group of four radiologists at New York University Medical Center to evaluate the impact of our most recent Lung CAD system at finding different kinds of nodules in the lung. The 54 cases used in the study had total of 395 nodules of which 234 were solid nodules, 29 were part-solid nodules, and 132 ground-glass opacities. Two readers read the 54 cases first without CAD and then with CAD. The study showed that (see Table 1) the CAD software resulted in a significant increase in sensitivity by 9.8 % for reader 1 and by 10.6 % for reader 2. The use of CAD did not increase the number of false positives for any of the readers.

Newer research prototype systems have also been studied although they have not yet been distributed commercially. A study (LungCAD_ARRS_2009) presented at the recent American Roentgen Ray Society 2009 annual meeting concluded that the use of our research prototype significantly increased the mean reader sensitivity in all subgroups ($p < 0.001$) (See Table 2).

Based on these and many other clinical studies, we have demonstrated that the *use of CAD as a second reader improves radiologist's detection of different kinds of pulmonary nodules*.

[3] *PE CAD & MammoCAD are only sold outside the US.*

	Sensitivity without CAD	*Sensitivity with CAD*	*Increase in sensitivity*
Reader 1	56.2 %	66.0 %	**9.8 %**
Reader 2	79.2 %	89.8 %	**10.6 %**

Table 1 Sensitivity for detecting different kinds of lung nodules without and with CAD for two different readers. Results are from (Godoy, et al. 2008).

	Mean sensitivity without CAD	*Mean sensitivity with CAD*	*Increase in sensitivity*
Solid Nodules	60%	85%	**15 %**
Part-solid Nodules	80%	95%	**15%**
Ground Glass Opacities	75%	86%	**11%**

Table 2 Mean sensitivity of four radiologists for detecting different kinds of lung nodules without and with CAD. Results are for a recent research prototype from (LungCAD_ARRS_2009).

4.2 Colon

For CT colonography we have developed what we call a Polyp Enhanced Viewer (PEV) system which helps the radiologist to properly visualize the polyps in three dimensions. In (Bogoni, et al. 2005) the utility of PEV was evaluated as part of a study involving data sets obtained from two sites, New York University Medical Center and the Cleveland Clinic Foundation. PEV resulted in 90% sensitivity for detection of medium and large sized polyps.

Another study (Baker, et al. 2007) was conducted to determine whether PEV can help improve sensitivity of polyp detection by less-experienced radiologist readers. Seven less-experienced readers from two institutions evaluated the CT colonographic images and marked polyps with and without PEV. *The average sensitivity of the seven readers for polyp detection was significantly improved with PEV—from 0.810 to 0.908 ($p = .0152$), a 9.8% increase in sensitivity*. The number of false-positive results per patient without and with PEV increased from 0.70 to 0.96. Results of this study suggest that our software significantly improves polyp detection among less-experienced readers.

4.3 Pulmonary Embolism

Several independent evaluations of our PE CAD system have been performed in real clinical settings. Dr. Das et al (Das 2008) conducted a clinical study whose objectives were to assess the sensitivity of our PE CAD system for the detection of pulmonary embolism and to assess the influence on radiologists' detection performance. Forty-three patients with suspected PE were included in this

retrospective study. Sensitivity for the radiologist with and without CAD software was assessed. The mean overall sensitivity for the CAD software alone was 83%. Table 3 summarizes the improvements in sensitivity obtained for the three readers.

Dr. Lake et al (Lake 2006) also examined our PE CAD system and investigated its influence on interpretation by resident readers. 25 patients with suspected pulmonary embolus were included in this study. Four radiology residents first independently reviewed all CT scans on a dedicated workstation and recorded sites of suspected PE and then reanalyzed all studies for a second time with the aid of the PE CAD. Overall, mean detection of PE by resident readers was increased from 53.5 % to 58.9 ($p<0.028$). Table 4 summarizes the improvements in sensitivity obtained for the four residents.

Dr. Buhmann and his team (Buhmann, et al. 2007) evaluated another system with 40 clinical cases. This system was designed for detecting peripheral emboli, because the central emboli can be easily detected by radiologists. The evaluation concluded that CAD detection of findings incremental to the radiologists suggests benefits when used as a second reader, especially for peripheral emboli. We have also received feedbacks from other clinical sites with our PE CAD installations for evaluations. The consensus is that our PE CAD system is of special value in the emergency room, as it boosts the physicians' confidence in negative reports and reduces missing diagnosis; a critical issue in current PE patient managements, as diagnosis has been missed in about 70% of the cases.

	Sensitivity without CAD	*Sensitivity with CAD*	*Increase in sensitivity*
Reader 1	87%	98%	**11%**
Reader 2	82%	93%	**11%**
Reader 3	77%	92%	**15%**

Table 3 Sensitivity for detecting pulmonary emboli in the lung without and with CAD for three different readers. Results are from (Das 2008)

	Sensitivity without CAD	*Sensitivity with CAD*	*Increase in sensitivity*
Resident 1	46.7 %	52.3 %	**5.6%**
Resident 2	57.9 %	59.8 %	**1.9%**
Resident 3	100 %	100 %	**0.0%**
Resident 4	91.7 %	100 %	**8.3%**

Table 4 Sensitivity for detecting pulmonary emboli in the lung without and with CAD for three different readers. Results are from (Lake 2006)

4.4 Breast

A study (Bamberger, et al. 2008) was conducted in order to assess the performance of a mammography algorithm designed to detect clusters, deemed actionable by expert radiologists. It was shown that algorithm achieved the goal of reproducing the performance of expert radiologists with 98% sensitivity and very few false marks. The algorithm performed equally well in dense and non-dense breasts.

4.5 Cardiac

A clinical study (Cannesson, et al. 2007) was conducted to test three hypotheses regarding our AutoEF product: (1) AutoEF produces similar results to manually traced biplane Simpson's rule. (2) AutoEF performs with less variability than visual EF calculated by expert and novice readers. The EF was calculated by visual assessment by expert readers using all available views. (3) AutoEF correlates favorably with EF calculated by using magnetic resonance imaging (MRI). The study made the following conclusions: (1) When comparing AutoEF to manual biplane Simpsons' rule the two methods were closely related (with correlation coefficient $r = 0.98$; $p < 0.01$). (2) AutoEF correlated well with visual EF by expert readers ($r = 0.96$; $p < 0.001$), with a bias of 2%. The novice readers achieved similar results to that of the experts when using AutoEF ($r = 0.96$; $p < 0.001$), even though they operated AutoEF for their first time. There was significantly lower inter-observer and intra-observer variability using AutoEF. (3) A favorable correlation was observed between AutoEF and MRI based estimation of EF.

5. CONCLUSIONS & LESSONS LEARNT

5.1 Impact of the commercially distributed products described in this paper

In an era of dramatic medical advances, radiologists now have access to orders of magnitude more data for diagnosing patients. Paradoxically, the deluge of data makes it more difficult & time consuming to identify key clinical findings for improving patient diagnosis & for therapy selection. This paper describes our commercially deployed software for mining medical images to identify or to enhance the viewing of suspicious structures such as nodules, possible polyps, possibly early stage breast cancers (masses, clusters of micro-calcifications etc), pulmonary emboli, etc. This paper also described our commercial software for the quantification of key clinical information contained in raw image data.

Every system described in this paper is marketed internationally[4] by Siemens Medical Solutions USA, Inc.

[4] *The Lung Nodule Enhanced Viewing software (predecessor to the current version of the LungCAD product) was launched in 2004, AutoEF software was launched in 2006, Polyp Enhanced Viewing & PE detection software were launched in 2007, and*

Together, several thousand units of these products have been installed in hospitals. With radiology data (images) expected to reach 25% of all data in the hospital within the next few years, it is critical to have key enablers like knowledge driven, role-based context sensitive data mining software in this domain. The field is undergoing explosive growth, and there is a key opportunity for data mining technologies to impact patient care worldwide.

5.2 Lessons Learnt

Along the way, while developing these systems we learnt several key points that are absolutely critical for large scale adoption of data mining systems in an area where there is initially a lot of skepticism about the abilities of computerized systems. One of the key lessons was that the systems are not successful just by being more accurate. Their true measure of impact is in terms of how much they improve the radiologists in *their diagnoses of patients*, assisted by software. This raises the need for extensive validation of how the radiologists' accuracy changes while using the system.

The second key lesson learnt was the need for first principles research innovation specific to the data domain. While we initially tried off-the-shelf methods like SVMs, we quickly learnt the need to focus on the specific data domain and the key data characteristics & requirements therein. We learnt that changing from an SVM to a boosting algorithm or a neural network really was not what improved system performance in a significant manner, it was absolutely essential to carefully analyze data, visualize and re-think the fundamental assumptions, evaluate which assumptions are appropriate for the problem, and study how we can change them while still retaining mathematical tractability. For example, we realized that the data is never independent and identically distributed (i.i.d.), a key assumption that is almost universal in most of the traditional classifier design technologies such as SVMs, neural networks etc.

Along the way we learnt that the interplay between image processing and data mining components was crucial, and it was important to understand the impact of each component on the other in order to jointly optimize the overall product. Indeed good image processing algorithms created the features that made subsequent data mining algorithms successful, and often a deep analysis of the fundamental ideas behind these algorithms would lead to a much better understanding of the statistical issues that would be faced by the classifier.

Driven by the needs of our data and our problem, we re-evaluated the assumptions and re-thought systems from first principles. This resulted in huge domain-specific improvements in system accuracy measures that are relevant for our products (as opposed to accuracy measures used in the data mining community based off the 0-1 loss for example). In all honesty, the initial approach of throwing a bunch of data mining algorithms at a problem and seeing what stuck simply led to initial disasters until we were humble enough to work on the problem we had rather than the method we liked. This was a second key lesson for us (most of the authors were practitioners who often came fresh from grad school trained in the data mining).

A final lesson learnt from our work in this area was the need for securing buy in (and leadership) from key clinical subject matter experts in order to have them drive the product features and capabilities. Many of the key product definition ideas were a result of collaborating with radiologists who identified the key capabilities in system that should be developed – our best guesses as data mining researchers were based on what we found technologically challenging or exciting, but often a feature which was much less time consuming and "low-tech" added much more value to end-users. The lesson was that while mining can add value, its use should be defined in collaboration with the end-user in order to fully exploit it in their workflow.

MammoCAD was launched in 2008. MammoCAD & PE CAD software are not sold in the USA, although they are commercially marketed in the rest of the world.

6. ACKNOWLEDGMENTS

We are grateful to our clinical collaborators who helped shape and validate these products. We would particularly like to thank Dr. D. Naidich, MD, of New York University, Dr. M. E. Baker, MD, of the Cleveland Clinic Foundation, Dr. M. Das, MD, of the University of Aachen, Dr. U. J. Schoepf, MD, of the Medical University of South Carolina, and Dr. Peter Herzog, MD, of Klinikum Grossharden, Munich. Any commercial product relies on the dedication of a large team. We are particularly grateful to our Siemens colleagues: Alok Gupta, Ph.D., Ingo Schmuecking, MD, Harald Steck, Ph.D., Maleeha Qazi, Stefan Niculescu, Ph.D., Romer Rosales, Ph.D., and the entire SISL team.

7. BIBLIOGRAPHY

Ambrose, J, and Hounsfield. "Computerized transverse axial scanning (tomography)." *British Journal of Radiology* 46, no. 552 (1973): 1016-1023.

Armato-III, S G, M L Giger, and H Mac Mahon. "Automated detection of lung nodules in CT scans: preliminary results." *Medical Physics* 28, no. 8 (2001): 1552-1561.

Baker, M E, et al. "Computer-aided detection of colorectal polyps: can it improve sensitivity of less-experienced readers? Preliminary findings." *Radiology* 245, no. 1 (2007): 140-149.

Bamberger, P, I Leichte, N Merlet, G Fung, and R Lederman. "A New Generation Algorithm for Digital Mammography Designed to Reproduce the Performance of Expert Radiologists in Detecting Actionable Clusters." *European Radiology Supplements* 18, no. 1 (2008).

Bi, J, and J. Liang. "Multiple Instance Learning of Pulmonary Embolism Detection with Geodesic Distance along Vascular Structure." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'07).* 2007.

Bi, J, et al. "Computer aided detection via asymmetric cascade of sparse hyperplane classifiers." *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2006.

Bi, J, T Xiong, S Yu, M Dundar, and R B Rao. "An Improved Multi-task Learning Approach with Applications in Medical Diagnosis." *Proceedings of the 18th European Conference on Machine Learning (ECML'08).* 2008.

Bogoni, L, et al. "Computer-aided detection (CAD) for CT colonography: a tool to address a growing need." *British Journal of Radiology*, no. 78 (2005): s57-s62.

Buhmann, S, et al. "Clinical evaluation of a computer aided diagnosis (CAD) prototype for the detection of pulmonary embolism." *Academic Radiology* 14, no. 6 (2007): 651-658.

Cannesson, M., et al. "A Novel Two-Dimensional Echocardiographic Image Analysis System Using Artificial Intelligence-Learned Pattern Recognition for Rapid Automated Ejection Fraction." *Journal of the American College of Cardiology* 49, no. 2 (2007): 217-226.

Chen, Y, J Bi, and J Z Wang. "MILES: Multiple-Instance Learning via Embedded Instance Selection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, no. 12 (2006): 1-17.

Das, M et al. "Computer-aided detection of pulmonary embolism: Influence on radiologists' detection performance with respect to vessel segments." *European Radiology* 18, no. 7 (July 2008): 1350–1355.

Dundar, M, and J Bi. "Joint Optimization of Cascaded Classifiers for Computer Aided Detection." *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'07).* 2007.

Dundar, M, G Fung, J Bi, S Sandilya, and R. B. Rao. "Sparse Fisher discriminant analysis for computer aided detection." *Proceedings of SIAM International Conference on Data Mining.* 2005.

Dundar, M, G Fung, L Bogoni, M Macari, A Megibow, and R B Rao. "A methodology for training and validating a CAD system and potential pitfalls." *Proceedings of CARS 2004 Computer Assisted Radiology and Surgery.* 2004.

Dundar, M, M Wolf, S Lakare, M Salganicoff, and V C Raykar. "Polyhedral Classifier for Target Detection A Case Study: Colorectal Cancer." *Proceedings of the 25th International Conference on Machine Learning (ICML 2008).* 2008. 288-295.

Fung, G, M Dundar, B Krishnapuram, and R B Rao. "Multiple instance algorithms for computer aided diagnosis." *Advances in Neural Information Processing Systems 19* . 2006. 425-432.

Georgescu, B, X S Zhou, D Comaniciu, and R B Rao. "Real-Time Multi-Model Tracking of Myocardium in Echocardiography using Robust Information Fusion." *Proceedings of MICCAI.* 2004.

Godoy, M C, T Kim, J P Ko, C Florin, A K Jerebko, and D P Naidich. "Computer-aided Detection of Pulmonary Nodules on CT: Evaluation of a New Prototype for Detection of Ground-glass and Part-Solid Nodules." *Radiological Society of North America scientific assembly and annual meeting program (RSNA 2008).* 2008.

Group, U.S. Cancer Statistics Working. *United States Cancer Statistics: 1999–2005 Incidence and Mortality Web-based Report.* http://www.cdc.gov/uscs, Atlanta: Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute, 2009.

Jemal, A, R Siegel, E Ward, T Murray, J Xu, and M J. Thun. "Cancer Statistics." *CA: A Cancer Journal for Clinicians* 57 (2007): 43-66.

Krishnapuram, B, et al. "Multiple instance learning improves CAD detection of masses in digital mammography." *Proceedings of the 9th international workshop on Digital Mammography (IWDM 2008).* 2008. 350-357.

Lake, et al. "Computer-aided detection of peripheral pulmonary embolus on multi-detector row CT: Initial experience and impact on resident diagnosis." *The 106th annual meeting of the American Roentgen Ray Society.* 2006.

Naidich, D P, J. P. Ko, and J Stoechek. "Computer aided diagnosis: Impact on nodule detection amongst community level radiologist. A multi-reader study." *Proceedings of Computer Assisted Radiology and Surgery.* 2004. 902 -907.

Park, S, T J Kim, V C Raykar, V Anand, M Deewan, and A Jerebko. "Assessment of Computer-aided Nodule Detection (CAD) Algorithm on Pathology Proved CT Data Sets." *Radiological Society of North America scientific assembly and annual meeting program (RSNA 2008),.* 2008.

Raykar, V C, B Krishnapuram, J Bi, M Dundar, and R B Rao. "Bayesian Multiple Instance Learning: Automatic Feature Selection and Inductive Transfer." *Proceedings of the 25th International Conference on Machine Learning (ICML 2008).* 2008. 808-815.

Raykar, V C, et al. "Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit." *Proceedings of the 26th International Conference on Machine Learning (ICML 2009).* 2009.

Stanton, A. "Wilhelm Conrad Röntgen On a New Kind of Rays: translation of a paper read before the Würzburg Physical and Medical Society, 1895." *Nature* 53 (1896): 274–276.

Suzuki, K, M Kusumoto, S I Watanabe, R Tsuchiya, and H Asamura. "Radiologic classification of Small Adenocarcinoma of the Lung: Radiologic-Pathologic Correlation and Its Prognostic Impact." *Annals of Thoracic Surgery* 81, no. 2 (2006): 413-419.

Vural, V, G Fung, B Krishnapuram, J G Dy, and B R Rao. "Using Local Dependencies within Batches to Improve Large Margin Classifiers." *Journal of machine learning research* 10 (2009): 183-206.

Wu, D, J Bi, and K Boyer. "A Min-Max Framework of Cascaded Classifier with Multiple Instance Learning for Computer Aided Diagnosis." *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'09).* 2009