Computer-aided detection and quantification of interstitial lung disease from computed tomography and chest radiography

Yulia Arzhaeva

This book was typeset by the author using LaTeX2$_\varepsilon$.

Cover was designed by the author using ZMatrix, a desktop enhancement tool that creates the streaming character effect of "The Matrix" movie. This program is freely available from `http://zmatrix.sourceforge.net`.

ISBN 978-90-393-5142-0

Printed by Wöhrmann Print Service, Zutphen.

Publication of this thesis was financially supported by: Delft Imaging Systems BV, the Netherlands, Philips Medical Systems Nederland B.V., the Röntgen Stichting Utrecht, 3mensio Medical Imaging BV, and ImagO.

Computer-aided detection and quantification of interstitial lung disease from computed tomography and chest radiography

Computerondersteunde detectie en kwantificatie van interstitiële longziekten vanuit computertomografie en röntgenbeelden van de thorax
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op donderdag 17 september 2009 des middags te 12.45 uur

door

Yulia Arzhaeva

geboren op 15 oktober 1976 te Minsk, Wit-Rusland

Promotoren: **Prof. dr. ir. M. A. Viergever**
Prof. dr. W. M. Prokop

Co-promotor: **Dr. B. van Ginneken**

The research described in this thesis was supported by the Dutch Technology Foundation (STW), applied science division of Netherlands Organisation for Scientific Research (NWO), and the technology programme of the Dutch Ministry of Economic Affairs under project number 6126.

Contents

Chapter 1

Introduction and outline

1.1 Introduction

This thesis presents computer-aided diagnosis (CAD) systems for automatic detection and quantification of interstitial lung disease (ILD) in conventional chest radiographs and computed tomography (CT) scans. The purpose of this introduction is to provide the reader with background information necessary for understanding the thesis's contents. Section 1.1.1 describes the basic lung anatomy and familiarizes the reader with interstitial lung disease. Two imaging modalities for diagnostics of ILD - conventional chest radiography and computed tomography - are introduced in Sections 1.1.2 and 1.1.3, respectively. In Section 1.1.4, fundamental principles and prerequisites in designing CAD systems presented in this thesis are explained.

1.1.1 The human lungs and interstitial lung disease

The lungs are the essential organs of respiration. They are two in number, placed one on either side of the chest, sometimes referred to as thorax. The left and right lungs are separated from each other by the heart, the great vessels of the heart, and the other organs of the mediastinum. The surface of each lung is covered by a thin membrane called pulmonary pleura[1]. The lung is connected to the trachea and heart through the bronchus and the pulmonary arteries and veins. A place where these organs enter the lung is called the hilum, and the region around it - the perihilum. The lung tissue is called parenchyma. The anatomical unit of the parenchyma is a primary lobule, which consists of an alveolar duct, tiny air cells, or alveoli, connected with it, their blood and lymphatic vessels, and nerves. The alveolar duct is connected to a bronchiole, the smallest subdivision of the intrapulmonary bronchi. The direct function of the lung, the exchange of gases between the atmosphere and the bloodstream, is accomplished in alveoli.

"Interstitial lung disease" is a term that encompasses the large group of disorders that primarily affect the lung parenchyma in a diffuse manner. The in-

[1]Medical terms related to the lungs often begin with *pulmo-*, or with *pneumo-*, originating from Latin and Greek words for *lung*, respectively.

terstitium is the connective tissue containing numerous elastic fibers and blood capillaries, that surrounds and separates the alveoli. The word comes from two Latin words: *inter* meaning *between*, and *sistere* meaning *to stand* - to stand between. Interstitial lung disease primarily involves the inflammation and fibrosis of the interstitium, but can also affect the alveolar space, peripheral airways and vessels [1]. The term "interstitial lung disease" is synonymous with "diffuse parenchymal lung disease" [2].

Clinically, the diseases have similar manifestation with increasing shortness of breath, cough, and widespread shadowing on a chest radiograph. The disease etiology, treatment and prognosis is, however, very different for different disorders. A disease can be chronic or acute, have a known or unknown cause. A group of chronic interstitial lung diseases with known causes includes disorders caused by occupational, environmental or drug agents, as well as disorders secondary to some systemic diseases. The most common chronic disorders are those whose cause is unknown, among them idiopathic interstitial pneumonias (IIPs) and sarcoidosis. Some types of ILD have poor prognosis and low survival rates, with only palliative treatment available, while the other types can be treated effectively. An accurate diagnosis is essential to the appropriate management of a patient. Medical treatments for some types of ILD are toxic and capable of producing severe side effects, therefore, should not be administered mistakenly. An up-to-date classification and recommendations on diagnostics and treatment of ILD can be found in Refs. [1, 2, 3].

TB is often regarded separately from chronic ILD. TB is an infectious disease and the major cause of illness and death worldwide, especially in developing countries. The World Health Organization (WHO) reported an estimated 9.27 million new cases of TB and 1.75 million deaths in 2007 [4]. In most cases, TB can be eliminated with modern antibiotics when diagnosed timely. One of the main targets of the WHO is to halve TB prevalence and death rates by 2015 compared with their level in 1990 [5].

Imaging plays an important role in diagnosing TB and ILD. Since TB is caused by a microorganism, only microbiological examination can produce a definitive diagnosis. Chest radiography is used in diagnosing such cases where clinical suspicion of TB exists but the sputum smear is negative [6]. With the advent of digital radiography, x-ray examination is increasingly employed in screening high-risk population groups for TB, in order to detect suspect cases that should undergo further examinations.

For diagnosing ILD a multidisciplinary approach is recommended, with the agreement of clinical, radiological and histological information. Previously, the final diagnosis was solely based on histology made on surgical biopsy, which is an invasive procedure associated with a lot of contraindications. Besides, histological findings are often not differential enough, i.e. could be associated with two or more diagnoses [2]. Nowadays, it is widely accepted that a diagnosis of many ILDs can be reasonably supported by analysis of patterns appearing at thin-section computed tomography of the lungs [2, 7]. In clinical practice, a conventional x-

ray is usually the first examination on which the initial detection of interstitial abnormalities is performed.

1.1.2 Conventional radiography

Until the beginning of the last century no means existed to investigate the internal world of the living human body. The discovery of penetrating x-rays by W.C. Röntgen in 1895 started the science of diagnostic medical imaging [8]. A chest radiograph, commonly called a chest x-ray, is a 2D projection image obtained by directing a beam of x-rays at the patient's chest. A fraction of x-ray photons passes through the body and onto some sensitive photon detector, e.g. a phosphor screen. The x-ray photon detector, in turn, emits light photons that can be either detected by a light-sensitive film (analog radiography), or released and digitized by a computed radiography reader device (digital computed radiography), or converted to electrical charge by photodiodes embedded in the photon detector (digital radiography). In the last two cases image acquisition results in a digital image stored as a computer file. Digital radiography has many advantages and is expected to completely replace the analog technique in the near future [9].

The contrast between the elements of the chest in a radiograph is achieved due to the different degrees of x-ray absorption by bones, soft tissue, and the air in the lungs while x-ray photons are traveling through the body. Different views of the chest can be obtained by changing the projection direction. The most common is a posterior-anterior (PA) view: the x-rays enter through the back of the chest and exit out of the front where they are detected. Figure 1.1 shows a PA chest radiograph that depicts healthy lungs, with several normal structures labeled. Since a radiograph is a 2D projection of a 3D subject, organs and structures situated at different depths in the thorax get superimposed in the final image. This is one of the major limitations of conventional radiography. Degraded contrast and image noise because of scattered radiation are the other limitations. Nevertheless, due to its simplicity and low cost, chest radiography remains the staple for diagnosis of many thoracic diseases.

The lung parenchyma affected by ILD exhibits a variety of abnormal texture patterns in the radiograph. In TB, abnormal changes in pleura or mediastinum can be present, in addition to textural abnormalities in the parenchyma. In this thesis we only focus on radiological findings related to the lung parenchyma.

Detection of interstitial abnormalities in chest radiographs is a clinically difficult task because of overlapping lung anatomy and low contrast of subtle abnormalities with diverse radiological presentations. Moreover, abnormal radiological findings are nonspecific - what looks similar in the x-ray can correspond to different ILDs [10]. In Figures 1.2 and 1.3 several close up views of TB and ILD lesions, respectively, are shown in order to illustrate the variety and, often, subtlety of interstitial abnormalities in chest radiographs. Note how the shadows formed by superimposed structures, such as ribs, blood vessels, scapulas and fat, degrade the contrast of lesions and obscure their borders.

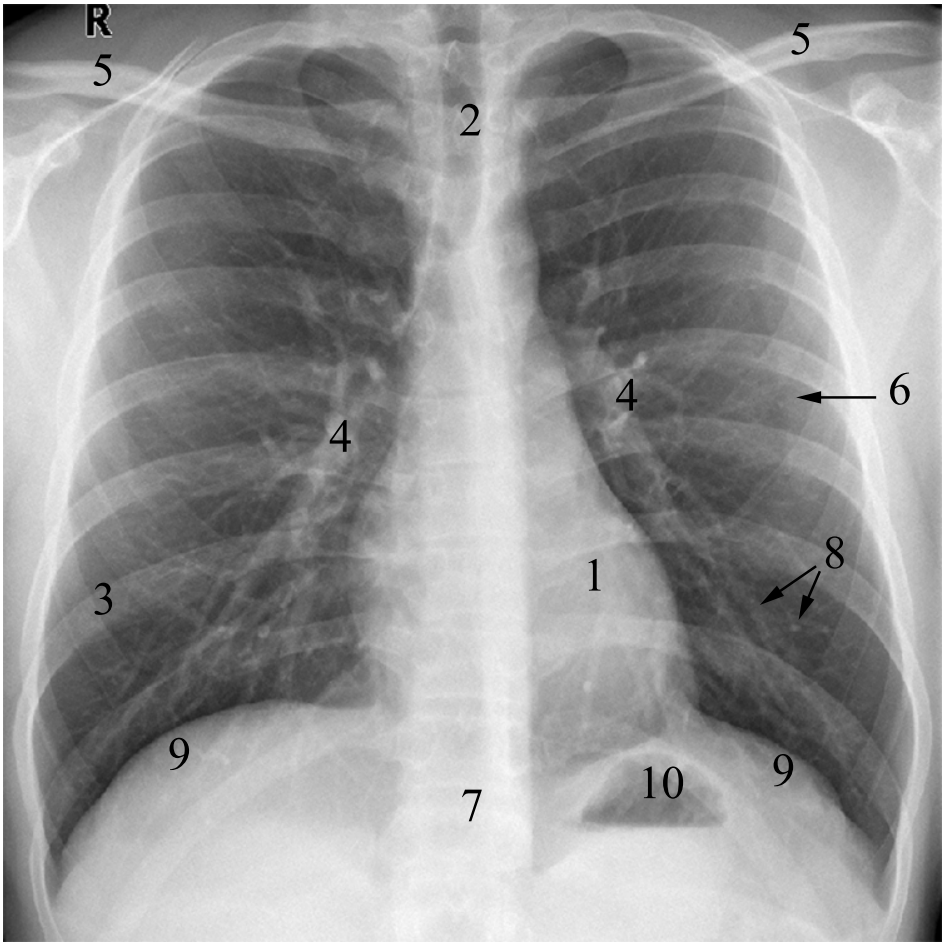
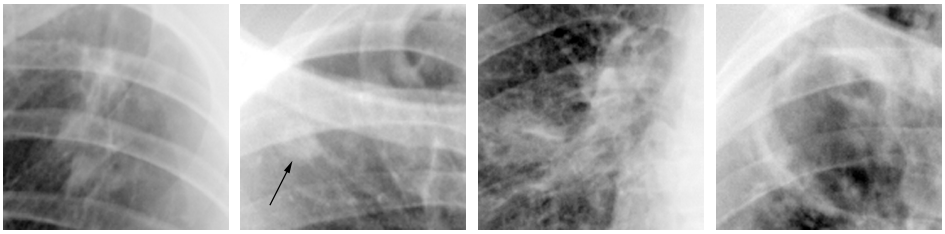
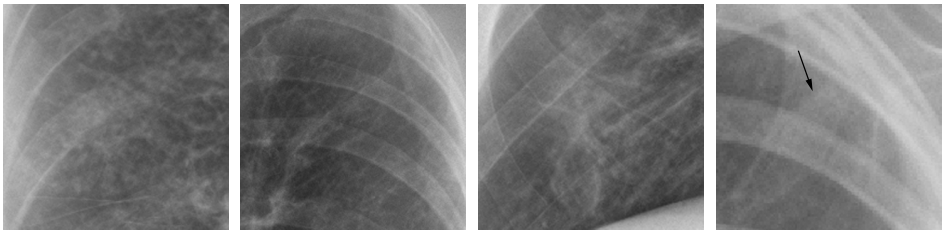


Figure 1.1: *A normal PA chest radiograph. The left side of the image shows the right lung of the patient. The lungs, filled with radiolucent air, show up dark in the image. They are transpierced with blood vessels. Bones that absorb a lot of radiation are the brightest structures in the image. Healthy interstitium should not be visible in the radiograph. Some important anatomical structures are labeled in the image. (1) The heart. (2) The trachea. (3) One of the ribs. The posterior (back) part of the rib is better visible in the PA radiograph than the anterior part. (4) The hilum in the right and left lungs. (5) The clavicles. (6) The arrow indicates the border of the vertical shadow produced by the scapula. (7) The spine. (8) Some blood vessels. The small dot pointed to by one of the arrows is a vessel running in the same direction as the x-rays. (9) The diaphragm. (10) A gas bubble in the stomach (often visible in the radiograph).*



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

Figure 1.2: *Examples of abnormal texture patterns often encountered in radiographs of TB patients. TB abnormalities are often located in the upper lobes (upper thirds of the lungs), like an infiltrate in (a), a nodule under the right clavicle in (b) and a cavity, a darkened areas surrounded by bright margins, in (d). An enlarged hilum, like one in (c), is another radiological finding characteristic for TB.*



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

Figure 1.3: *Examples of abnormal texture patterns encountered in chronic ILD patients. While diffuse and consolidated abnormalities can be quite obvious ((a) and (c)), the examples in (d) and (d) depict more subtle lesions. The line in (b) going from bottom left to top right is an abnormal fibrotic scarring. The arrow in (d) points to a very subtle abnormality hidden behind the rib. This a difficult lesion to spot for radiologists and CAD systems.*

1.1.3 Computed tomography

The second most important landmark in diagnostic imaging was the invention of x-ray tomography scanner by G. Hounsfield in 1972 [11]. The computed tomography scan, or, simply, the CT scan, is the 2D reconstruction of a cross-section of a patient's anatomy rather than a projection of the shadows cast by overlapping organs and bones. During CT scanning, the source of x-ray radiation circles around the patient, and the detectors on the opposite side measure the fraction of the radiation that passes through the body. From these measurements, the cross-section image is reconstructed. Although the earliest mathematical description of a reconstruction method appeared in 1917 [12], the implementation of the method was not possible without the computer. Therefore, computed tomography is an inherently digital imaging modality.

A modern chest CT scan is a stack of 2D cross-section images (slices) obtained at every 0.7-1.0 mm, each of these corresponding to a section of thorax as thin as 1 mm. Such scans are generated by multidetector CT (MDCT) scanners first introduced in 1998 [13]. Nowadays, MDCT scanners can acquire up to 320 transversal slices simultaneously and produce an isotropic picture of the thorax or other body parts in a few seconds. Another key advantage of computed tomography, besides eliminating overlapping structures, is an improved contrast between tissues. Figure 1.4 shows one of transversal slices of a normal thoracic CT scan and two other planar views, coronal and sagittal, that can be easily computed from a stack of transversal slices and are routinely used in clinical diagnostics.

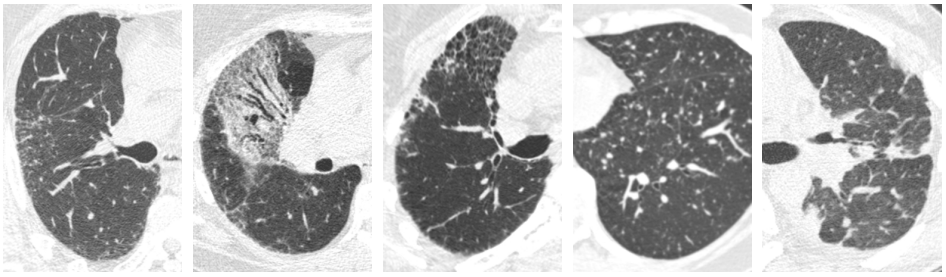
Thin-section CT, such as MDCT, is valuable in detecting ILD in patients for whom the diagnosis is uncertain after chest radiography and clinical assessment [2]. As in convention radiography, interstitial abnormalities in CT look like patches of abnormal texture, only much better defined than in the radiograph. The presence of a certain abnormal texture pattern in the patient's CT, or a combination of several patterns, as well as their spatial 3D distribution in the lungs, were found characteristic for many ILDs, especially those with unknown cause [7]. Some abnormal patterns typical for ILD are illustrated in Figure 1.5.

1.1.4 Computer-aided diagnosis

Radiologists do not detect all abnormalities that are found in retrospective image reviews, and they do not always correctly make a diagnosis based on those abnormalities that they find. The application of computers to medical image interpretation has been investigated since the 1960s. With the increasing power and spread of computers and the advent of digital imaging modalities, the growth of the CAD research area has been tremendous for the last 20 years. In thoracic imaging, studies on CAD include the development of computerized methods for the detection of lung nodules in chest radiographs and CT, detection and classification of ILD, detection of pneumothorax, and temporal subtraction of chest radiographs to detect interval changes. General historical overviews of CAD sys-

(a) (b) (c)

Figure 1.4: *Examples of 2D views taken from normal thoracic (chest) CT. In (a) a transversal, or axial, CT section is depicted. This particular slice corresponds approximately to the center of the chest. The left side of the image depicts the right lung. A modern CT scan consists of several hundreds of transversal slices, from which the other planar views can be easily restored at any level of depth. A coronal view in (b) is a frontal view of the chest. In this example it is restored at the chest level marked by the horizontal line in (a). The patient in this image is facing the viewer. A sagittal view in (c) is a lateral view of the chest. In this example it is restored at the chest level marked by the vertical line in (a). The back of the patient is on the left side of this image.*



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

Figure 1.5: *Examples of abnormal texture patterns typical for different ILDs, as seen in axial CT slices: reticular and linear patterns (a), ground glass opacity with reticular pattern (b), honeycombing (c), nodular pattern (d), and consolidations (e). One may notice that some of these image are quite noisy. This is because they were taken with a decreased amount of radiation (it is called a low-dose CT) to reduce the chance of damage to the patient's tissues. Nevertheless, the quality of such images is good enough for performing ILD diagnostics.*

tems can be found in Refs. [14] and [15]. Refs. [16], [17] and [18] provide specialized reviews on CAD applications for the lungs.

The CAD output is intended to be used as a "second opinion", and radiologists make the final decisions. An underlying assumption is that combining the competence of the radiologist with the capability of the computer system can improve interpretation results. Obviously, the higher the performance of the system, the better an overall effect on the final decisions. However, the performance of the computer doesn't have to be higher or equal to that of the radiologist. An improved performance of radiologists who used CAD in making their decisions was demonstrated for a number of thoracic CAD applications [19, 20, 21, 22]. Several CAD systems for lung nodules detection have become commercially available.

CAD systems can be classified into two types - computer-aided detection and computer-aided diagnosis. The former outputs estimated locations of abnormalities, or simply estimates a likelihood that the image contains anything suspicious. The major area of the utilization of such CAD systems are screening programs, e.g., a TB screening program for high-risk population groups, or lung cancer CT screening for smokers. In screening programs, an enormous amount of images have to be interpreted, with most being normal. This is a time-consuming and error-prone task for radiologists. For already-detected lesions, computer-aided diagnosis systems assist in their characterization, e.g., in estimating the malignancy of lung nodules or the types of ILD.

In this thesis we consider computer-aided detection systems. The basic working principles of such systems are described in the next section.

Principles of CAD systems

From an engineering point of view, CAD systems are based on the combination of sophisticated image processing and pattern recognition techniques. While the image processing part is responsible for extracting useful information from images, the pattern recognition part is responsible for the classification of extracted information. What sort of information should be extracted depends on the task at hand. For detection of interstitial abnormalities, for example, the texture appearance of the lung fields is important, because it distinguishes the normal lungs from abnormal. Therefore, some texture properties should be derived from the patches of texture in the lungs. Whereas for the lung nodules detection, first, the objects resembling nodules (nodule candidates) should be located within the lung fields, and then, their properties should be measured. e.g., their shape and size. When distinguishing properties of the image, or the objects of interest in the image are measured, these measurements are concatenated in a vector, thereby representing each sample in a high-dimensional space. By convention, such a vector is called *a feature vector*, and its elements are called *features*.

The CAD system is often represented as a black box that takes a feature vector as an input and produces a sample label as an output. What happens inside the black box of the CAD systems of this thesis, as well as inside many other CAD

systems, is called *two-class supervised classification*. It uses *a classifier* to make decisions. The classifier is a function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, where p is the dimensionality of the feature space, that is applied to a feature vector $\mathbf{x}$ and predicts the class label of a sample represented by $\mathbf{x}$. The class label can either be a binary decision, such as normal or abnormal, or denote a degree of abnormality on a given scale, such as a probability of being abnormal. The binary decision is obtained by thresholding the function f:

$$h(\mathbf{x}; \mathbf{w}) = \begin{cases} 1 & \text{if } f(\mathbf{x}; \mathbf{w}) > \theta, \\ 0 & \text{otherwise,} \end{cases} \tag{1.1}$$

where $\mathbf{w}$ are the free parameters of the function f. They determine the boundary between two classes in the feature space. The estimation of the free parameters $\mathbf{w}$ precedes the application of the classifier. In the learning, or training, phase, parameters $\mathbf{w}$ get optimized, for instance, by minimizing the classification error on the set of samples whose true class labels are known. Such set of samples is called *a training set*, and the process of obtaining the optimal $\mathbf{w}$ is called *training*. In the classification, or testing, phase, new, previously unseen data are classified using the trained classifier, that is, their class labels are determined by applying functions f or h. Popular classifiers are linear and quadratic discriminant classifiers, nearest neighbor classifiers, support vector machines, neural networks, etc [23].

The choice of a classifier, as well as the choice of features, are two important issues to consider while designing a CAD system for a particular task. In choosing features the goal is to allow feature vectors from different classes to occupy compact, disjoint regions in the feature space. Then, finding the true boundaries between classes is feasible, provided enough training samples are available. However, it is difficult to find good, discriminating features for real-life complex problems. One of the ways is to compute a lot of candidate features that might be useful according to some idea or assumption about a given problem, and then dispose of meaningless ones. Several automated procedures have been developed to select an optimal subset from candidate features, or to compute a reduced set of new features from the initial set. In chapters 3, 5 and 6 of this thesis we apply *principle component analysis* (PCA) to the initial feature set. PCA is a linear transformation that effectively reduces the dimensionality of the feature set by extracting the most expressive features responsible for most of the variance in the training data.

As far as classifiers are concerned, there is no best performing classifier for all problems. In designing a CAD system, one often needs to experiment with a number of classifiers to find the most suitable one for a given task. Some of the classifiers make an assumption on the form of the classes' probability density functions (e.g., linear and quadratic discriminant classifiers), while the others don't (e.g., nearest neighbor classifiers). Linear classifiers, such as the linear discriminant classifier and support vector machine, look for a linear boundary be-

tween classes, while boundaries constructed by nearest neighbor classifiers could be rather complex. For all the CAD systems presented in this thesis, linear classifiers were the best performing ones. Chapter 2 of this thesis investigates a novel linear classifier whose training is based on other principals than the minimization of classification error.

There is a lot more to be said about pattern classification that can be found in textbooks and reviews like [23, 24].

Prerequisites

Two other aspects involved in building the CAD system are data acquisition and preprocessing. They greatly depend on the problem domain. The preprocessing of thoracic images, whether they are conventional radiographs or CT scans, includes segmentation of the lung fields and, optionally, other anatomical structures. Segmentation of the lung fields results in a binary mask image, where the lungs and non-lungs areas are labeled differently. Such a mask restricts computation of features to the lung fields. There exist a considerable number of methods for 2D and 3D lung segmentation. To mention a few, there are active shape models and pixel classification methods for 2D segmentation [25], an automatic rule-based approach using thresholding and regiongrowing for 3D segmentation [26], and an atlas-based segmentation-by-registration approach [27, 28] that is especially useful in segmenting pathological lungs from CT scans. In this thesis, for lung segmentation in chest radiographs, the active shape model algorithm was used in chapters 2 and 3, and the pixel classification method in chapter 6. In chapter 5, the automatic rule-based approach was used in combination with the atlas-based approach to segment the lungs in 3D CT scans.

Image registration is another possible preprocessing step. Registration means alignment of anatomical structures of two different images. When used for segmentation, an atlas, i.e. an image with already known segmentation mask, is registered to a target image. The resulting deformation is applied to the atlas's mask in order to create a mask for the target image. When there is more than one atlas and, subsequently, more than one registration can be computed, the target mask is generated by combining all the masks obtained with available deformations. Another possible utilization of registration is image comparison. Aligning two CT scans enables the comparison of CT sections of matching anatomy. This is used in detecting temporal changes in a patient's successive scans in chapter 5 of this thesis.

The annotation of images, such as the delineation of abnormalities, is a part of a data acquisition process. Since training of the CAD system requires a set of samples with known labels, such labels have to be obtained beforehand. They are called *the reference standard*, or *the ground truth*. The form of the reference standard depends on a given task. For the detection of interstitial abnormalities, an ideal reference standard would be the exact outlines of lesions in the image. Such a reference standard is usually provided by an expert radiologist, or a panel of

radiologists, who manually (with the assistance of a dedicated software) delineate all the lesions they are able to find in the training images. Defuse interstitial abnormalities in chest radiographs typically have ill-defined borders. This impedes their delineation and results in unreliable and irreproducible outlines. Manual delineation is also a tiresome task, especially in 3D. A CAD system in chapter 4 of this thesis employs a superior reference standard for interstitial lesions in chest radiographs which is based on CT findings.

A reduced but more reliable ground truth is an opinion whether the image contains any ILD- or TB-related abnormalities. Such an opinion, given by a radiologist, can be also supported by clinical evidences or outcomes of laboratory tests. Images, for whom only the presence or absence of the disease is known, but not the delineation of specific lesions, are called *weakly labeled data*. In practice, this form of the ground truth is natural and, therefore, easier to obtain, because giving a general opinion about an image is a typical daily task for a radiologist. For example, the data that comes from a TB screening program, are usually accompanied by the reading results of two or more physicians, indicating whether the image is normal or exhibits some suspicious patterns associated with TB. The delineations of suspicious regions are not provided, and a specialized observer study has to be conducted to obtain them. But obtaining exact abnormality outlines may not be possible or practical for the reasons mentioned above. Classification of weakly labeled data is the subject of research in chapters 2, 3 and 6 of this thesis.

Evaluation

When the ground truth is provided for the test data as well, the classification performance of the CAD system can be evaluated. The performance of two-class classification can be assessed either in terms of *accuracy*, *sensitivity* and *specificity*, or by building the *receiver operating characteristic* (ROC) curve and computing the *area under curve* (denoted as either AUC or A_z in this thesis). To explain these measures, let us introduce some terminology first. Normal and abnormal (i.e., diseased) samples are conventionally called *negatives* and *positives*, respectively. With two-class classification, four types of outcome can be distinguished: *true positives*, which are correctly classified normal samples; *false negatives*, which are abnormal samples misclassified as normal; *true negatives*, which are correctly classified abnormal samples; and, finally, *false positives*, which are normal samples misclassified as abnormal.

The CAD system strives for minimizing mistakes, that is, minimizing the total amount of false negatives and false positives. This is equivalent to maximizing a classification accuracy. Accuracy is the fraction of correctly classified samples, i.e., a ratio of the sum of true positives and true negatives to the total number of samples. When a classification system makes no mistakes, the maximum accuracy of 1 is achieved. Sensitivity and specificity are used to characterize how well positive and negative samples are classified. Sensitivity is a fraction of true positives,

that is, a ratio of true positives to the sum of true positives and false negatives. Specificity is a fraction of true negatives, that is, a ratio of true negatives to the sum of true negatives and false positives. The closer sensitivity and specificity are to 1 the better classification performance is.

Accuracy, sensitivity and specificity are suitable performance measures for a classification system that produces binary labels. According to Eq. 1.1, a discrimination threshold θ should be fixed to some value to convert the continuous output of a classifier to the binary one. If the output of the classifier denotes the probability of a sample of being abnormal, a typical value for θ is 0.5. However, this choice of θ is empirical and does not guarantee the highest possible classification accuracy of a particular CAD system. By varying θ and computing corresponding accuracies, one can find an optimal discrimination threshold. If classified samples are ordered correctly, i.e., every positive sample gets a higher classification output than every negative sample, it is possible to find a θ that results in the accuracy of 1.

The area under the ROC curve is another measure of a classifier's performance [29, 30]. The ROC curve is obtained when the true positive fraction is plotted as the function of the false positive fraction. This is equivalent to plotting sensitivity against 1-specificity computed for varying thresholds. Figure 4.4 in chapter 4 shows several ROC curves. AUC characterizes the performance of a classification system independently from a chosen discrimination threshold and possesses other desirable properties, such as invariance to prior class probabilities. AUC gives an indication of how well the negative and positive classes are separated by the classifier. When the negative and positive samples are ordered correctly, AUC equals to 1. As a single-number performance measure, AUC is preferred to accuracy.

1.2 Outline of the thesis

This thesis contributes to the development of pattern classification methods employed by CAD, with the application of these methods to the automated analysis of ILD and TB. We consider three different applications. In **chapters 2, 3** and **6**, we aim at detecting the presence of disease in chest radiographs. The purpose of a CAD system described in **chapter 4** is to pinpoint the locations of abnormalities in radiographs. In **chapter 5**, a CAD system is described that estimates the progression of ILD in follow-up CT scans.

In **chapters 2** and **3**, two novel classification approaches are presented that deal with weakly labeled data. In **chapter 2**, the absence of local ground truth is overcome by assuming that every pixel in an abnormal image is abnormal. Such an assumption usually results in highly overlapping normal and abnormal classes in the feature space. This chapter focuses on training the classifier with such ill-separable data. Instead of minimizing the classification error, AUC is explicitly optimized, which is shown to be a more robust strategy when two classes have a

large overlap. When all pixels in the lung fields are classified, their probabilities of being abnormal are integrated in a single decision about the whole image.

Chapter 3 describes an entirely different approach to handling weakly labeled data. We assume that the distributions of local texture measurements are different for normal radiographs and radiographs with diffused abnormalities. Dissimilarity-based features that estimate these differences per measurement are proposed as an image representation. A number of CAD systems can be trained, each with image representations obtained from comparison with a different training image. Subsequently, a test image can be classified a number of times. A final image decision is obtained by combining the results of all such classifications.

In **chapter 4**, a new method to set a reference standard for interstitial abnormalities in chest radiographs is described. Abnormality outlines are manually delineated in selected coronal slices of a thoracic CT scan and automatically mapped to a radiograph of the same patient. To train a CAD system, lung pixels that fall inside delineated areas are considered positives, while those fallen outside are considered negatives. The CAD system performs pixel-wise classification of the lung fields and produces a color-coded probability map accentuating areas highly probable of being abnormal.

Chapter 5 presents a CAD system for automated estimation of ILD progression in serial thoracic CT scans. The system compares corresponding 2D axial sections from the baseline and follow-up scans and yields an opinion whether this pair of sections represents regression, progression or unchanged disease. The CAD system performs classification in two stages. In the first stage, image pairs exhibiting any change in the state of disease are separated from unchanged cases. In the second stage, the direction of an estimated change is classified into regression or progression. Different features are exploited in each classification stage.

In **chapter 6**, another application that uses the approach described in **chapter 3** is considered: the analysis of radiographs from TB mass screening programs. The dissimilarity-based approach is extended by applying it to fixed lung partitions, as well as to the complete lung fields, and merging the local and global classification results into a single image decision. The CAD system yields a probability for an image to contain TB-related abnormalities.

Chapter 7 is the final chapter that provides a summary and general discussion.

Chapter 2

Optimization of the Area under the ROC curve, with an application to the detection of interstitial lung disease in chest radiographs

D.M.J. Tax, Y. Arzhaeva, R.P.W. Duin and B. van Ginneken, "AUC optimization by subsampling constraints," *in preparation.*

Abstract

The area under the ROC curve (AUC) is a popular performance measure for classifiers because it is insensitive to class imbalance and skewed misclassification costs. Unfortunately, AUC optimization is difficult for large sample size problems and most subsampling methods for optimizing the AUC are ill-suited for the situations in which optimizing the AUC is preferable over the classification error, i.e. when the classes are highly overlapping and imbalanced. We propose a classifier that optimizes the AUC using a linear programming formulation. This formulation has the advantage that the large set of classification constraints can easily be subsampled, without significantly decreasing the stability of the result. Furthermore, the non-used constraints can be used to optimize the complexity parameter of the classifier. In a set of experiments the good performance of this classifier is shown and compared with other methods.

2.1 Introduction

In medical screening programs, or in detection problems in general, one has to deal with unknown class priors or misclassification costs. For these problems, the Area under the ROC curve (AUC) is a better performance measure than the classification error [30, 31, 32, 33]. Not only is it insensitive to class priors and costs, it also appears to be more stable for small sample sizes, making it a better measure to compare different classifiers than accuracy [33].

In the optimization of the AUC one focuses on the correct ordering of objects from two classes, often called the positive and negative class. Objects from the positive class should get a higher classifier output than objects from the negative class. When the classifier outputs of the positive objects have the same distribution as the outputs of the negative objects then both classes overlap and cannot be separated. On the other hand, when all positive objects have a higher output than all the negative objects, the two classes are fully separable. The AUC directly represents the fraction of object pairs that are correctly ordered. Optimizing the AUC therefore minimizes classification error but because it uses the object pairs instead of the class distributions, it is insensitive to class priors or misclassification costs.

Several methods are proposed for optimizing the AUC, inspired by, for instance, boosting algorithms [34], decision trees [35] or support vector classifiers [36]. In [37], a simplified linear version of the classifier in [36] is presented. This classifier is simpler to optimize than the original version from [36] because it reduces the quadratic optimization problem to a linear programming problem. Furthermore, it is sparse in the features such that it can perform well in high dimensional feature spaces.

Although the optimization of the AUC seems very attractive from the theoretical point of view, in practice there are two important factors that limit the widespread use of the AUC optimization. The first factor is that the practical performance increase is often small, compared to the minimization of classification error. The AUC results in [38, 36, 39] show that for many data sets it does not pay off to explicitly optimize the AUC. It is shown in [40] that the expected value for the AUC is linearly related to the classification error. It appears that the advantage of the AUC over the classification error is that for imbalanced and highly overlapping classes the variance of the AUC estimate is much smaller. This higher stability of the AUC is very valuable when classifiers are evaluated in noisy and low sample size situations.

The second factor that limits the applicability of the AUC is that it is computationally a hard optimization problem. For the optimization of the AUC one has to consider all pairs of positive and negative objects. For each pair a constraint is introduced that enforces the correct ordering of these two objects. The complexity increases therefore from the sample size n, for a standard classification problem, to $n^+ \cdot n^-$ for the AUC optimization, where n^+ and n^- are the number of positive and negative training examples respectively. Several heuristics have

been proposed to alleviate the problem. A successful one is proposed in [36] where a k-nearest neighbor approach is used. For each training object the nearest object from the other class is selected and all other objects are ignored. Results show that this is a very successful approach, in particular for clearly separable classes. Unfortunately, for heavily overlapping classes the nearest neighbors are not very informative and this subsampling method does not perform well.

In section 2.2 we discuss situations when the AUC optimization could be preferred to the minimization of the classification error. Then, we show the formulation of the AUC optimization that uses random subsampling of *constraints*, instead of subsampling of objects. This makes the method efficient in cases when the AUC optimization is advantageous. Furthermore, the discarded constraints can be used to optimize a free complexity parameter (this complexity parameter determines the trade-off between the sparsity of the solution, and the error on the training set)[1]. In section 2.4 we compare the proposed classifier with other methods and we finish with conclusions in section 2.5.

2.2 The use of optimizing the AUC

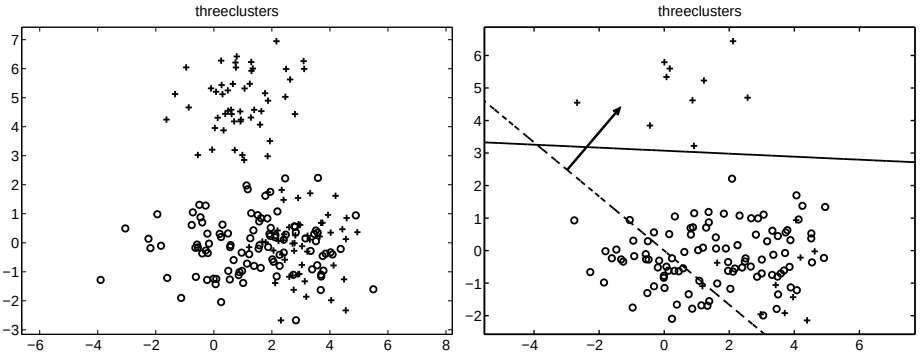


Figure 2.1: *Artificial data set containing strong clustering characteristics and class overlap. The left figure shows a scatter plot using 100 objects per class. In the right subplot the two class sizes are 100 and 20.*

It is often observed that the explicit optimization of the AUC is in many cases not significantly better than optimizing the classification performance [38, 36, 39]. In [40] it is shown that the expected value of the AUC for a fixed error rate is a direct function of the error rate[2]. However, the analysis of the variance shows

[1] A preliminary paper with this idea of subsampling the constraints was first presented in [37]. In this chapter we analyzed why this method is expected to perform better and we extend the formulation by incorporating the automatic optimization of the complexity parameter.

[2] Two classifiers with identical classification errors can have different ROC curves. When one fixes the classification error, one can integrate over all ROC curves that are consistent with this

that the AUC estimation has smaller variance than the classification error, in particular in the case that the classes are overlapping and imbalanced. To clarify this, an artificial data set is constructed to show these characteristics.

In the left subplot of Figure 2.1, a scatter plot of an artificial data set is shown, that incorporates the features that make the minimization of classification error not suitable. The data set contains two classes. The first class contains two Gaussian clusters of equal size, centered at $(3, 0)$ and $(0, 0)$. The second class also contains two Gaussian clusters, but centered at $(3, 0)$ and $(1, 5)$. Most standard classifiers tend to focus on the gap between the upper and two lower clusters. In particular when one of the classes is severely undersampled and the misclassification costs are not adjusted accordingly, the few overlapping objects in cluster $(3, 0)$ can be ignored. For standard classification this signifies that the classifier is robust against noise and outliers. In the case that these objects convey important information on the class distribution, these objects cannot be considered noise and cannot be ignored in the optimization of the AUC.

This is shown in the right subplot. The class sizes are $n^- = 100$ and $n^+ = 20$. The straight solid line is the decision boundary obtained by the linear support vector classifier [41]. The resulting AUC for this classifier is 0.61. Without the adjustment and tuning of the trade-off parameter C for each of the classes, the classifier considers the few erroneous objects to be noise. The dashed line on the other hand indicates the solution found after optimizing the AUC. Here the objects in the overlapping clusters are not considered noise and they influence the solution. By that, the AUC for this type of classifier is improved to 0.84.

We can summarize that the optimization maybe particularly useful when a few objects of one class are 'hidden' in a much larger group of objects from the other class, and these objects are still important to detect for an optimal AUC. For classification problems with low class overlap and/or small class imbalance, it may not be advantageous to perform the AUC optimization. It is sufficient to optimize a standard classifier minimizing the classification error. For many of the relatively simple UCI data sets [42] this may be the case. On the other hand, many real world problems indeed suffer from class imbalance and overlap, making the AUC optimization still a worthwhile effort.

Therefore, a good AUC optimizer should be robust against high class overlap and (large) class imbalance. Furthermore, for its practical application it is also desirable that the number of free parameters that has to be set by the user is reduced, and that it is robust against high-dimensional feature spaces (or, that it can perform feature selection automatically). In the next section we propose an L_1 AUC optimization formulation, and a new strategy to handle large data sets.

classification error and find the average AUC. It appears that this is AUC is exactly one minus the classification error.

2.3 L_1 AUC optimization

Assume that we have a data set $\mathcal{X}^{tr} = \{(\mathbf{x}_i, y_i), i = 1, ..., N\}$ with two classes, a positive and a negative class, where each object is represented by p features $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$. The classes contain n^+ and n^- objects respectively. It is assumed that index k^+ runs over all positive objects (with $y_i = +1$) and k^- over all negative objects (with $y_j = -1$).

In the traditional approach to learning, a classifier is defined to predict from a new object $\mathbf{x}$ its label, by applying a function $f : \mathbb{R}^p \to \mathbb{R}$ to the object, and thresholding this function output by threshold θ to obtain a class label:

$$h(\mathbf{x}; \mathbf{w}) = \begin{cases} +1 & \text{if } f(\mathbf{x}; \mathbf{w}) > \theta, \\ -1 & \text{otherwise,} \end{cases} \tag{2.1}$$

where $\mathbf{w}$ are the free parameters of function f. Notice that by varying the value for θ, the ROC curve is obtained [29].

In most cases the classifier h is optimized by optimizing the parameters $\mathbf{w}$ in f. This can be done for instance by minimizing the (apparent) classification error:

$$\hat{e}_{emp}(h, \mathcal{X}^{tr}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(h(\mathbf{x}; \mathbf{w}) \neq y_i) \tag{2.2}$$

where $\mathbf{1}(.)$ is the indicator function. Instead of the indicator function, the error can be estimated by minimizing the mean squared error. In the definition of these errors, the contribution of each object is counted independently of the other objects in the training set.

An alternative performance measure is the Area under the ROC curve (AUC). This measure counts how often an object of class $+1$ ($\mathbf{x}_+$) is ranked higher than an object of class -1 ($\mathbf{x}_-$):

$$AUC = Pr(f(\mathbf{x}_+) > f(\mathbf{x}_-)). \tag{2.3}$$

Clearly, a perfect separation of the two classes is obtained when $AUC = 1$. The AUC performance can be simply estimated on a finite data set $\mathcal{X}^{tr}$:

$$\hat{AUC}(h, \mathcal{X}^{tr}) = \frac{1}{n^+ n^-} \sum_{k^+=1}^{n^+} \sum_{k^-=1}^{n^-} \mathbf{1}\left(f(\mathbf{x}_{k^+}) > f(\mathbf{x}_{k^-})\right). \tag{2.4}$$

One of the classifiers that directly optimizes the AUC is the support vector machine as defined in [39]. It minimizes the L_2 norm of $\mathbf{w}$ with constraints on the ordering of the objects. The optimization problem is defined as follows:

$$\min \|\mathbf{w}\|^2 + C \sum_{k^+} \sum_{k^-} \xi_{k^+k^-} \tag{2.5}$$

$$\text{s.t.} \quad \forall k^+, k^- : f(\mathbf{x}_{k^+}) - f(\mathbf{x}_{k^-}) \geq 1 - \xi_{k^+k^-}, \quad \xi_{k^+k^-} \geq 0. \tag{2.6}$$

The constraints (2.6) express the pairwise differences between objects from different classes $f(\mathbf{x}_{k^+}) - f(\mathbf{x}_{k^-})$. The slack variables $\xi_{k^+k^-}$ in these constraints approximate the indicator function $\mathbf{1}(.)$ that is part of (2.4). This approximation has the drawback that the number of constraints is quadratic in the number of objects, so it becomes very large.

The optimization problem can be cast into a linear programming form, by replacing the L_2 norm by the L_1 norm, (similar to the L_1-SVM, [43] or [44]) and by using a linear function f: $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T\mathbf{x}$. In [38] the L_1 formulation was used such that large data sets can be considered.

$$\min \|\mathbf{w}\|_1 + C \sum_{k^+} \sum_{k^-} \xi_{k^+k^-} \tag{2.7}$$

$$\text{s.t.} \quad \forall k^+, k^- : \mathbf{w}^T(\mathbf{x}_{k^+} - \mathbf{x}_{k^-}) \geq 1 - \xi_{k^+k^-},\ \xi_{k^+k^-} \geq 0. \tag{2.8}$$

This can easily be recast in a linear programming formulation:

$$\min \sum_i (u_i + v_i) + C \sum_{k^+} \sum_{k^-} \xi_{k^+k^-} \tag{2.9}$$

$$\begin{aligned} \text{subject to:} \quad & \forall k^+, k^- :\ (\mathbf{u}^T - \mathbf{v}^T)(\mathbf{x}_{k^+} - \mathbf{x}_{k^-}) \geq 1 - \xi_{k^+k^-}, \\ & \xi_{k^+k^-} \geq 0, \quad \forall i :\ u_i \geq 0,\ v_i \geq 0. \end{aligned} \tag{2.10}$$

The threshold θ is not defined, but can be derived when the misclassification costs and class priors have been supplied. We refer to this as the optimized AUC linear programming classifier, or AUC-LPC classifier.

This AUC-LPC is simple to kernelize, similar to [38]. For this, we have to assume that we can write $\mathbf{u}^T\mathbf{z} = \sum_k \alpha_k \mathbf{x}_k^T\mathbf{z}$ and $\mathbf{v}^T\mathbf{z} = \sum_k \beta_k \mathbf{x}_k^T\mathbf{z}$, where k runs over all training objects. This results in:

$$\min \sum_i (u_i + v_i) + C \sum_{k^+} \sum_{k^-} \xi_{k^+k^-} \tag{2.11}$$

$$\begin{aligned} \text{subject to:} \quad & \forall k^+, k^- :\ \sum_k (\alpha_k - \beta_k) \cdot (K(\mathbf{x}_{k^+}, \mathbf{x}_k) - K(\mathbf{x}_{k^-}, \mathbf{x}_k)) \geq 1 - \xi_{k^+k^-}, \\ & \xi_{k^+k^-} \geq 0, \forall i :\ u_i \geq 0,\ v_i \geq 0. \end{aligned} \tag{2.12}$$

2.3.1 Subsampling the constraints

A serious problem is that the number of constraints in (2.6) and (2.8) is quadratic in the number of objects, or more precisely n^+n^-. To cope with this, different strategies have been proposed.

The first strategy, as is used in [38], is by training the classifier in batches ('chunks'), retaining the problematic pairs from each of the batches. In practice, it is still a computationally expensive procedure, and the authors of [38] are not very content with it either. The second strategy, that the authors then propose, is to start by *randomly drawing objects* from both classes, and to iteratively update

this set by considering the objects that are violating many constraints. In this case, only very hard to classify objects are considered. To avoid that the classifier flips its labels because only objects that have violated constraints are included, the training set is extended to include some well-ordered object pairs as well.

In [36], the third subsampling strategy is suggested, inspired by [45] and [46]. There, only the *objects and their nearest neighbors from the other class* are considered. The user has to define a number m indicating how many neighbors are assumed to be informative. Then, only constraints between each object $\mathbf{x}$ and m of its nearest neighbors are imposed. This subsampling heuristic can fail when there is a large class overlap and the number of neighbors is not sufficiently large. A similar approach is taken in [47], where the data are clustered, and the cluster centers are used as positive and negative examples. It has similar characteristics to using the nearest neighbor approach, for it selects or generate objects in the feature space.

In this chapter, we suggest the fourth approach that avoids subsampling objects. By utilizing the primal formulation (2.6) or (2.8), it is very simple to *randomly subsample the constraints* in (2.8). This random sampling with M constraints avoids focusing on the local structure in the data (as given by the m nearest neighbors or clusters), but characterizes the structure of the constraints on a larger scale. Therefore, the approximation is less biased.

Unfortunately, when constraints are randomly subsampled from (2.6), its dual formulation is still very large and the kernel matrix is still of size $n^+n^- \times n^+n^-$. Only the elements that correspond to the object pairs $\mathbf{x}_{k^+}, \mathbf{x}_{k^-}$ that are not considered will be zero. In practice the optimization of the dual still becomes impractical. The L_1 variant, on the other hand, does not suffer from that (see equation (2.12)), and, therefore, we will focus on the L_1 version in this chapter.

More insight can be gained by considering the difference vectors, representing the constraints in equation (2.8):

$$\tilde{\mathbf{x}}_{k^+k^-} = \mathbf{x}_{k^+} - \mathbf{x}_{k^-}. \tag{2.13}$$

For each constraint a single object $\tilde{\mathbf{x}}$ is defined. For a given weight vector $\mathbf{w}$, a constraint k^+k^- is satisfied when:

$$\mathbf{w}^T \tilde{\mathbf{x}}_{k^+k^-} \geq 1. \tag{2.14}$$

The constraints can be represented by points with the same dimensionality as the original objects. More remarkably, the linear classifier $\mathbf{w}$ in the original feature space can directly be plotted in this "difference" space. Equation (2.14) shows that the plane (passing through the origin), for which the highest number of objects has a projection larger than 1 on the normal of this plane, is the optimal AUC plane. The normal of this plane $\mathbf{w}$ is the ranking direction in the original feature space.

In Figure 2.2 the original artificial data set from section 2.2 is scattered (the left subplot), together with a scatter plot of the difference vectors (the right subplot).

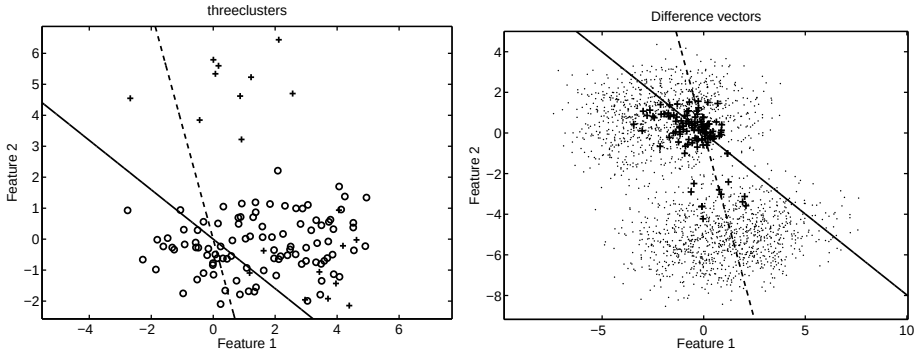


Figure 2.2: *On the left, The scatter plot of a data set containing two overlapping classes, indicated by the circles and the crosses. An AUC-LPC using the k-nearest neighbor heuristic is plotted using a dashed line, an AUC-LPC using the subsampling heuristic is plotted with a solid line. On the right, the scatter plot of all the difference vectors where each dot represents one constraint* $\tilde{\mathbf{x}}_{k^+k^-}$*, see text. The pluses indicate the difference vectors selected by the 1-nearest neighbor approach. The dashed line is obtained by using the constraints given by the pluses. The solid line is obtained using all constraints.*

When all the difference vectors are used, the AUC-LPC defines a $\mathbf{w}$ perpendicular to the solid line in the left subplot. The AUC-LPC using the 1-nearest neighbor subsampling results in the (suboptimal) dashed line. The overlapping classes result in a relatively poor sampling of the constraints (the lower cluster in the right subplot), and it is reflected in the resulting AUC. The first (optimal) classifier has AUC=0.84, while the second only obtains AUC=0.61.

2.3.2 Constraint subsampling or object subsampling

A full analytic analysis of the difference between constraint subsampling and object subsampling is complicated because it involves unknown distributions for the two classes, and the summation over the violated constraints (2.14). Instead, we consider how much the distribution of the difference vectors changes when the original objects $\mathbf{x}_{k^+}, \mathbf{x}_{k^-}$ are subsampled instead of the difference vectors $\tilde{\mathbf{x}}_{k^+k^-}$, for some given class distributions. The essential observation is that by subsampling the objects instead of the constraints, one does not obtain independent constraints. This harms the convergence of the estimates on the constraint distribution. In particular, the estimates of the variance converge much slower when objects are subsampled than when the constraints are subsampled. This results in a less stable estimate for the AUC.

Assume we have two random variables X^+ and X^-, distributed according to $p^+(\mathbf{x})$ and $p^-(\mathbf{x})$, respectively. We define a new random variable, the difference

between the variables $Z = X^+ - X^-$. The distribution of Z is now given by:

$$p(\mathbf{z}) = \int p^+(\mathbf{x})p^-(\mathbf{x}-\mathbf{z})d\mathbf{x} \tag{2.15}$$

When X^+ and X^- are normally distributed,

$$p^+(\mathbf{x}) = \mathcal{N}(\mathbf{x};\mu,\sigma^2), \qquad p^-(\mathbf{x}) = \mathcal{N}(\mathbf{x};\nu,\tau^2), \tag{2.16}$$

the distribution of the difference vectors can be determined exactly by using (2.15):

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mu-\nu,\sigma^2+\tau^2). \tag{2.17}$$

Given a sample $\{\mathbf{x}_i^+\}, i = 1,..,qn^+$ from p^+, $0 < q < 1$, the maximum likelihood estimates for the first and second moments are $\widehat{\mu^+} = \frac{1}{qn^+}\sum_i^{qn^+}\mathbf{x}_i^+$ and $\widehat{\sigma^2} = \frac{1}{qn^+-1}\sum_i^{qn^+}(\mathbf{x}_i^+ - \widehat{\mu^+})^2$. The variance of this variance estimate is:

$$< \mathrm{var}(\widehat{\sigma^2}) >= \frac{(qn^+-1)^2}{(qn^+)^3}m_4^+ - \frac{(qn^+-1)(qn^+-3)}{(qn^+)^3}\widehat{\sigma^2} \sim O(\frac{1}{qn^+}) \tag{2.18}$$

where m_4^+ is the estimate of the fourth order moment of p^+. Similarly, $< \mathrm{var}(\widehat{\sigma^2}) >$ can be assessed.

When a fraction $0 < q < 1$ from (2.16) is sampled, the estimates of the variances converge with $\frac{1}{qn^+}$ and $\frac{1}{qn^-}$, making the total estimate for $\sigma^2+\tau^2$ converging with $\frac{1}{\min(qn^+,qn^-)}$:

$$< \mathrm{var}(\widehat{\sigma^2}+\widehat{\tau^2}) >\sim O(\frac{1}{\min(qn^+,qn^-)}) \tag{2.19}$$

On the other hand, when $\sigma^2+\tau^2$ is estimated directly from (2.17) using the $q^2n^-n^+$ samples, it converges with:

$$< \mathrm{var}(\widehat{\sigma^2+\tau^2}) >\sim O(\frac{1}{qn^+qn^-}) \tag{2.20}$$

The estimate (2.20) therefore converges much faster than (2.19), showing that the subsampling in (2.17) is much more efficient than in (2.16).

To verify that a higher variance in the variance estimates of difference vectors lead to a poorer estimate of the AUC, we simulate it using an artificial, 2D banana-shaped data set. This data set contains only $n^- = n^+ = 50$ objects, and therefore, the full sum in (2.4) can be computed. For varying fractions q, the ratio between the estimated AUC and the sum (2.4) is computed. The estimated AUC is obtained by subsampling q objects from p^+ and p^-, or by subsampling q^2 from (2.15).

In Figure 2.3 the resulting two ratios are shown, averaged over 50 runs. The picture shows that the bias in this AUC estimate is small for both sampling approaches. The variances of the two approaches differ significantly though. The AUC estimate using random object sampling has a significantly larger variance than that using the constraint sampling.

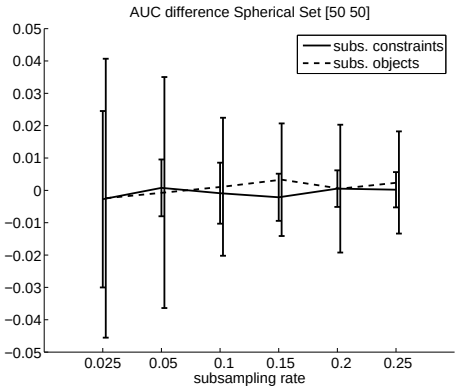


Figure 2.3: *The relative deviation from the true sum* (2.4) *for the estimate using the subsampled constraints (solid line) and the subsampled objects (dashed line). The results are averaged over* 50 *runs.*

2.3.3 Using the unused constraints for the optimization of C

By considering a random sample of the constraints for the optimization leaves a (large) set of constraints unused. The AUC is an estimate of the fraction of constraints that are satisfied (see equation (2.4)). The remaining constraints can therefore be used to evaluate the model. Although these data are not independent of the training data (or training constraints), they give an indication of the suitability of C in (2.7), without using extra validation data. When M constraints are used in the optimization of (2.8), the fraction of satisfied constraints can be estimated by:

$$q(\mathbf{w}; \mathcal{X}^{tr}, M) = \frac{1}{n^+n^- - M} \mathbf{1}\left(\sum_{\overline{\{k^+,k^-\}}} \mathbf{w}^T(\mathbf{x}_{k^+} - \mathbf{x}_{k^-}) > 1 \right) \tag{2.21}$$

where $\overline{\{k^+, k^-\}}$ is the set of all constraints that are not used in (2.8).

In order to test this approach, we estimate the AUC performance on an independent test set and compare this with the number of constraint violations among the remaining constraints. In figure 2.4 the AUC performance on an independent test set is shown as function of the trade-off parameter C (indicated by the solid line). The data set is the same as used in section 2.2, with $n^- = 100$ and $n^+ = 20$. The AUC-LPC is fitted using $M = 500$ constraints. The fraction of satisfied training constraints depends on the class overlap, and, in this case, it is around 0.84 (indicated by the dotted line). The dashed line shows the fraction of satisfied unused constraints. The figure shows that this fraction fits well with the AUC on the test set. The fraction of satisfied constraints on the training set, on the other hand, is positively biased. For this example the optimum fraction of

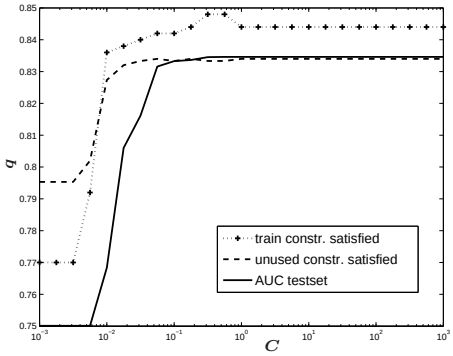


Figure 2.4: *The AUC performance on an independent test set (solid line), the fraction of satisfied unused constraints q (dashed line) and the fraction of satisfied training constraints q (dotted line) as function of C.*

satisfied constraints is obtained for $C = 0.5$, resulting in an AUC of around 0.83. Although the estimate may suffer from some noise, the shape of the curve still allows the optimization of C.

2.4 Experiments

In this section we compare first the AUC-LPC with other classifiers on some artificial data set, then on a larger set of standard classification problems, and, finally, in a real world application.

2.4.1 Artificial data set

In Table 2.1, the performances of some classifiers are compared on the artificial data shown in Figure 2.1. We compare standard supervised classifiers, such as the linear discriminant analysis (LDA), quadratic discriminant (QD) [23], the support vector machine using the L_1 norm (L_1 SVM) [48], the standard (linear) support vector machine using the L_2 norm (L_2 SVM) [41], and a classifier called Least Errors in Sparse Subspaces (LESS) [49]. These classifiers are compared to classifiers that optimize the AUC directly. These are the Rankboost algorithm [34], the AUC-LPC with the full set of constraints, the AUC-LPC with the nearest neighbor heuristic and the k-means heuristic, and the AUC-LPC with the random subsampling heuristic and the optimized parameter C.

In the third column the results are shown for the balanced case. Simple classifiers, like the LDA or QD, correctly classify only the pure clusters, and obtain AUC performances around 0.66. More advanced classifiers, like both SVM classifiers and the AUC optimizers, do not ignore the overlapping clusters and

Table 2.1: *AUC performances (×100) on the artificial data shown in Figure 2.1. Results are averaged over 10 runs, the standard deviation is given in parenthesis. The parameter C is fixed to $C = 10$, except for the last two classifiers, where C is optimized on the remaining constraints. Results in bold are the best performances, and the performances that are not significantly worse than the best (tested using a single sided t-test with a confidence of $\alpha = 0.05$).*

classifier		balanced 100:100	imbalanced 100:20
LDA		66.0 (3.8)	61.7 (4.3)
QD		70.8 (2.2)	71.4 (2.2)
linear L_1 SVM	$C = 10$	**84.4 (0.2)**	79.2 (0.6)
linear L_2 SVM	$C = 10$	**84.0 (0.7)**	79.2 (0.6)
LESS	$C = 10$	**84.4 (0.2)**	78.5 (1.5)
Rankboost	$B = 500$	**85.0 (0.7)**	**83.0 (1.7)**
AUC-LPC full	$M = n^- n^+$	**84.5 (0.4)**	**84.0 (0.6)**
AUC-LPC k-nn	$k = 1$	83.1 (1.8)	74.4 (7.5)
AUC-LPC k-nn	$k = 5$	83.9 (0.8)	77.4 (5.2)
AUC-LPC k-means	$k = 10$	75.3 (0.5)	74.7 (0.5)
AUC-LPC k-means	$k = 25$	**83.8 (1.5)**	**83.3 (1.6)**
AUC-LPC subsamp.	$M = n^- + n^+$	**84.5 (0.4)**	**83.8 (0.8)**
AUC-LPC subsamp.	$M = 5(n^- + n^+)$	**84.5 (0.4)**	**83.8 (1.0)**
AUC-LPC optim. C	$M = n^- + n^+$	**84.4 (0.6)**	83.6 (0.9)
AUC-LPC optim. C	$M = 5(n^- + n^+)$	**84.4 (0.5)**	**83.9 (0.8)**

obtain a significantly higher AUC. The cluster overlap causes the k-nn and k-means subsampling approaches to represent the constraint distribution poorly (as shown in Figure 2.2). The good results of the Rankboost indicate that a non-linear decision boundary may be preferred for this data set. The last column shows the results when the class balance is skewed from 100 : 100 to 100 : 20. The SVM classifiers fall back to an AUC of less than 0.8. The AUC optimizers are stable in their performances maintaining the AUC at around 0.84.

2.4.2 Standard UCI data sets

To compare the different approaches in optimizing the AUC for a wider variety of data sets, the classifiers are applied to a set of problems, listed in Table 2.2, obtained from the UCI repository [42]. The data sets vary in the number of objects, number of features and class imbalance. All problems are reduced to two-class classification problems by selecting one class (given in Table 2.2) and combining all other classes. All features were rescaled to unit variance.

In Tables 2.3 and 2.4, AUC performances are given for a large set of classifiers.

Table 2.2: *Summary of the used data sets.*

name	data set	n^+-n^-	p	class
heart	Heart Cleveland data set	139-164	13	Diseased
biomed	Biomedical data set	127-67	5	Ill
auto	Automobile database	88-71	25	Risk rating > 0
abal	Abalone data set	1407-2770	10	1-8
thyr	Thyroid data set	191-3581	21	Hyperfunctioning
glass	Glass identification	76-138	9	Building non-float
diab	Diabetes data set	500-258	8	Presence
ecoli	Ecoli data set	52-284	7	Periplasm
colon	Colon gene expression [50]	22-40	1908	Tumor
leuk	Leukemia gene expression [51]	25-47	3571	Cancer

The first four classifiers are standard classifiers: the LDA, the QD, the Parzen density estimator (where the width parameter is optimized using a leave-one-out estimation [52]), and the 1-nearest-neighbor classifier (where the class posterior probability is inversely proportional to the distance to the nearest neighbor of that classe). Next, there are the L_1 and L_2 SVM, both with the linear kernel. For the L_2 SVM, the LIBSVM implementation is used [53]. The L_1 SVM and the AUC-LPC use a linear programming optimizer, GLPK [54]. After that, the Rankboost is shown, with a varying number of weak learners B.

Finally, the AUC-LPC is applied with different settings. In "AUC-LPC full", all constraints are used. In "AUC-LPC obj.", training objects are randomly subsampled. And in "AUC-LPC k-nn", constraints are subsampled by only considering the k nearest neighbors from the other class (as mentioned in section 2.3.1). Random subsampling results are listed with "AUC-LPC subsamp.". Then, the results for the AUC-LPC using the optimization of C are shown. To obtain the optimal C, 25 different AUC-LPC's with 25 values of C between 10^{-3} and 10^{+3} are trained. The AUC-LPC with the lowest number of violated constraints is then used as the optimized AUC-LPC. The last lines of Tables 2.3 and 2.4 show the results for the kernelized AUC-LPC that uses an RBF kernel. The kernel width parameter is optimized using 5-fold cross validation, where 25 values between $\sigma = 0.1 d_{1NN}$ and $\sigma = 50 d_{1NN}$ are tried, where d_{1NN} is the averaged first nearest neighbor distance in the training set. For the support vector machines and the AUC-LPC methods (except for "AUC-LPC optim. C"), the parameter C is set to 10.

Among data sets in Table 2.2, two different types of classification problems can be distinguished roughly: the first, where linear classifiers perform very well already, and the second, where non-linear classifiers are required. For the latter problems (represented by auto, thyr, glass, and ecoli), the nearest neighbor classifier often performs very well, followed by the Rankboost algorithm. The

other classifiers, like the SVM and AUC-LPC, are not flexible enough, and, therefore, their kernelized variants have to be used. This requires the optimization of extra kernel parameters, which results, in almost all cases, in overfitted classifiers. Consequently, their performances are not better (except for `ecoli`) then the results of the 1-nn or Rankboost.

Comparing the linear classifiers, the AUC-LPC and the L_2 SVM are often very close in performance, although there are situations where the AUC-LPC is considerably better (for instance, `thyr` and `glass`). The standard LDA is often much worse than more sophisticated linear classifiers.

The AUC-LPC with the full set of constraints does not significantly outperform the AUC-LPC with randomly subsampled constraints for most of the data sets. When the default value of $C = 10$ is not optimal for a data set, which is the case for the high dimensional data sets `colon` and `leuk`, the AUC-LPC with optimized C outperforms the AUC-LPC using all constraints but with the fixed $C = 10$. For the other data sets, the default C give satisfactory results. Furthermore, the AUC-LPC that randomly subsamples objects is, practically, always significantly worse than the AUC-LPC subsampling the constaints (except for, possibly, `colon`). The AUC-LPC that randomly subsamples the constraints is a bit better than the AUC-LPC that uses the k-nn subsampling.

In most cases the variances in the results of different AUC-LPC classifiers are similar, with the exception of the "AUC-LPC obj." that exhibits a larger standard deviation for all the data sets. One may also notice that a very large standard deviation appears in the results of the Rankboost algorithm for `auto` data set. All of the individual features in the `auto` data set have large class overlap, and combined with the relatively small sample size this results in a very unstable performance. a huge standard deviation explains why the performances of the Rankboost with this data set (an AUC of 0.771 and 0.778) are still not significantly different from that of the 1-Nearest neighbor classifier (an AUC of 977). It means that the results of many runs of the algorithm were very good, but there were a few runs where the method completely collapsed. This results in a huge variance in the performance, but the method is finally not significantly worse than the best method.

Finally, the results in Tables 2.3 and 2.4 show that the AUC optimizers perform better than the standard classifiers on highly unbalanced data, such as `thyr`. But the standard classifiers can be equally good or better than the AUC optimizers for balanced and mildly unbalanced data sets.

In Tables 2.5 and 2.6, the training times for all methods are listed. The "AUC-LPC full" is the most computational expensive method, often requiring more than 10 times more computational time than the other methods. The "AUC-LPC optim. C" is still slow, because optimization is repeated 25 times in order to find an optimal C. As expected, the "AUC-LPC optim. C $n^- + n^+$" is roughly 25 times slower than the "AUC-LPC subsamp. $n^- + n^+$". Computing the AUC-LPC with the nearest neighbor subsampling is slightly faster than computing the AUC-LPC with the random constraint subsampling, although the

total number of constraints is the same. It is, probably, caused by the fact that the nearest neighbor subsampling procedure generates dependent and redundant constraints. Therefore, the optimization problem is slightly simpler than the problem where the constraints are independent and diverse. For the (very) high dimensional feature sets, such as `colon` and `leuk`, where the computation of the inverse of the covariance matrix is very expensive, training of AUC-LPC classifiers is even faster than training of the LDA and QD.

2.4.3 Lung disease detection

To show the utility of the AUC optimization in a real world example, we discuss the detection of interstitial lung disease in chest radiographs. The task is to classify patients as being healthy or being ill, based on the classification of individual pixels in a radiograph. The experiments are performed on a database obtained at the University of Chicago hospitals [55]. It contains 100 normal chest radiographs and 100 abnormal radiographs with interstitial lung disease. ILD manifests itself on the radiographs with a variety of abnormal texture patterns. The normal cases were selected based on consensus of the panel of experienced radiologists. The abnormal cases were selected based on radiological findings, computer tomography scans, clinical data and/or follow-up radiographs, by consensus of the same radiologists. Any image that contained possibly abnormal or definitely abnormal areas was labeled abnormal. The radiographs were digitized to 2000 by 2000 pixels with 0.175 mm pixel size and 10 bits intensity.

In each of the radiographs lung segmentation is performed using the Active Shape Model algorithm, description of which can be found elsewhere (e.g., in [56]). Then, the mean lung shape is calculated from the available images with the previously segmented lung fields, and every 5th pixel in the X and Y directions is selected within the mean lung fields, in total 7103 pixels. Correspondent pixels are subsequently obtained in every image using a warping function computed between the mean lung shape and a given lung segmentation, as described in [57]. From each radiograph, a number of features are computed for subsampled pixels. First, images are filtered with the Gaussian derivatives, up and including the second order, at five different scales. An original pixel intensity and pixel intensities from all the filtered images are included in the feature set. Furthermore, for each pixel in the original and filtered images, four central moments, the mean, standard deviation, skewness and kurtosis, are computed from a small circular region around the pixel. These features characterize the textural appearance of a pixel neighborhood. Together with the two positional features and a binary feature indicating if a pixel is inside a rib or not (found by applying rib segmentation from [58]), this results in 158 features per pixel.

In order to classify a patient, the pixel classification outputs $h(\mathbf{x})$ for pixels from the radiograph $\mathcal{X}_i$ have to be integrated into a single output $g(\mathcal{X}_i)$. A simple combination rule is used: the outputs $h(\mathbf{x})$ for the pixels of $\mathcal{X}_i$ are sorted, and a certain quantile level output (for instance, the 95% percentile) is extracted. This

Table 2.3: *AUC performances. Results are averaged over five 10-fold cross validation runs, the standard deviation is given in parenthesis. Results for the "AUC-LPC full" on* `abal` *and* `thyr` *cannot be given because the number of constraints was to large.*

	data set				
classifier	`heart`	`biomed`	`auto`	`abal`	`thyr`
LDA	56.1 (0.4)	86.4 (0.6)	58.0 (1.2)	84.2 (0.1)	58.3 (0.2)
QD	57.2 (0.4)	89.9 (0.7)	72.9 (1.1)	84.9 (0.1)	60.5 (0.1)
Parzen Density	64.2 (0.6)	89.9 (0.6)	84.5 (0.5)	85.3 (0.2)	63.3 (0.2)
1-Nearest neighbor	82.4 (0.7)	92.4 (0.5)	**97.7 (0.7)**	85.5 (0.2)	83.0 (0.4)
L_1 SVM	**89.7 (0.5)**	**95.7 (0.5)**	96.2 (0.8)	90.7 (0.1)	50.5 (6.9)
L_2 SVM	**89.8 (0.6)**	**95.7 (0.4)**	93.9 (0.7)	90.7 (0.1)	61.0 (0.7)
Rankboost $B = 25$	87.9 (1.7)	89.5 (1.8)	76.4 (19.7)	87.7 (0.6)	98.9 (0.0)
Rankboost $B = 50$	87.6 (1.5)	92.1 (1.6)	**77.1 (20.2)**	88.5 (0.8)	99.7 (0.0)
Rankboost $B = 100$	87.4 (1.7)	92.4 (1.6)	**77.8 (21.1)**	89.0 (0.9)	**99.8 (0.0)**
AUC-LPC full	**89.6 (0.7)**	**95.3 (0.7)**	95.6 (0.5)	- (-)	- (-)
AUC-LPC obj. $n^- + n^+$	79.9 (3.6)	90.3 (2.4)	84.4 (5.4)	89.4 (0.5)	81.1 (8.2)
AUC-LPC obj. $5(n^- + n^+)$	85.2 (1.8)	94.1 (0.8)	88.6 (2.5)	90.2 (0.3)	91.2 (2.7)
AUC-LPC 1-nn	87.7 (1.1)	94.8 (0.7)	86.5 (2.4)	73.8 (0.7)	86.8 (0.7)
AUC-LPC 5-nn	**89.7 (1.0)**	**95.3 (0.2)**	94.5 (1.1)	89.0 (0.1)	92.7 (0.9)
AUC-LPC subsamp. $n^- + n^+$	**88.0 (1.8)**	**95.0 (0.9)**	93.3 (0.6)	**90.8 (0.1)**	96.7 (0.2)
AUC-LPC subsamp. $5(n^- + n^+)$	**89.3 (0.6)**	**95.3 (0.6)**	95.9 (0.7)	**90.9 (0.1)**	96.3 (0.2)
AUC-LPC optim. C $n^- + n^+$	88.5 (0.8)	95.0 (0.4)	93.3 (0.9)	**90.8 (0.1)**	96.2 (0.4)
AUC-LPC optim. C $5(n^- + n^+)$	**89.3 (1.0)**	**95.2 (0.5)**	95.0 (1.1)	**90.8 (0.1)**	96.3 (0.3)
AUC-LPC optRBF $n^- + n^+$	**89.3 (0.9)**	92.8 (0.9)	93.2 (1.9)	89.6 (0.3)	97.1 (0.3)

	data set				
classifier	glass	diab	ecoli	colon	leuk
LDA	60.6 (0.6)	70.8 (0.3)	93.1 (0.5)	50.0 (0.0)	79.9 (1.5)
QD	64.2 (0.9)	70.8 (0.2)	93.6 (0.4)	50.0 (0.0)	78.3 (0.8)
Parzen Density	68.2 (0.5)	75.3 (0.2)	94.8 (0.2)	50.0 (0.0)	50.0 (0.0)
1-Nearest neighbor	**87.7 (1.4)**	72.2 (0.4)	**95.3 (0.4)**	78.9 (2.3)	97.7 (1.1)
L_1 SVM	59.0 (0.6)	83.0 (0.1)	93.4 (0.5)	**82.9 (3.7)**	**98.8 (1.3)**
L_2 SVM	65.2 (1.7)	**83.0 (0.1)**	93.5 (0.4)	**84.5 (1.1)**	98.6 (0.4)
Rankboost $B = 25$	73.0 (2.4)	81.0 (0.3)	93.0 (0.8)	77.0 (1.8)	**99.2 (1.4)**
Rankboost $B = 50$	74.2 (2.7)	82.3 (0.4)	94.0 (0.6)	80.1 (2.3)	**99.8 (0.6)**
Rankboost $B = 100$	77.3 (2.4)	**83.3 (0.3)**	94.4 (0.5)	**81.3 (3.6)**	**100.0 (0.0)**
AUC-LPC full	72.4 (1.0)	- (-)	93.9 (0.4)	**82.9 (3.7)**	98.3 (1.6)
AUC-LPC obj. $n^- + n^+$	64.6 (4.3)	78.3 (1.8)	91.7 (1.9)	79.4 (5.3)	93.2 (4.6)
AUC-LPC obj. $5(n^- + n^+)$	68.9 (2.6)	80.8 (1.1)	92.4 (1.2)	**82.5 (4.7)**	97.5 (0.6)
AUC-LPC 1-nn	67.8 (1.7)	82.4 (0.4)	91.0 (0.4)	80.1 (4.2)	98.5 (0.5)
AUC-LPC 5-nn	70.3 (0.5)	82.7 (0.2)	93.5 (0.3)	**83.2 (4.6)**	98.3 (1.6)
AUC-LPC subsamp. $n^- + n^+$	67.2 (2.1)	82.7 (0.2)	93.6 (0.4)	**87.9 (5.9)**	98.7 (0.9)
AUC-LPC subsamp. $5(n^- + n^+)$	71.7 (1.2)	**83.1 (0.2)**	93.7 (0.4)	**83.2 (3.0)**	98.4 (1.5)
AUC-LPC optim. C $n^- + n^+$	70.9 (1.3)	**83.1 (0.2)**	93.3 (0.6)	**87.7 (1.7)**	**99.3 (0.9)**
AUC-LPC optim. C $5(n^- + n^+)$	72.2 (0.8)	**83.2 (0.2)**	93.9 (0.4)	**86.5 (2.5)**	98.6 (1.0)
AUC-LPC optRBF $n^- + n^+$	82.1 (1.0)	80.3 (0.7)	**96.1 (0.7)**	**83.8 (5.6)**	95.4 (1.4)

Table 2.4: *AUC performances. The performance of the "AUC-LPC full" on* `diab` *is not given because the number of constraints was too large. Results are averaged over five 10-fold cross validation runs, the standard deviation is given in parenthesis.*

Table 2.5: *Training time (s), averaged over five 10-fold cross validation runs, with the standard deviation in parenthesis. Times for the "AUC-LPC full" on* `abal` *and* `thyr` *are not given because the number of constraints was to large.*

	data set				
classifier	`heart`	`biomed`	`auto`	`abal`	`thyr`
LDA	0.030 (0.000)	0.030 (0.011)	0.028 (0.005)	0.020 (0.001)	0.041 (0.007)
QD	0.030 (0.000)	0.026 (0.001)	0.027 (0.001)	0.020 (0.001)	0.038 (0.001)
Parzen Density	0.263 (0.001)	0.121 (0.011)	0.092 (0.003)	15.767 (0.630)	21.438 (0.716)
1-Nearest neighbor	0.048 (0.000)	0.041 (0.000)	0.042 (0.001)	0.032 (0.002)	0.053 (0.001)
L_1 SVM	0.097 (0.001)	0.057 (0.005)	0.073 (0.004)	5.158 (0.105)	11.935 (0.035)
L_2 SVM	0.060 (0.001)	0.039 (0.001)	0.041 (0.000)	2.034 (0.066)	0.901 (0.005)
Rankboost $B = 25$	0.039 (0.000)	0.027 (0.002)	0.030 (0.001)	0.996 (0.046)	0.419 (0.003)
Rankboost $B = 50$	0.070 (0.000)	0.046 (0.000)	0.052 (0.001)	1.998 (0.077)	0.825 (0.008)
Rankboost $B = 100$	0.163 (0.000)	0.104 (0.000)	0.118 (0.001)	4.984 (0.241)	2.040 (0.014)
AUC-LPC full	363.195 (1.492)	23.958 (0.260)	36.285 (0.436)	- (-)	- (-)
AUC-LPC obj. $n^- + n^+$	0.099 (0.001)	0.052 (0.000)	0.067 (0.001)	4.242 (0.279)	5.251 (0.467)
AUC-LPC obj. $5(n^- + n^+)$	1.105 (0.014)	0.274 (0.007)	0.568 (0.012)	250.269 (5.227)	234.192 (9.445)
AUC-LPC 1-nn	0.081 (0.001)	0.054 (0.001)	0.065 (0.000)	3.652 (0.079)	4.609 (0.030)
AUC-LPC 5-nn	0.710 (0.006)	0.245 (0.001)	0.465 (0.003)	145.575 (3.579)	187.623 (1.449)
AUC-LPC subsamp. $n^- + n^+$	0.090 (0.001)	0.053 (0.000)	0.066 (0.000)	4.688 (0.122)	4.675 (0.144)
AUC-LPC subsamp. $5(n^- + n^+)$	0.986 (0.011)	0.285 (0.002)	0.530 (0.003)	333.453 (10.054)	207.140 (1.986)
AUC-LPC optim. C $k = 1$	1.194 (0.030)	0.425 (0.006)	0.734 (0.009)	115.492 (2.853)	108.132 (4.059)
AUC-LPC optim. C $k = 5$	23.365 (0.256)	6.143 (0.046)	11.563 (0.128)	8206.7 (257.4)	5024.9 (133.8)
AUC-LPC optRBF $n^- + n^+$	1.499 (0.006)	0.335 (0.004)	0.689 (0.004)	58528 (1940.7)	43020 (810.2)

	data set				
classifier	glass	diab	ecoli	colon	leuk
LDA	0.026 (0.001)	0.030 (0.000)	0.024 (0.000)	51.534 (0.040)	427.551 (0.541)
QD	0.026 (0.000)	0.030 (0.000)	0.024 (0.000)	100.793 (0.117)	862.616 (6.556)
Parzen Density	0.107 (0.004)	0.975 (0.009)	0.262 (0.008)	0.124 (0.001)	0.286 (0.004)
1-Nearest neighbor	0.041 (0.000)	0.047 (0.000)	0.039 (0.000)	0.071 (0.001)	0.133 (0.002)
L_1 SVM	0.053 (0.000)	0.192 (0.003)	0.083 (0.000)	1.025 (0.012)	3.145 (0.048)
L_2 SVM	0.051 (0.001)	0.191 (0.003)	0.042 (0.000)	0.101 (0.001)	0.238 (0.004)
Rankboost $B = 25$	0.028 (0.000)	0.082 (0.001)	0.029 (0.000)	0.194 (0.012)	0.407 (0.002)
Rankboost $B = 50$	0.049 (0.000)	0.152 (0.005)	0.051 (0.000)	0.374 (0.029)	0.772 (0.002)
Rankboost $B = 100$	0.112 (0.000)	0.370 (0.012)	0.117 (0.001)	0.911 (0.073)	1.866 (0.004)
AUC-LPC full	26.334 (0.263)	- (-)	80.400 (0.404)	24.235 (0.367)	93.253 (1.166)
AUC-LPC obj. $n^- + n^+$	0.059 (0.001)	0.191 (0.022)	0.073 (0.002)	0.737 (0.008)	1.809 (0.024)
AUC-LPC obj. $5(n^- + n^+)$	0.311 (0.019)	4.340 (0.356)	0.703 (0.019)	5.694 (0.045)	16.329 (0.271)
AUC-LPC 1-nn	0.053 (0.000)	0.208 (0.002)	0.061 (0.000)	0.928 (0.008)	2.661 (0.054)
AUC-LPC 5-nn	0.221 (0.002)	3.146 (0.013)	0.537 (0.002)	5.797 (0.113)	18.102 (0.042)
AUC-LPC subsamp. $n^- + n^+$	0.051 (0.001)	0.192 (0.006)	0.068 (0.001)	0.882 (0.010)	2.489 (0.020)
AUC-LPC subsamp. $5(n^- + n^+)$	0.245 (0.005)	4.350 (0.088)	0.678 (0.007)	6.022 (0.034)	19.167 (0.642)
AUC-LPC optim. C $k = 1$	0.419 (0.006)	3.780 (0.135)	0.862 (0.018)	18.122 (0.235)	53.786 (0.465)
AUC-LPC optim. C $k = 5$	4.802 (0.127)	104.783 (2.312)	16.569 (0.098)	140.479 (1.434)	442.860 (4.086)
AUC-LPC optRBF $n^- + n^+$	0.454 (0.013)	14.725 (0.422)	1.107 (0.032)	1.550 (0.011)	27.627 (0.551)

Table 2.6: *Training time (s), averaged over five 10-fold cross validation runs, with the standard deviation in parenthesis. The training time of the "AUC-LPC full" on* `diab` *is not given because the number of constraints was too large.*

value is used as the final output $g(\mathcal{X}_i)$ for the patient i. This means that when at least 5% of the pixels are classified as being abnormal with high confidence, the output for this patient is labeled "abnormal".

In Table 2.7 the results of the patient classification at three different quantile levels are shown. In this data set both classes are balanced: 90 normal and 90 abnormal radiographs are used for training in each iteration of 10-fold cross validation. For training purposes, all the pixels in the abnormal images are considered abnormal. Consequently, the two classes are highly overlapping in the feature space. The first four classifiers are supervised classifiers, the LDA and QD, the linear L_2 SVM, and the L_2 SVM using the radial basis kernel. The parameters C and σ in the RBF kernel of the SVM are optimized using two-fold cross validation on a rough 25×25 grid of values C, σ.

All supervised classifiers show relatively poor performances in Table 2.7. The results of the AUC-LPC classifiers are much better. We evaluated three different subsampling strategies, the k-means subsampling, the k-nn subsampling with different k, the random constraint subsampling and the constraint subsampling with C optimization. For all the AUC-LPC mappings but the last one, $C = 1$ is chosen, which is close to the value of C chosen by the optimization procedure.

The subsampling strategy can make a difference. The k-means subsampling strategy gives poor performance, because of the relatively small number of clusters that are used. When a larger number of clusters are used, many of these clusters are empty. For this data set, only about 25 clusters can be found reliably. The difference between the k-nn and random subsampling approaches is not very large, but still significant for a higher threshold. The subsampling approach is more stable, the variance of its outcomes is often lower than that of the other classifiers. The optimization of C improves the performance only slightly.

Table 2.7: *AUC results on the interstitial lung disease data with the balanced classes. Results are obtained using 10-fold cross validation. Performances indicated in bold are the best, or not significantly worse than the best for a given quantile level.*

	quantile level output		
classifiers	0.05	0.5	0.95
LDA	56.2 (11.6)	68.6 (17.0)	63.7 (17.2)
QD	53.6 (10.4)	66.2 (16.7)	59.7 (17.7)
linear L_2 SVM	49.1 (17.1)	61.8 (21.3)	63.4 (21.5)
RBF L_2 SVM $\sigma = 500$	71.7 (19.1)	67.9 (17.5)	66.0 (16.9)
AUC-LPC k-means $k = 25$	57.7 (17.3)	64.0 (18.5)	45.5 (15.5)
AUC-LPC knn $k = 1$	**92.6 (7.6)**	89.2 (9.8)	37.0 (18.0)
AUC-LPC knn $k = 5$	**93.1 (7.5)**	**92.3 (7.5)**	37.9 (17.4)
AUC-LPC subs. $M = 2500$	**91.7 (6.8)**	**94.9 (5.9)**	**80.0 (15.2)**
AUC-LPC opt C $M = 2500$	**92.9 (6.7)**	**95.5 (5.7)**	**78.0 (16.3)**

2.5 Conclusions and discussion

Although the AUC optimization is often proposed as a more stable alternative to classification error minimization, in practice it is advantageous only when classes are severely overlapping or imbalanced in size. For these cases, standard classifiers often ignore the few objects from the small class that are located inside the distribution of the larger class. When these objects cannot be considered noise, the optimization of the AUC is preferred over the minimization of the classification error.

Unfortunately, the optimization of the AUC has a complexity quadratic to the training set size. Heuristics have been developed to subsample the training objects, but, in particular for the situation when the classes overlap significantly, suboptimal results have been obtained. In this chapter another approximate AUC optimization procedure was proposed to handle such situations. Instead of subsampling the training set, it randomly subsampled the pairwise constraints. This avoids a biased sampling of the constraints which can happen when objects are subsampled. Furthermore, the constraints that were not used in the optimization could be used to evaluate the classifier. This enables the optimization of meta-parameters that influence the complexity of the classifier. This is particularly useful when the performance is relatively sensitive to the setting of the complexity parameter, and the training set size is small. In this chapter, the proposed subsampling strategy was applied to a (sparse) linear classifier, but it can be applied to any AUC optimizer that explicitly lists the constraints, e.g., a kernelized version.

The experimental results on the standard data sets and the real world medical imaging data demonstrated the benefit of the AUC optimization for highly unbalanced or overlapping classes. The proposed subsampling strategy considerably decreased the training time without compromising the classification performance of the AUC optimizer. In most of the experiments presented in this chapter, the AUC optimizer using this strategy produced significantly better results than the results obtained using the other subsampling heuristics.

Chapter 3

Dissimilarity-based classification in the absence of local ground truth and its application to the diagnostic interpretation of chest radiographs

Y. Arzhaeva, D.M.J. Tax and B. van Ginneken, "Dissimilarity-based classification in the absence of local ground truth: application to the diagnostic interpretation of chest radiographs," *Pattern Recognition*, vol. 42, no. 9, pp. 1768–1776, 2009.

Abstract

In this chapter classification on dissimilarity representations is applied to medical imaging data with the task of discrimination between normal images and images with signs of disease. We show that dissimilarity-based classification is a beneficial approach in dealing with weakly labeled data, i.e. when the location of disease in an image is unknown and therefore local feature-based classifiers cannot be trained. A modification to the standard dissimilarity-based approach is proposed that makes a dissimilarity measure multi-valued, hence, able to retain more information. A multi-valued dissimilarity between an image and a prototype becomes an image representation vector in classification. Several classification outputs with respect to different prototypes are further integrated into a final image decision. Both standard and proposed methods are evaluated on data sets of chest radiographs with textural abnormalities and compared to several feature-based region classification approaches applied to the same data. On a tuberculosis data set the multi-valued dissimilarity-based classification performs as well as the best region classification method applied to the fully labeled data, with an area under the ROC curve (A_z) of 0.82. The standard dissimilarity-based classification yields $A_z = 0.80$. On a data set with interstitial abnormalities both dissimilarity-based approaches achieve $A_z = 0.98$ which is closely behind the best region classification method.

3.1 Introduction

Computer-aided diagnosis (CAD) is an important pattern recognition application. Statistical and structural pattern recognition methods as well as artificial neural networks have been employed in the diagnostic interpretation of medical images of different modalities and organs. The choice of a classification method for a particular application can be influenced by many factors, among them the availability of well-annotated training data. In this chapter we consider weakly labeled medical images with diffuse local textural abnormalities and the task of distinguishing them from images without abnormalities.

In object recognition, weakly labeled data is often defined as images labeled only according to the presence or absence of the objects of interest. For the diagnostic interpretation of medical images, data is weakly labeled in the sense that the absence or presence of disease in an image is known, however, the location of a lesion and its precise delineation are not available. This is, in fact, a common situation in practice because manual annotation of lesions is laborious or even impracticable. Ill-defined diffuse abnormal changes in the local textural appearance of an organ is a clear example in case. Manual segmentation of textural abnormalities is unreliable due to high inter-observer variability. However, texture features extracted from small local patches are potentially very informative in this case. In this work we show that local information alone, without local labels, can give good discriminatory results.

To detect images with abnormalities, we address the absence of local ground truth for training by combining local texture features extracted from a large number of regions of interest (ROIs). In the context of object recognition, a similar classification problem was considered in [59]. The authors focused on combining local information as well and developed generative and discriminative approaches to the task. They showed that the generative model gave a higher classification accuracy but required some fully labeled images for initialization. The discriminative model was considerably less accurate than the generative one. In this work we have chosen a very different approach that does not require any parametric modeling or careful initialization.

We assume that normal images of the same organ bear more similarity to each other with respect to their textural appearance than normal images and images with abnormalities. We also assume that images with the same type of abnormalities are more similar to each other than to normal images. Therefore we propose to reflect the common nature of images belonging to the same class by using dissimilarity representations. In [60] this is defined as the representation of objects by their pairwise comparisons instead of feature vectors. A pairwise comparison is done by computing a measure of dissimilarity, or distance, between two objects. To construct a classifier on dissimilarities one represents each training object as a vector of distances to a set of prototype objects. The standard dissimilarity-based classification is described in [61] and [60]. In this chapter we propose a modification to this strategy.

In the standard approach, a single dissimilarity measure is computed between two images, a test image and a prototype, thus reducing the abundance of local textural information to one quantity. Since an image in our approach is represented by a set of texture features each of which is computed at multiple locations, we propose to retain more information by computing dissimilarities for each feature separately. Then, instead of combining feature-based comparisons into one value as in [62] and [63], we use them as a vector to train a classifier. That allows us to build as many classifiers as there are prototypes, and to classify each test image several times. Subsequently, we combine the outputs of all classifiers into one posterior probability value.

Experimentally, we will focus on two specific data sets. These data sets have already been used in other research papers which enables us to compare our results with previously reported ones. Both data sets present challenging ill-defined textural abnormalities in chest radiographs. The first one is a database from a tuberculosis (TB) mass screening program, the second database contains images with interstitial lung disease (ILD). In [55, 64, 65, 66] one or both of them were used as test data for algorithms to distinguish between normal and abnormal images. In the two best performing classification schemes applied in [65] and [66] local labeling was used for training. That allowed the supervised classification of ROIs as normal or abnormal, and the subsequent integration of local decisions into a decision about the whole radiograph. Local labeling, provided it is correct, provides more information for training the system, and hence, such systems are potentially more powerful than ones where local labeling is unavailable. Although the labeling implemented in [66] or [65] might not have been a perfect ground truth, we will use their results as benchmarks for our study.

This chapter is organized as follows: Section 3.2 introduces the dissimilarity representation and classification in dissimilarity space, as well as dissimilarity measures we intend to use. The proposed classification approach is also explained in this section. In Section 3.3 we compare the results of different classification strategies applied to medical images. We discuss the results and methods in Section 3.4. Section 3.5 draws conclusions.

3.2 Dissimilarity representations

In statistical pattern recognition objects are usually described by feature vectors. When we consider images described by sets of texture features extracted from a large number of patches, the description of the whole image becomes extremely high-dimensional and therefore inefficient for learning. Dissimilarities provide a convenient alternative for an image representation. Moreover, the proximity-based representation is a natural way of describing the class of similar objects. A profound discussion on this subject can be found in [60, 61], and we borrow notation from that work.

If T is a training set of size n, and R is a set of prototype objects of size

$r, R = \{p_1, \ldots, p_r\}$, then any $x, x \in T$, is represented by a vector of dissimilarities $D(x, R) = \{d(x, p_1), \ldots, d(x, p_r)\}$, where d is a dissimilarity measure. Thereby any traditional classifier operating on a feature space can be built on the $n \times r$ dissimilarity matrix $D(T, R)$. Usually, R is a subset of T, or the same set as T. Objects in R can be randomly selected from T, or selected using a systematic approach (see [67] for a discussion on possible approaches). A test set S of s objects is also described in terms of their distances to R, i.e. by $s \times r$ dissimilarity matrix $D(S, R)$.

Defining a discriminative dissimilarity measure is as difficult as defining good features in traditional feature-based classification. It is logical to demand that the measure is non-negative. Another natural requirement for dissimilarities is to be relatively small for similar objects. To ensure that, is desirable for the measure to satisfy the triangle inequality condition, $d(x, y) \leq d(x, z) + d(z, y)$, for all x, y, z, or else the compactness of dissimilarity representations might be violated [60]. If the measure is also symmetric and definite, it becomes a metric. Metrics are preferred as measures of dissimilarity because many classification methods work in metric spaces. In this chapter we consider only measures that are metrics.

For clarity, we use the terms "feature" and "feature vector" only for original measurements extracted from an object. In our case, these are texture measurements computed from a large number of image ROIs. In the dissimilarity-based classification framework, a vector of dissimilarities constitutes an image representation and is passed to a classifier.

3.2.1 Dissimilarity measures

Let us introduce several common measures suitable for the task of image classification. In the context of image retrieval images are often characterized by multi-dimensional histograms of their features. An example of such features is the distribution of pixel intensities in an image and in filtered versions thereof. In this study, dedicated texture features are extracted from numerous and uniformly placed ROIs in images. Similarly to pixel intensity histograms we can build one- or multi-dimensional histograms to estimate the probability density of these features or the density of their joint distribution. Several non-parametric measures of dissimilarities between two histograms $h = \{h(i)\}$ and $k = \{k(i)\}$, i being a bin index, will be experimentally investigated in this chapter.

Minkowski, or l_p, distance:

$$d_p(h, k) = \Big(\sum_i \big| h(i) - k(i) \big|^p \Big)^{1/p}. \tag{3.1}$$

For $p = 1$, this becomes the city block distance, and for $p = 2$, the Euclidean distance.

χ^2 statistics:

$$d_{\chi^2}(h,k) = \sum_i \frac{\Big(h(i) - m(i)\Big)^2}{m(i)}\,, \tag{3.2}$$

where $m(i) = \frac{h(i)+k(i)}{2}$. This measure calculates how unlikely it is that both histograms represent the same distribution.

Jeffrey divergence:

$$d_J(h,k) = \sum_i \left(h(i)\log\frac{h(i)}{m(i)} + k(i)\log\frac{k(i)}{m(i)}\right), \tag{3.3}$$

where again $m(i) = \frac{h(i)+k(i)}{2}$. The Jeffrey divergence is a modification of the Kullback-Leibler divergence [68] and is numerically stable, symmetric and robust with respect to noise and the size of histogram bins [62].

Match distance:

$$d_M(h,k) = \sum_i |\hat{h}(i) - \hat{k}(i)|\,, \tag{3.4}$$

where $\hat{h}(i) = \sum_{j\le i} h(j)$ and $\hat{k}(i) = \sum_{j\le i} k(j)$ are cumulative histograms of h and k respectively.

Kolmogorov-Smirnov distance:

$$d_{KS}(h,k) = \max_i \left(|\hat{h}(i) - \hat{k}(i)|\right), \tag{3.5}$$

where again $\hat{h}(i)$ and $\hat{k}(i)$ are cumulative histograms. The match and Kolmogorov-Smirnov distances are only defined for one-dimensional histograms, because the ordering relation $j \le i$ is arbitrary in more than one dimension [69].

A one-dimensional histogram is obtained by a suitable binning of the range of feature values. However, we do not apply binning for multidimensional joint feature distributions in order to avoid sparse and unstable histograms. Instead, we use a clustering algorithm such as k-means to partition the feature space into a fixed number of bins [69].

A dissimilarity between two images x and y can be expressed as the dissimilarity of their joint feature distributions, $d(x,y) = d(h,k)$, where h and k are the multi-dimensional histograms of the features of x and y respectively. Dissimilarity measures from Eq. 3.1–3.3 are used for this purpose in the experimental part of this study. This is by no means an exhaustive list of suitable dissimilarity measures (see [60] and [69] for more). Besides comparing their joint feature distributions, a dissimilarity between two images can be computed by combining independently evaluated comparisons of individual feature distributions. Just as in [62], we use

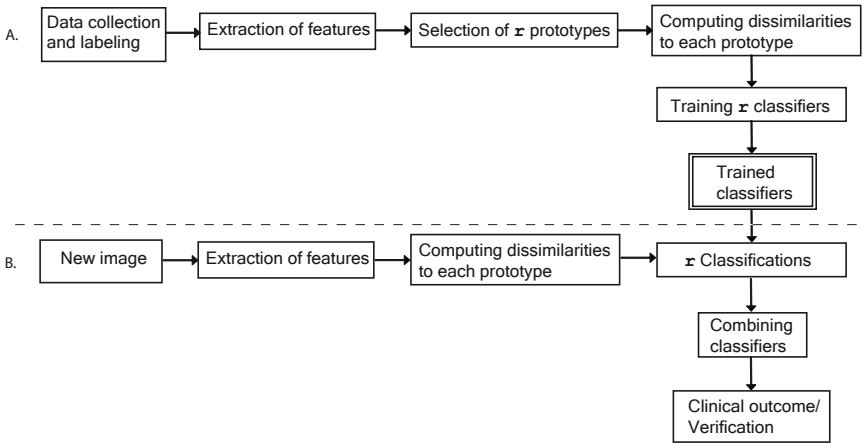
the Minkowski norm of order 1 to combine them, $d(x,y) = \sum_f d(h_f, k_f)$, where $d(h_f, k_f)$ is any of the measures from Eq. 3.1–3.5 computed for the histograms h_f and k_f of the feature f of the images x and y respectively.

3.2.2 Proposed approach

We propose, however, not to combine the individual comparisons $d(h_f, k_f)$ into one value but to construct a vector $D(x,y) = \{d(h_f, k_f)\}$ and use it as a new image representation. The main difference between this approach and the standard dissimilarity-based classification lies in how an image is represented through its comparisons with the prototypes. In the standard approach, each element of the image representation vector $D(x,R)$ expresses the dissimilarity of the image x to a different prototype from the set R. In our representation, each element of the vector $D(x,p_k)$ is a dissimilarity between the image x and the same prototype p_k, $p_k \in R$, computed with respect to a different image characteristic. Here those characteristics are the distributions of various features extracted from both images.

Thus, we can obtain as many image representations as we have prototypes, each representation vector having the same dimensionality as the set of original features. With r prototype images, r representations are obtained for each training image, and consequently r classifiers can be trained. A test image, subsequently, can be classified r times using its prototype-bound representations. We suggest combining the outputs of all classifiers to obtain a final solution. In Figure 3.1 the training and testing phases of the proposed approach are schematically depicted.

Figure 3.1: *Flow chart of the proposed approach. (A) Training phase. (B) Testing phase.*



The combination of classifiers benefits from complementary information provided by different image representations. Various fixed, trainable and adaptable combiners have been described in the literature (see [24] for references). In the absence of a large pool of training data we opt for a fixed combination rule and leave the exploration of trainable schemes for further research. It is shown in [70] that the sum (or average) rule outperforms other fixed rules (such as the voting and product rules) for combining classifiers that use different representations of the patterns to be classified. The sum rule proved to be less sensitive to the errors of individual classifiers. In this chapter we combine the image posterior probabilities resulting from different classifiers with the sum rule:

$$P(c|x) = \frac{1}{r}\sum_{k=0}^{r-1} P_k(c|x), \tag{3.6}$$

where $P(c|x)$ is a posterior probability that the image x belongs to a class $c, c = \{0, 1\}$, and $P_k(c|x)$ is a posterior probability yielded by the classifier k.

3.3 Comparative experiments

In this section we apply both standard and proposed approaches of Section 3.2 to the classification of two sets of medical images exhibiting textural abnormalities. Additionally, we compare the dissimilarity-based methods to a region classification strategy adapted to weakly labeled data. The classification task is discrimination between normal images (of healthy subjects) and images containing disease (we refer to such images as abnormal).

3.3.1 Data

The TB database was collected from a tuberculosis screening program in the Netherlands. Posterior-anterior (PA) chest radiographs were digitized to 932 by 932 pixels and 12-bit intensity. More technical details can be found in [66]. The data set used in our experiments contains 241 normal cases and 147 abnormal cases with textural abnormalities. These cases were selected from a larger database by exclusion of images with non-textural abnormalities as well as images with artefacts (e.g. clothing artefacts). The same subset was used in [66]. The ground truth for the images was set by two radiologists. The image was considered abnormal if one of them judged the image to be abnormal.

The ILD database consists of 100 normal and 100 abnormal PA chest radiographs obtained from the daily clinical practice of the University of Chicago hospitals [71]. The abnormal radiographs exhibited various interstitial lung diseases and were selected on the basis of radiological findings, clinical and computed tomography data, biopsy and the consensus of the panel of experienced radiologists. Each normal case was chosen based on the consensus of the same panel. The radiographs were digitized to 2000 by 2000 pixels.

In both data sets the lung fields were segmented from the rest of the image using the Active Shape Model algorithm, description of which can be found elsewhere (e.g. in [56]). The segmentation was performed with the same settings of parameters as in [25]. Prior to feature extraction the resolution of images in both data sets was sub-sampled to 700 by 700. In Figures 3.2(a) and 3.2(b) an example radiograph from the ILD database is shown, together with its lung mask obtained from the lung segmentation.

3.3.2 Texture features

In order to extract discriminative texture features the images are filtered with a multiscale filter bank of Gaussian derivatives, and the moments of histograms are calculated from regions in the derived images. Using multiple scales enables the characterization of texture elements of different sizes, and the analysis of local histograms considers the texture primitives regardless of their spacial distribution. This is a general approach to texture characterization [72, 73]. The histogram moments were successfully used for automatic detection of textural abnormalities in chest radiographs [66, 74] and for texture analysis in thoracic computed tomography scans [75].

Prior to filtering the image, pixel values in the lung fields are mirrored outside the lungs symmetrically with respect to the lung borders. Namely, for each pixel outside the lungs, the pixel value is substituted by its counterpart inside the lungs with the nearest pixel on the lung border as a center of symmetry. This prevents a major distortion in the filter output near the lung borders which is normally caused by a large difference in appearance inside and outside the lungs. Additionally, the left lung is flipped to resemble the right lung in orientation of various anatomical and texture elements. Figures 3.2(c) and 3.2(d) illustrate the mirroring and flipping preprocessing steps.

The lung fields are subdivided into overlapping regions of interest (ROIs). We use an 8 by 8 pixel spacing to define the centers of circular ROIs, each of which have a radius of 32 pixels. The number of ROIs per radiograph ranges from 1400 to 4100 approximately depending on the size of individual lungs. Radiographs are filtered with Gaussian derivatives of orders $0, 1$ and 2 at five scales, $\sigma = 1, 2, 4, 8, 16$ (illustrated in Figure 3.3). Then four central moments of the pixel intensity distribution, namely, the mean, standard deviation, skewness and kurtosis are calculated from each ROI in the original and filtered images, amounting to 124 features in total. These are the same features that were successfully used in the classification of small regions in [76] and in [74] for localization of interstitial abnormalities.

In those works local texture features were complemented by two position features, namely x and y coordinates of the ROI centers relative to the center of the mass of the lung field. For this study we have assumed that the histograms of ROI locations covering lung fields uniformly would not be informative attributes in distinguishing between normal and abnormal lungs. For the multi-dimensional

Figure 3.2: *Data preprocessing steps are shown on an example chest radiograph: (a) the original radiograph; (b) the lung mask obtained from the lung segmentation, with distinct mask values for the right and left lung fields; (c) the right lung, delineated, with its exterior substituted by corresponding pixel values from the lung inside; (d) the left lung, delineated and flipped, with its exterior substituted by corresponding pixel values from the lung inside.*

histograms of joint feature distributions adding position features could have resulted in the mistaken estimation of two abnormal images as dissimilar when their abnormalities were located in different lung regions. However, in practice, including position features brings minor improvements to dissimilarity classification results for both approaches. This could possibly be explained by an observation made in [66] that the spacial distribution of abnormal areas in the TB and ILD databases is not uniform. Tuberculosis is known to often affect the upper lung fields, while interstitial abnormalities are more likely to occur in the lower lung fields.

The original features are normalized. In the set of prototypes, each feature is translated and scaled to have zero mean and unit standard deviation. Then the same normalization parameters are applied to feature vectors in the rest of the images.

3.3.3 Histograms

Before applying dissimilarity classification methods, the probability densities of local texture features have to be estimated. We represent the probability density of individual features by a histogram with 128 bins. The bin partitioning is fixed on a set of prototype images. Namely, the range of possible values of each feature is estimated and split into 128 equal intervals.

To construct the multi-dimensional histogram of the joint distribution of features we first run a k-means algorithm with 128 clusters on the combined distribution of ROIs from all the prototypes. Then, for any image in a database, each feature vector is assigned to the closest cluster in the partitioned feature space.

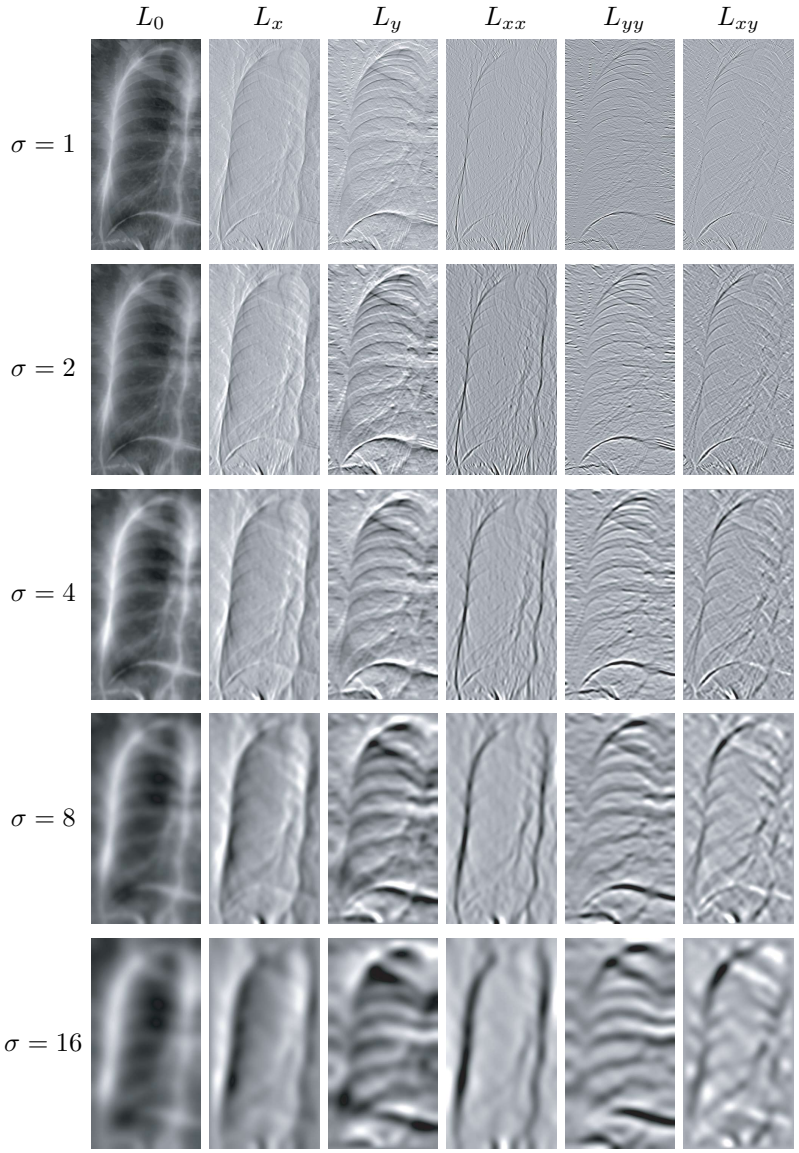


Figure 3.3: *Illustration of the 30 filtered images for the input image of a right lung from Figure 3.2(c). The input image is convolved with Gaussian derivatives of orders* $0, 1$ *and* 2 *at five scales. Each row shows the resulting images,* L_0*,* L_x*,* L_y*,* L_{xx}*,* L_{yy} *and* L_{xy}*, at one scale.*

3.3.4 Comparison with standard dissimilarity classification

In both databases we randomly selected 10 normal and 10 abnormal radiographs to serve as prototype images, and computed dissimilarities between each prototype and all the images in a database (including the prototypes themselves). For use with the standard dissimilarity-based classification approach each image was represented by a 20-dimensional vector computed with every appropriate dissimilarity measure described in Section 3.2.1. For the proposed multi-valued dissimilarity-based method each image was described by twenty 126-dimensional image representation vectors computed using every dissimilarity measure for one-dimensional histograms from Eq. 3.1– 3.4.

The classification experiments were conducted by means of cross-validation. Each database, with exclusion of the prototypes, was divided into 4 folds, with equal amounts of normal and abnormal images in each fold. Classification was performed 4 times, each time with a different fold as a test set and the other three folds together as a training set. The prototype images were always appended to the training set. We estimated the classification performances of both methods by means of receiver operating characteristic (ROC) analysis [29]. The ROC curve plots the sensitivity of a classifier against its 1-specificity at varying confidence thresholds. A_z, the area under the ROC curve, was used as a classification performance measure.

For both dissimilarity-based approaches we compared the linear discriminant analysis (LDA), quadratic discriminant analysis , k-nearest neighbor classifier (k = 15), and support vector machine (radial basis function kernel, the kernel parameter g =1.0 and penalty parameter C =1.0). Details of these classifiers can be found elsewhere, i.e. in [23]. We found that the LDA performed considerably better than the other classifiers did with the same fixed test, training and prototype sets. We think that one of possible causes for this is the simplicity of the LDA classifier. Another likely explanation could be that the dissimilarity measures are based on the summation over many components, and therefore tend to be normally distributed. Hence, normal density-based classifiers, such as the linear and quadratic discriminants, should perform well in dissimilarity spaces, as was already observed in [61]. Moreover, since the LDA is a linearly weighted combination of dissimilarities, it is less sensitive to errors caused by some individual dissimilarities. The quadratic discriminant analysis might have performed well in our experiments had we had more training samples to accurately estimate the class covariance matrices.

For the LDA, the number of features might still be too large relatively to the number of training samples in the multi-valued dissimilarity-based experiment, especially in the ILD database. In an attempt to resolve this we applied principal component analysis (PCA) retaining 99% of variance to the image representation vectors. This improved the multi-valued dissimilarity classification performance on the ILD data.

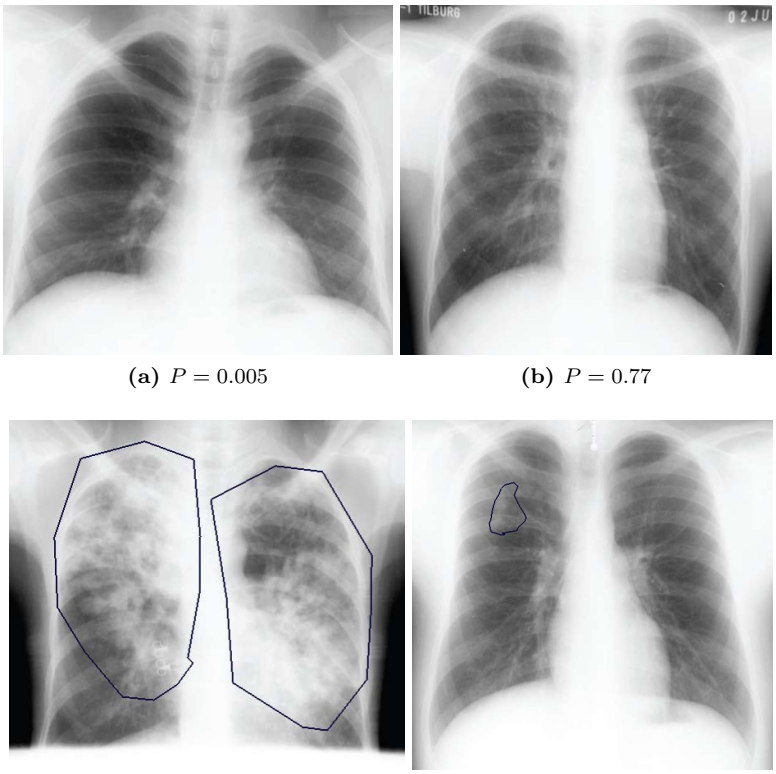
Table 3.1 displays the results of the standard dissimilarity-based classification

for all the dissimilarity measures under consideration. The results of the proposed method for different dissimilarity measures are presented in Table 3.2. Note, that the results in Table 3.2 on the ILD database were obtained by application of PCA while no PCA was applied to the TB data. In both tables A_z values are averaged over the folds and accompanied by the standard deviation. Examples of radiographs, correctly classified or misclassified by the proposed method, are given in Figure 3.4.

The image in Figure 3.4(d) is the instructive example of a situation where the dissimilarity-based methods can fail. Subtle abnormalities with a small size relative to the whole area of interest are unlikely to be discernibly reflected in global measures, such as histograms over the whole lung fields. On the other hand, the normal image shown in Figure 3.4(b) exhibits an enlarged perhilar region in the upper and middle right lung with a bright and pronounced texture pattern. We assume that the contribution of that pattern into global histograms caused the misclassification of this image as abnormal by our system. Correct classification of the perihilar region is also difficult for a region-based classification because of its bright and pattern-rich manifestation and large variability [74].

Overall, the standard dissimilarity-based classification performed as well as the multi-valued approach on the ILD data. On the TB data the standard dissimilarity-based classification was less accurate than our modification. From Table 3.1 we can also conclude that the performance of the standard dissimilarity-based method largely depended on what type of dissimilarity measures were employed. All combined measures were superior to the measures computed between multi-dimensional histograms. It is likely that 128 clusters in the 126-dimensional feature space was a rather coarse histogram partitioning which made histograms of different classes less distinctive; with 128 clusters we got from 11 to 32 entries per histogram bin on average. Given a fixed number of samples, a considerable increase in the number of bins could lead to a histogram sparseness which, in turn, could make a histogram less discriminative as well. The one-dimensional histogram binning did not have that problem and produced, possibly, more discriminatory histograms.

Concerning the utilized dissimilarity measures, the χ^2 statistics and Jeffrey divergence were consistently successful in both classification approaches. The match and Kolmogorov-Smirnov distances were the least stable of all the measures. All the measures performed better in the multi-valued approach than in the standard approach when applied to the TB data. In the application to the ILD data, the Euclidean and match distances showed slightly better performances in the standard settings. The differences in performance between the measures were rather small in both approaches applied to the ILD data. On the contrary, when applied to the TB data, the χ^2 statistics and Jeffrey divergence performed considerably better than the other combined measures in the standard approach. The performance differences between measures are less striking in the multi-valued approach, with four measures showing nearly identical results.



(a) $P = 0.005$ **(b)** $P = 0.77$

(c) $P = 0.996$ **(d)** $P = 0.07$

Figure 3.4: *Two normal, (a) and (b), and two abnormal, (c) and (d), cases from the TB database. Abnormalities are roughly outlined on images (c) and (d). The proposed dissimilarity-based method with the city block dissimilarity measure found image (a) most normal and image (b) most abnormal of all normal images. Image (c) was found most abnormal and image (d) most normal of all abnormal images. Estimated probabilities of being abnormal are indicated below each image.*

Table 3.1: *The performances of the standard dissimilarity-based classification applied to the TB and ILD data sets for the dissimilarity measures described in Section 3.2.1. The table presents the average and standard deviation of the areas under the ROC curve.*

Dissimilarity measure	TB data set	ILD data set
City block	0.715 (0.052)	0.936 (0.05)
Euclidean	0.693 (0.069)	0.924 (0.037)
χ^2 statistics	0.719 (0.034)	0.929 (0.047)
Jeffrey divergence	0.724 (0.033)	0.931 (0.04)
Combined city block	0.771 (0.04)	0.971 (0.014)
Combined Euclidean	0.767 (0.042)	0.976 (0.016)
Combined χ^2	0.797 (0.044)	0.974 (0.018)
Combined Jeffrey	0.798 (0.041)	0.974 (0.019)
Combined match distance	0.747 (0.08)	0.964 (0.014)
Combined Kolomogorov-Sm.	0.748 (0.067)	0.966 (0.015)

Table 3.2: *The performances of the proposed multi-valued dissimilarity-based classification applied to the TB and ILD data sets for different dissimilarity measures. The table presents the average and standard deviation of the areas under the ROC curve.*

Dissimilarity measure	TB data set	ILD data set
City block	0.820 (0.026)	0.974 (0.011)
Euclidean	0.817 (0.012)	0.970 (0.008)
χ^2 statistics	0.812 (0.020)	0.975 (0.010)
Jeffrey divergence	0.817 (0.018)	0.974 (0.014)
Match distance	0.793 (0.038)	0.961 (0.033)
Kolomogorov-Sm. distance	0.775 (0.025)	0.978 (0.011)

Table 3.3: *Comparison of different classification strategies in terms of A_z. The best results of the standard and multi-valued dissimilarity-based approaches are taken from Tables 3.1 and 3.2, respectively.*

Study or method	TB data set	ILD data set
van Ginneken [66]	0.820 (0.022)	0.986 (0.006)
Ishida [65]	n.a.	0.976 (0.012)
Naive region classification	0.786 (0.035)	0.962 (0.031)
Standard dissimilarities (best result)	0.798 (0.041)	0.976 (0.016)
Multi-valued dissimilarities (best result)	0.820 (0.026)	0.978 (0.011)

3.3.5 Comparison with region classification

To put the dissimilarity-based approaches in perspective we compare them with another strategy to deal with weakly labeled data. It is based on a naive assumption that every pixel in an abnormal image is abnormal. Hence, every region extracted from the lung fields as described in Section 3.3.2 is labeled according to the radiograph it belongs to. In this way the absence of local labels is circumvented. The image classification task then can be considered as a region classification and subsequent fusion of regional posterior probabilities. Such an approach was first proposed in [77].

In the region classification experiment the same division of the data into folds was applied. The images used as prototypes in the dissimilarity-bases experiments were added to the training sets of each fold. Each ROI was described by a 126-dimensional normalized feature vector consisting of 124 texture features and 2 position features (see Sections 3.3.2 and 3.3.3). For practical reasons we randomly selected 20% of the ROIs from each training image to train a classifier. As we already saw in [74], the LDA was a good choice of a classifier to discern between normal and abnormal ROIs. The classification yielded the posterior probability of being abnormal for each ROI in the image. The overall image decision was obtained by integrating the regional posterior probabilities using the 90% percentile rule.

The classification performance was evaluated in terms of A_z and averaged over the folds. We achieved $A_z = 0.786$ with a standard deviation of 0.035 on the TB database, and $A_z = 0.962$ with a standard deviation of 0.031 on the ILD database.

3.4 Discussion

The merits of the two dissimilarity-based methods in this particular application can be evaluated in the light of previous research. In Table 3.3 the results of different classification strategies applied to the TB and ILD databases are given. The first two rows present the results from the studies where, as mentioned in Section 3.1, ROIs were labeled and classified, and then the results of region classification were integrated into an overall image decision. The third row holds the results of region classification cf. Section 3.3.5. In the last two rows the results of the two dissimilarity-based approaches are presented, taking the best results from Tables 3.1 and 3.2.

It is interesting to note that the best and worst approaches for both data sets are region classification techniques. The best approach described in [66] used the manual annotation of lesions in the images. Although it was a rough annotation, it was obviously a better ground truth than the naive region abnormality assumption employed as a labeling strategy in Section 3.3.5. It should be noticed, however, that the gap between the best and worst classification results is relatively small. The multi-valued dissimilarity-based approach shows the same result as the best performing region classification on the TB data. On the ILD

data both dissimilarity-based methods perform similarly to the region-based classification described in [65]. This comparison gives reason to hypothesize that the dissimilarity-based approaches are advantageous in dealing with weakly labeled textural data because they are capable of achieving results close to or even equal to those obtained with data labeled as fully as possible.

We suppose that the dissimilarity-based methods might not be equally useful in dealing with weakly labeled images with other types of abnormalities, e.g. in detecting chest radiographs with lung nodules. Similar to the abnormality in Figure 3.4(d), many lung nodules are too small to possibly make a difference in a global measure, unless such a measure is a dedicated lung nodule filter of some sort. A more important consideration, however, is how well- or ill-defined a type of abnormality is. From the beginning, the application of the dissimilarity-based methods was motivated by the difficulty or impracticability of obtaining the local ground truth. When a reliable local ground truth is available to train a CAD system, we would suggest to use detection algorithms that can directly employ local information.

Next we discuss how the results of the standard dissimilarity-based approach compared with its multi-valued modification. The former used one-dimensional histograms of individual feature distributions in order to produce combined dissimilarity measures, while the latter used comparisons between feature histograms directly in the image representation vectors. Both strategies yielded comparable classification performances, which is not unexpected since image posterior probabilities are conveyed through a weighted sum of feature dissimilarities by both approaches.

The LDA, applied in the standard approach, results in the linear combination of dissimilarities to all the prototypes, where a dissimilarity to each prototype is, in turn, a linear combination of individual feature dissimilarities. In the multi-valued approach, the LDA first yields the linear combination of individual feature dissimilarities to one prototype. Then, by averaging the LDA results over all the prototypes in the classifier combination phase, we obtain the same type of additive solution as in the standard approach. It seems that the order of application of the LDA is what makes the proposed multi-valued approach slightly more accurate than the standard classification on dissimilarities. This assumption conforms with our initial idea that a classifier applied to feature-based dissimilarities rather than to image-based dissimilarities should benefit from more information.

Regarding the dissimilarity measures, it was noted in Section 3.3.4 that the χ^2 statistics and Jeffrey divergence measures showed comparable classification results in both approaches on both the TB and ILD data sets. While the city block and Euclidean distances could be considered as general-purpose measures in the Euclidean space, the χ^2 statistics and Jeffrey divergence are dedicated measures for probability distributions originating from statistical and information theory, respectively. That might explain their reliable performance in classifying features based on comparisons between distributions. The match and Kolmogorov-Smirnov distances are special distance measures for cumulative histograms and

are known to produce better results with finer binning [69]. In our experiments they performed less satisfactorily than the other measures, with the exception of the Kolmogorov-Smirnov distance yielding the best result on the ILD data in the multi-valued approach. They were also less stable, generally, exhibiting some of the largest standard deviations. We may only hypothesize that with finer histogram binning and more samples, the results of these two measures could improve.

It was beyond the scope of this study to investigate the use of prototype selection methods. In [67] it is argued that the systematic selection of prototypes in general does better than the random selection. With the random selection of prototypes we have already achieved results that are closely comparable with those obtained in the presence of local ground truth. It could be an intriguing future study to investigate whether adding prototype selection notably improves the classification performance. Achieving better results on a weakly labeled data set than on fully labeled data would indicate either unsatisfactory local ground truth used by the feature-based approaches in classification of regions, or some room for improvement in the region classification method itself.

3.5 Conclusion

In conclusion, we successfully applied a dissimilarity-based classification approach to weakly labeled chest radiographs with textural abnormalities. The obtained results were similar to those obtained by feature-based methods on fully labeled data. Our proposed modification to the standard dissimilarity-based approach was preferable in the classification of tuberculosis, while both dissimilarity-based methods performed equally well in the classification of interstitial abnormalities. The application of these techniques to other weakly labeled image data is of interest for future research.

Chapter 4

Detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography

Y. Arzhaeva, M. Prokop, D.M.J. Tax, P.A. de Jong, C.M. Schaefer-Prokop and B. van Ginneken, "Computer-aided detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography," *Medical Physics*, vol. 34, no. 12, pp. 4798–4809, 2007.

Abstract

A computer-aided detection (CAD) system is presented for the localization of interstitial lesions in chest radiographs. The system analyzes the complete lung fields using a two-class supervised pattern classification approach to distinguish between normal texture and texture affected by interstitial lung disease. Analysis is done pixel-wise and produces a probability map for an image where each pixel in the lung fields is assigned a probability of being abnormal. Interstitial lesions are often subtle and ill-defined on x-rays and hence difficult to detect, even for expert radiologists. Therefore a new, semi-automatic method is proposed for setting a reference standard for training and evaluating the CAD system. The proposed method employs the fact that interstitial lesions are more distinct on a computed tomography (CT) scan than on a radiograph. Lesion outlines, manually drawn on coronal slices of a CT scan of the same patient, are automatically transformed to corresponding outlines on the chest x-ray, using manually indicated correspondences for a small set of anatomical landmarks. For the texture analysis, local structures are described by means of the multi-scale Gaussian filter bank. The system performance is evaluated with ROC analysis on a database of digital chest radiographs containing 44 abnormal and 8 normal cases. The best performance is achieved for the linear discriminant and support vector machine classifiers, with an area under the ROC curve (A_z) of 0.78. Separate ROC curves are built for classification of abnormalities of different degrees of subtlety versus normal class. Here the best performance in terms of A_z is 0.90 for differentiation between obviously abnormal and normal pixels. The system is compared with two

human observers, an expert chest radiologist and a chest radiologist in training, on evaluation of regions. Each lung field is divided in four regions, and the reference standard and the probability maps are converted into region scores. The system performance does not significantly differ from that of the observers, when the perihilar regions are excluded from evaluation, and reaches $A_z = 0.85$ for the system, with $A_z = 0.88$ for both observers.

4.1 Introduction

Conventional chest radiography is an important diagnostic examination for a variety of lung disorders including interstitial lung disease (ILD). In recent years computer tomography (CT) has become the modality of choice for the diagnostics of ILD [78]. However, chest radiography remains the first and most common examination in clinical practice. In comparison to CT, it is simple to perform and inexpensive. Therefore the role of chest radiography is to provide an initial detection of abnormalities and a preliminary diagnosis, and to give a recommendation for a subsequent CT examination [1]. Techniques for automated detection and characterization of abnormalities in chest radiography have been developed for about two decades [9, 16]. In recent studies [19, 79], computer-aided detection (CAD) systems for chest radiographs have been demonstrated to be potentially useful tools leading to more accurate diagnoses for various lung diseases including detection of ILD. Currently the results of computer analysis are considered to play a complementary role in clinical practice as a second opinion.

ILD, also known as diffuse parenchymal lung disease, is the common term for more than 150 types of disorders, which may cause considerable morbidity and mortality [1, 80]. The interstitium of the lung is the tissue between the air sacs, and when the interstitium is damaged the 'textural appearance' of the lung is changed in radiological images. Detection and differentiation of ILD is an exceptionally difficult task, even for an experienced chest radiologist. The key radiological finding is widespread or focal shadowing with specific underlying patterns. The majority of interstitial diseases exhibit a reticular, nodular or ground-glass pattern [10, 81], or a combination of these. Whereas a large variation of abnormal patterns can represent one type of ILD, radiographs of patients with different types of ILD may look alike. Moreover, the difference between normal and abnormal texture patterns is ambiguous even for human experts, which is revealed by high inter-observer variability [1, 82]. As a result, development of a CAD system for the detection of ILD in chest radiographs is an extremely challenging task.

The majority of works in this field [55, 83, 84, 65, 85, 66, 77] used an approach that could be roughly divided into three steps. First, regions of interest (ROIs) were manually or automatically selected within the lung fields. From each ROI a set of texture features was computed. Then classification was performed using rule-based or pattern recognition methods, and as the result of classification an "opinion" (a class label or probability of being normal/abnormal) about each ROI was obtained. Finally, probabilities over regions were fused to yield a conclusion for the whole image, determining whether it contained any interstitial abnormalities or not.

The CAD systems exploiting this method were evaluated using receiver operating characteristic (ROC) analysis, and showed high performances when evaluated either as a stand-alone system or as an assistant to radiologists in the task of discrimination between normal chest radiographs and radiographs that contained

signs of interstitial abnormalities [86, 66, 19]. Although the classification of images was based on classification of regions, only in van Ginneken *et al.* [66] was the CAD performance on regions itself evaluated. It appeared there that the classification performance was poorer at region level than at image level (the area under ROC curve values ranged from 0.67 to 0.93 for different regions, and reached 0.99 for images), in other words, the CAD system could not always distinguish between samples of healthy lung texture and regions affected by ILD. In [66], the reference standard for a region was set visually by one radiologist and may therefore not be considered highly reliable. In other previous studies such an evaluation was not carried out at all, possibly because of the absence of a region-level reference standard to compare with.

The accurate delineation of abnormalities is an anticipated ability of such a CAD system. The distribution of ILD throughout the lung is often related to a type of ILD (e.g. extrinsic allergic alveolitis is commonly found in the upper lobes whereas usual interstitial pneumonia is seen mainly in the lower lobes and the lung periphery), and this has important differential diagnostic implications [87]. The main obstacle, in our opinion, that prevents the construction of a system that localizes ILD abnormalities in chest radiographs is the difficulty of obtaining a reliable local reference standard. Without this a CAD system cannot be verified and in many cases cannot be properly trained. Well-defined lesions, e.g. tumors, nodules, calcifications, can be manually segmented or pinpointed and often histologically proven to be a lesion. However, manual segmentation is less suited to the delineation of interstitial lesions due to their diffuse and ambiguous appearance. In this paper we use an alternative way to establish a more reliable reference standard for chest radiographs.

In our previous work [76] we demonstrated with small-size peripheral regions from digitized chest radiographs that local analysis could yield a high classification performance. In this paper we present a CAD system for detection of ILD lesions in complete posterior-anterior chest radiographs. The system is trained and tested on a relatively small set of digital chest radiographs to show its ability to locate interstitial abnormalities. An innovative method for obtaining the local reference standard is presented. The reference standard is established by using a CT scan of the same patient to estimate the positions of interstitial disease in a chest radiograph and, consequently, label each pixel within the lung fields on a radiograph either normal or affected by ILD.

The labels are used to train the CAD system on a subset of our x-ray database (training data). In the rest of the database (test data) the labels are used as the reference standard for evaluation of the system performance. When the trained system is applied to a new chest radiograph, a probability of being abnormal is assigned to each lung pixel. A CAD outcome, either as a color-coded probability map or as regional scores, can be presented to a radiologist as an assisting tool. The performance is evaluated by means of ROC analysis and compared with the performance of two human observers on the same set of data.

The paper is organized as follows: section 4.2 describes the data and the CAD

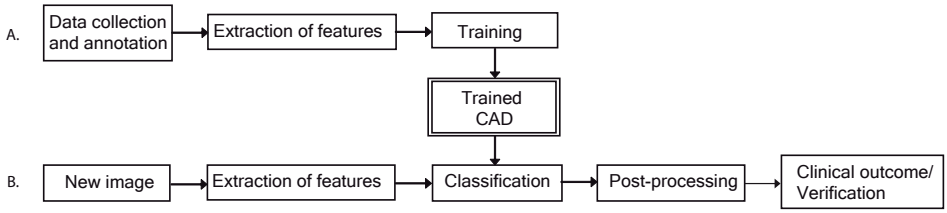
system starting with the system outline and the methodology of data collection and continuing with detailing of different parts of the system. Section 4.3 gives details of experimental setup and evaluation methods. In section 4.4 the results are presented. They are discussed in section 4.5, and conclusions are drawn in section 4.6.

4.2 Materials and Methods

4.2.1 System outline

From an engineering point of view, most CAD systems, including the one presented in this paper, have a typical design relying on a combination of image-processing and pattern recognition or artificial intelligence techniques. In Figure 4.1 the scheme of our system is depicted. In the training phase (Figure 4.1, row A) image data that represents the diversity of ILD manifestations is collected, together with normal image data. Images with pathology are annotated, i.e. lesions are delineated and given a subtlety rank. These are training examples used to train the CAD system to distinguish between normal and abnormal patterns. To be able to train the system, pixels within the lung fields are represented by vectors of features computed from their neighborhood. After extraction of features a statistical decision model is constructed. The trained CAD system can yield an opinion about the presence of ILD lesions in a new image that was not included in the set of training examples. Pixels from a new chest radiograph, also represented by feature vectors, are classified according to the decision model and receive a probability of being abnormal (depicted in Figure 4.1, row B). In the post-processing stage a probability map is produced that accentuates areas with a high probability of being affected by ILD. If abnormality annotations (a reference standard) for this image are available, the outcome of the system can be verified.

Figure 4.1: *Flow chart of the CAD system. (A) Training phase. (B) Testing phase.*



4.2.2 Data set

For the study presented here we collected a number of digital posterior-anterior (PA) chest radiographs from the Picture Archiving and Communication System

(PACS) of University Medical Center Utrecht, the Netherlands. These images were acquired in a daily clinical practice between 2004 and 2006. Direct radiography units (Digital Diagnost, Philips Medical Systems) with a cesium iodine scintillator, a matrix of 3000×3000 pixel and 0.143 mm pixel size were used for acquisition. Images were exported to the PACS with 15 bits data depth. The PACS provides a single point of entry for all images and their associated data related to the same patient. In this study, patient data was treated in accordance with the Declaration of Helsinki. All patients registered in this system between 2004 and 2006 were considered, and among them chest radiographs were selected based on two criteria.

Firstly, for a patient with a chest radiograph, a multislice chest CT scan, which was taken within one month before or after the x-ray examination, was also required. This time interval was chosen by an expert radiologist since the majority of interstitial diseases are known to progress rather slowly.

Secondly, for an x-ray to be classed as normal, radiological reports associated with both images (x-ray and CT) were required to clearly indicate healthy lungs. For an x-ray to be classed as abnormal, either both reports or the CT report were required to refer to ILD or describe textures typical for ILD.

All CT images selected for the study were acquired on one of several multislice scanners (Philips Medical Systems, the Netherlands), namely, Brilliance-16P, Brilliance-40, Brilliance-64 and Mx8000 IDT 16, with standard parameters for high-resolution volumetric CT scanning. Collimation varied between 0.625 mm (40- and 64-slice scanners) and 0.75 mm (16-slice). Images of 0.9 mm thickness (40- and 64-slice) or 1 mm thickness (16-slice) were reconstructed every 0.7 mm. Exposure settings were 120 kVp and between 100 and 170 mAs, depending on a scanner and patient size.

Normal and abnormal radiographs, selected in this manner, together with accompanying CT scans were subsequently examined by an expert chest radiologist (MP, the second author, with more than 15 years of experience) who decided whether a chest x-ray and the corresponding multislice CT scan were normal (not containing interstitial abnormalities) or abnormal (containing signs of ILD), and whether the extent of abnormality was similar in both modalities. We included in the study the following types of chronic interstitial abnormalities:

- focal pulmonary opacities,
- diffuse interstitial reticular or linear changes,
- diffuse nodular changes,
- diffuse changes with increased parenchymal density,
- pulmonary diseases with cystic changes.

Diffuse lung changes with decreased density were excluded.

From the initially selected 50 cases 6 cases were excluded by the radiologist. They were excluded either because of an unclear diagnosis (the radiologist did not agree with the reported presence of ILD) or owing to a varying amount of disease manifestation on the CT scan and radiograph. The final set of abnormal images contained 44 cases. The average patient age was 58 (range 26–87 years, standard deviation 15 years). There is a higher prevalence of ILD in older patients. The gender distribution of patients was 21 males and 23 females.

After delineation of pathology in the abnormal images, 8 normal radiographs were added to balance the total quantity of normal and abnormal tissue in the data. There were 4 males and 4 females, with an average age of 46 (range 20–81 years, standard deviation 19 years).

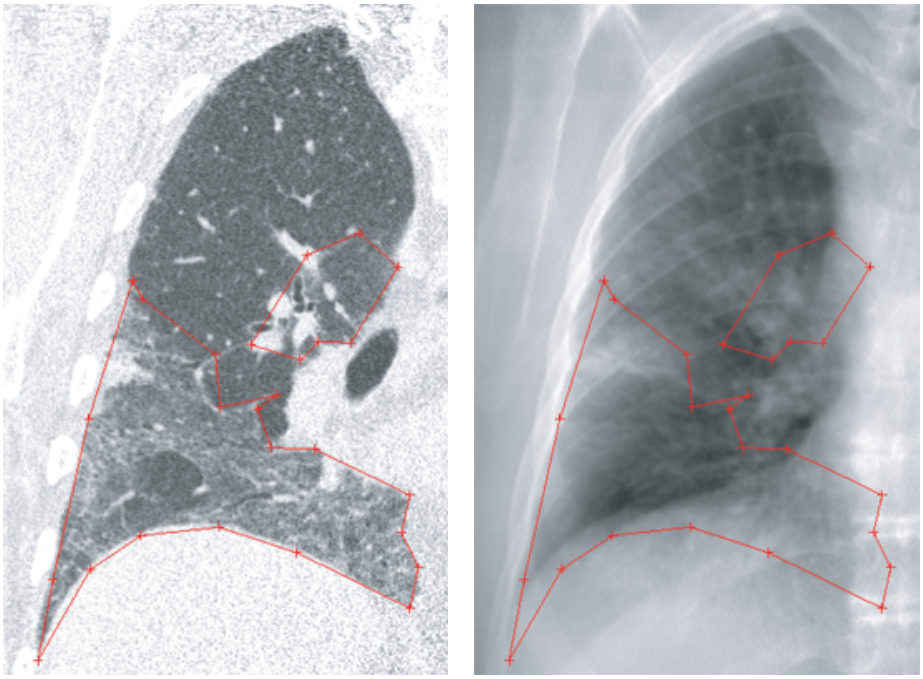
4.2.3 Reference standard

A multislice CT scan accompanying each chest radiograph is not only used to confirm a diagnosis but also to set up a reference standard on the corresponding radiograph. Thin section width and overlapping image reconstruction of multislice CT result in good quality 2D image reformations in all directions. Moreover, such a 2D view is non-superimposed and has excellent contrast resolution. In Figure 4.2, a one-voxel thick coronal plane of a CT scan (Figure 4.2a) is compared with a PA chest radiograph of the same patient (Figure 4.2c). Note that pathological areas stand out much clearer against the background lung tissue on the CT slice than on the radiograph. Manual, and even automatic (e.g. in [88]), segmentation of interstitial abnormalities on CT slices is feasible whereas definite borders between pathological and normal lung tissue in conventional radiographs can barely be found. This point is clearly illustrated in Figure 4.2.

The proposed method uses CT data as a superior gold standard and is based on delineation of abnormalities on coronal CT sections (the same orientation as the conventional chest radiograph) which are then transferred to the corresponding radiograph. Thus we circumvent the inability to segment abnormalities directly in radiographs.

Note that the annotation of radiographs with use of CT scans of the same patients happens during the collection of training data in order to obtain a set of well-annotated radiographs that can be used to train a CAD system. Once the CAD system is trained the analysis of a new radiograph does not require a supporting CT scan. For the test data we used the same method to establish a reliable reference standard in order to evaluate the system performance.

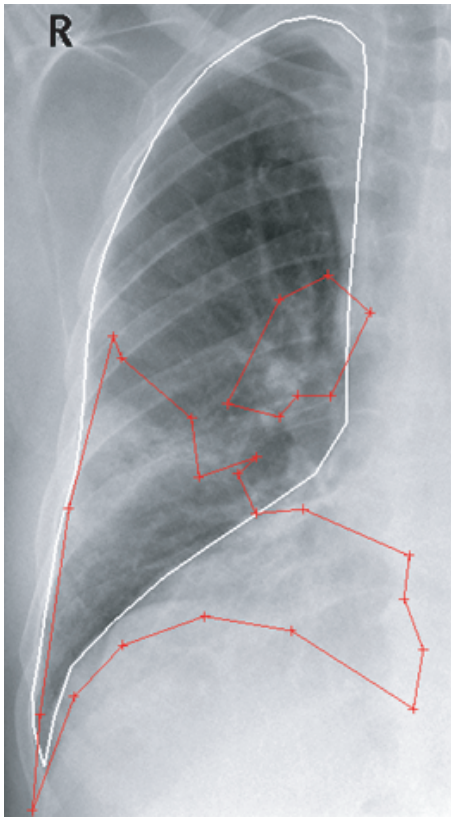
For each pair of an abnormal x-ray and a CT scan, interstitial abnormalities were manually delineated by the expert chest radiologist (MP) with a dedicated computer program built for this study. Delineations were performed on single, 0.7 mm thick coronal slices selected at every 10 mm. In order to translate delineations to an x-ray a mapping function is established between a coronal projection from the CT volume and the x-ray. The coronal projection is obtained for this purpose by averaging CT numbers in the coronal direction. In this way the coronal



(a) *A lesion delineation on a coronal CT slice.*

(b) *Its corresponding position on the coronal CT projection.*

Figure 4.2: *An example of an interstitial lesion delineation. The borders between normal and abnormal texture are clearly visible on a CT slice (a) in contrast to an indistinct border on an x-ray (c). In order to segment all abnormal areas on an x-ray, several delineations have to be made on different CT slices that will combine into one or more lesion delineations on an x-ray.*



(c) *The result of mapping the lesions outlined on (a) and (b) to an x-ray of the same patient.*

(d) *A final abnormality delineation. Colors of outlines indicate zones of different abnormality subtlety. Red corresponds to obviously abnormal areas, pink - to relatively abnormal ones, while cyan and blue denote subtle and very subtle abnormalities.*

Figure 4.2: *continued. A combined projection of all lesion delineations made on CT slices is mapped onto an x-ray (in (d)). An abnormality segmentation obtained in this way is divided into areas of different abnormality subtlety. The lung field is outlined in white. Areas lying within abnormality delineations but outside the lung fields are not considered in our system.*

projection approximates the radiograph (see Figure 4.2b). Deformation between the CT projection and the radiograph is found using radial basis functions as described in [57]. This method requires a set of known corresponding points (control points). The same radiologist who segmented abnormalities indicated several (from 6 to 10, depending on the image) anatomically corresponding landmarks in the radiograph and CT projection to be used as control points. The mapping function is constructed based on the control points and applied to the vertices of the abnormality outlines. As a result, the corresponding outlines in the radiograph are obtained (see Figure 4.2c). Their shapes can be corrected manually, if deemed necessary. Superimposed outlines were replaced by their union. Usually, the final delineation of a lesion on the x-ray corresponds to a combination of several delineations made on different CT slices.

The radiologist made no corrections to outlines transferred to the radiographs. However, in 14 cases one or both lungs were completely affected by ILD, and per slice delineations deemed unnecessary. In those cases the lung boundaries in a radiograph were used to define an abnormal area. In 10 cases the final delineations were slightly corrected for smoothness.

In Figure 4.2d an example of a final outcome of the described segmentation procedure is presented. Additionally, the abnormal areas in each radiograph were divided into areas of different abnormality subtlety of disease. This was done by the same radiologist who also had defined the abnormality areas based on the CT findings. For subtlety assigning no information from the CT scan was used. The expert radiologist's judgement was based on his visual assessment of the radiograph. Four levels of subtlety in detection of an interstitial abnormality are recognized: a) obvious (detection of abnormality is easy); b) relatively obvious (detection is relatively easy); c) subtle (detection is difficult); d) very subtle (detection is very difficult, almost impossible). These categories are further used to evaluate the performance of the CAD system and human observers on areas that differ in the visible amount of ILD signs.

4.2.4 Features

Segmentation

In order to analyze lung fields, they have to be segmented from the rest of the radiograph. For this study the lung fields were delineated manually. In spite of the existence and availability of automatic segmentation methods for PA chest radiographs on the research site (e.g. see [25]), these supervised methods were previously trained with digitized films and appeared to perform imperfectly when applied to digital images.

Feature extraction

Since ILD mostly manifests itself in radiographs through a distortion of the normal appearance of the lung texture, it is important to extract discriminative texture

features. A powerful method for local texture analysis is filtering the image with a multiscale filter bank of Gaussian derivatives and calculating the moments of histograms from regions in the derived images. Using multiple scales allows us to characterize texture elements of different sizes, and analysis of local histograms considers the texture primitives regardless of their spacial distribution. This is a general approach to texture characterization [72, 73]. The histogram moments were successfully used for automatic detection of interstitial abnormalities in chest radiographs in [66, 74], and by Sluimer *et al.* [75] for texture analysis in high resolution thoracic CT.

In our work, prior to filtering the image, pixel values are mirrored with respect to the lung borders. Namely, a pixel value outside the lungs is substituted for its counterpart inside the lungs with the nearest pixel on the lung contour as a center of symmetry. This step is taken to avoid a major distortion in the filter output near the lung borders which is normally caused by a large difference in appearance inside and outside the lungs. Next, the left lung is flipped to resemble the right lung. Chest radiographs are filtered with Gaussian derivatives of orders $0, 1$ and 2 at five scales, $\sigma = 1, 2, 4, 8, 16$.

$$\begin{aligned} G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad &- \text{ zero order Gaussian derivative} \\ G_x(x,y), \quad G_y(x,y) \quad &- \text{ 1st order Gaussian derivatives} \\ G_{xx}(x,y), \quad G_{xy}(x,y), \quad G_{yy}(x,y) \quad &- \text{ 2nd order Gaussian derivatives} \end{aligned}$$

Four central moments of the pixel intensity distribution, namely, the mean, standard deviation, skewness and kurtosis are calculated from a circular neighborhood of selected pixels (pixels of interest, POIs). These are pixels lying on a 10x10 grid within the lung fields. The features are computed from all the filtered images and from the original image. The radius of the neighborhood is chosen to be 128 pixels for the final system setup. Other radii (64, 96 and 160) appeared slightly less successful in pilot classification experiments. The chosen radius also approximates the sizes of regions in [76]. Unlike [76], in this work the local analysis is performed on automatically selected regions that covered the lung fields completely. Two position features are added to the feature set, namely, the x and y coordinates of the POI relative to the center of mass of the lung containing it. The position features are scaled to have unit standard deviation per image. In total, 126 features are extracted for each POI.

4.2.5 Classification

In the next step a soft classification of POIs is performed in the feature space. Prior to classification, features are normalized. In a training set, each feature is translated and scaled to have zero mean and unit standard deviation. Then the same normalization parameters are applied to feature vectors of test samples.

A supervised classification method is used, which means that a classification function (a classifier) is first trained on labeled samples from both normal and

abnormal classes. There is a wide choice of classifiers available in the literature with no superior learning method overall (e.g. see [23] or [24]). The type of problem and prior knowledge determine which classifier provides a better performance. Since no prior knowledge was available we evaluated and compared several different classifiers. For convenience, we restricted ourselves to four different types of classifiers, namely, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), a k-nearest-neighbors classifier (k-NN), and a support vector machine (SVM).

Both LDA and QDA assume Gaussian distributions for the samples of each class. The LDA additionally assumes equal covariance matrices for each distribution. The k-NN is a non-parametric classifier, with a free parameter k that has to be found empirically. In the k-NN rule, the posterior probability for each of the classes is estimated by the fraction of training samples among the k nearest neighbors of a test sample that belong to that class. In this work, the fast implementation of the k-NN classifier by Arya and Mount [89] was used.

The SVM is a family of classifiers that has gained popularity in recent years. In the case of an ideal linear separability of a training set the SVM finds an optimal discriminative plane by maximizing the margin between the nearest samples, also known as support vectors, of both classes. For linear non-separable data, the plane found by the SVM is a tradeoff, controlled by a penalty parameter, between the classification error on the training set and margin maximization. A useful feature of the SVM is that this method can be kernelized if linear discriminants are not appropriate for a given data set. By mapping original feature vectors into a higher-dimensional feature space and solving an SVM optimization problem there, a highly nonlinear classification function can be obtained in the original feature space. The success of the SVM for a particular classification problem depends on a correctly estimated penalty parameter and a suitable kernel function. For the SVM implementation, the LIBSVM library was used [53].

4.2.6 Post-processing

The classification of an image results in the estimation of posterior probabilities for POIs. To convert this into a probability map we compute new pixel posterior probabilities by averaging posterior probabilities of neighboring POIs. For each pixel i in the lungs, including a POI, its posterior probability p_i is calculated as

$$p_i = \frac{1}{N_{R_i}} \sum_{r \in R_i} p_r^c,$$

where R_i is a neighborhood of i, p_r^c is an estimated by a soft classification posterior probability in a POI $r, r \in R_i$, and N_{R_i} is the number of all POIs lying in the neighborhood R_i. The neighborhood is defined in the same way as the one used to calculate the features.

4.3 Experiments

4.3.1 Cross validation

All experiments were performed by cross validation. We randomly divided 52 images into four folds, with the condition that normal images and images containing abnormalities were equally spread among the folds (2 normal and 11 pathological radiographs in each fold). The CAD system made 4 iterations. In each iteration a different fold was used as the test set and three other folds together as the training set. This setup guaranteed the optimal use of the available data, as well as an unbiased evaluation, because at no time did training and test sets contain samples originating from the same images.

4.3.2 Generation of training set

As mentioned in section 4.2.3, interstitial lesions were divided into four different categories of abnormality subtlety. A straightforward approach would be to use samples from all four categories to train the CAD system. However pilot experiments showed that this approach would not give the best possible performance. The best performance was achieved when the system was trained with normal samples taken from both normal images and images containing ILD lesions, and abnormal samples from the 'obvious' category. Approximately the same performance was reached when the abnormal class was represented by samples from both 'obvious' and 'relatively obvious' categories. Adding samples with less pronounced abnormalities worsened performance. Therefore all results in this paper were obtained with training sets that contained only normal and obviously abnormal samples. Approximately half of the normal samples in the training set in each fold came from normal training images, and the other half came from normal parts of abnormal training images. Note that for the evaluation of the system no selection of test samples was made. Normal samples from images containing some abnormality were included in the training set only if they were calculated from neighborhoods that did not overlap with any outlined lesions.

4.3.3 Choice of system parameters

The k-NN and SVM classifiers require some parameter tuning. For the k-NN, the parameter $k = 39$ was chosen experimentally, with negligible differences in the system performance for the whole range of k between 25 and 45. For the SVM, we considered the radial basis function (RBF) kernel. The SVM requires a long tuning and training time, therefore the number of samples in the training set was reduced by random sub-sampling. The penalty and kernel parameters were found using a five-fold cross validation grid search on the training set.

No feature selection was applied in the final experimental setup. In pilot experiments we found out that the system did not gain in performance when run

with a subset of features selected by the standard approaches, like Sequential Forward Search and Sequential Backward Search.

4.3.4 Evaluation

The evaluation of pixel classification was done by means of the receiver operating characteristic (ROC) analysis [29]. The ROC curve plots the sensitivity against 1-specificity of a system at varying confidence thresholds. A_z, the area under an ROC curve, was used as a classification performance measure. Outcomes of all iterations of the cross validation were analyzed together yielding a single ROC curve that estimated an overall system performance.

An individual ROC curve was computed at each of four levels of subtlety, with abnormal samples of that subtlety level as positives and all normal samples as negatives. Such analysis may yield better understanding of the relationship between the abnormality subtlety and the detection abilities of the system or humans. Additionally, a generalized ROC curve was calculated that considered all abnormal samples together as positives.

4.3.5 Observer study

An observer study was performed that closely resembles usual clinical routine. Each lung field was automatically divided into 4 equal-sized regions (see Figure 4.3), altogether 8 regions per radiograph. The observers were asked to diagnose each region separately, stating whether it contained any interstitial abnormalities.

The division did not accurately correspond to any anatomical landmarks but was guided by the notion that the top, perihilum, middle periphery and bottom of a normal lung field exhibit different textural patterns. According to our division algorithm, the region around hilum (perihilum) included lung pixels overlapping with a circle placed at the lungs' center of mass. The radius of the circle was separately chosen for the left and right lung, such that the overlap covered one quarter of the pixels of that lung. The rest of the lung field was horizontally divided into three equal-sized parts – the top, middle and bottom regions.

The observers assessed regions using the discrete scale of grades from 1 to 5, where 1 corresponded to 'normal' (a region looked completely normal, or the amount of ILD was negligible, i.e. less than 10% of the region area) and 5 to 'obviously abnormal' (more than 10% of the region area clearly contained interstitial abnormalities). The grades 2, 3 and 4 corresponded to intermediate levels of observers' certainty whether a region contained interstitial abnormalities. The 10% threshold was a subjective choice of the expert chest radiologist (MP). According to his experience, setting a lower threshold would cause a large inter-observer variability.

In order to compare the human performance with the CAD system, region scores for the system were computed from the probability maps by averaging

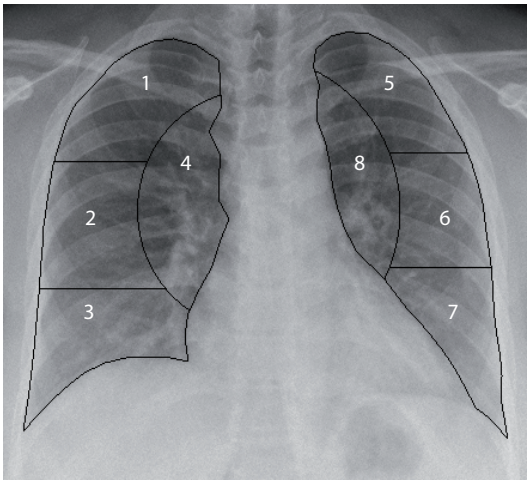


Figure 4.3: *An example of the lung fields automatically divided in eight regions for the observer study. The regions are: (1) and (5) - top; (2) and (6) - middle periphery; (3) and (7) - bottom; (4) and (8) - perihilum.*

posterior probabilities within each region. The reference standard was also converted into region labels. A label for each region was determined based on the total amount of abnormal tissue present in that region. If less than 10% of the region area fell within any abnormality outline then the region was considered normal. Otherwise the region was considered abnormal. Similarly to the pixel classification, a subtlety category was assigned to the region in order to evaluate the system performance on regions of different complexity. Such a category was assigned in accordance with the most frequent subtlety type in the region. For example, if the majority of abnormal pixels in the region had been previously ranked as 'subtle' than the region received the label 'subtle'. The quantitative distribution of regions among the normal class and different abnormality subtlety categories is shown in Table 4.1.

With the reference standard obtained for each region ROC analysis becomes possible. It is performed at each of four levels of abnormality subtlety, similarly to the evaluation of pixel classification. Four ROC curves are computed, each considering a subset of abnormal regions that have a certain subtlety level as positives, and all normal regions as negatives. Thus, a number of false positives is the same for each of the curves, while a number of false negatives vary depending on which subset of abnormal images is considered. An overall ROC curve was also calculated considering regions of any abnormality subtlety as positives.

Table 4.1: *The numbers of normal regions and regions of different abnormality subtlety is presented in the first column. A distribution limited to peripheral regions only is presented in the second column.*

Abnormality subtlety	**All regions**	**Excluding perihilum**
Normal	117	91
Obvious	119	74
Relatively Obvious	79	60
Subtle	75	62
Very Subtle	26	25
All categories	416	312

Table 4.2: *CAD performances for different classifiers in terms of the area under the ROC curve. The performance is estimated separately for different abnormality subtleties vs. normal class, as well as for all abnormality types together vs. normal class.*

Abnormality subtlety	**Classifier**			
	LDA	QDA	kNN	SVM
Obvious	0.90	0.84	0.88	0.90
Relatively Obvious	0.77	0.71	0.73	0.77
Subtle	0.65	0.61	0.62	0.66
Very Subtle	0.58	0.52	0.55	0.59
All categories	0.78	0.73	0.76	0.78

4.4 Results

For the pixel classification, the CAD system was trained and tested with each of four classifiers described in section 4.2.5. The values of A_z for different classifiers and different levels of abnormality subtlety are listed in Table 4.2. It is shown that the CAD outcomes are comparable for the LDA and SVM. Both classifiers outperform the QDA and k-NN. Further on in this paper we refer only to the system that uses the LDA, because the LDA is a simpler classifier than SVM. ROC curves measuring the performance of the system with the LDA are plotted in Figure 4.4. The curves are clearly distinguished for the different degrees of the abnormality subtlety.

4.4.1 Probability maps

The output of the CAD system is a probability map that assigns to each pixel in the lung fields a probability p to being abnormal, $0 \leq p \leq 1$. For each of 52 radiographs in the data set such a map has been generated by the system. The

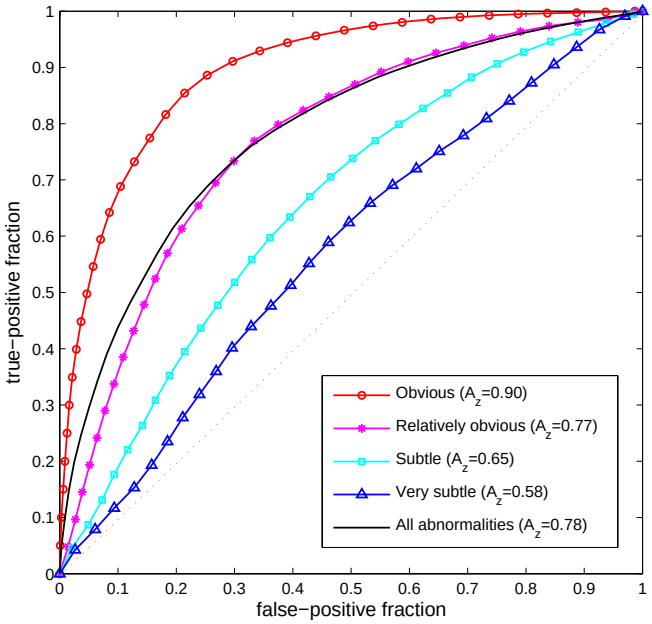


Figure 4.4: *ROC curves for the evaluation of pixel classification using the linear discriminant analysis. Curves are plotted for different abnormality subtleties vs. normal class.*

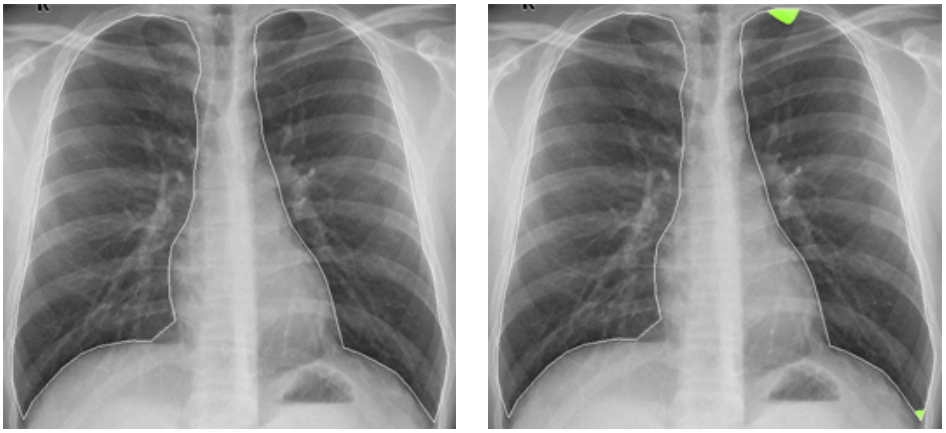
map is visualized as a color-coded overlay, where a certain color corresponds to a articular probability range. In Figure 4.5 examples of normal and abnormal radiographs and their probability maps are shown. The color-coded overlays in all maps are thresholded i.e. pixels within the lung fields that have been assigned a posterior probability lower than a threshold value $p = 0.15$ stay transparent. After sorting probability maps according to their mean probabilities, separately for normal and abnormal images, three examples from each image class have been randomly chosen from the lower, middle and upper parts of the mean probability range. Figures 4.5a, 4.5f and 4.5e are the examples of one normal and two abnormal radiographs for which the system output matches the reference standard well. In Figure 4.5d the findings of the system are misplaced, and the opacity in the bottom of the left lung is not found. The probability maps in Figures 4.5b and 4.5c demonstrate the most common mistake that the system makes, namely, misclassification of the perihilar regions.

4.4.2 CAD performance compared to human observers

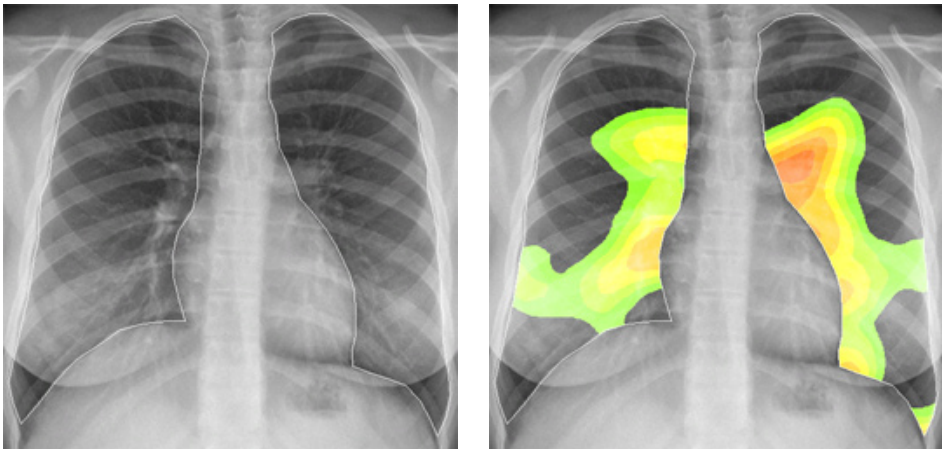
We compared the system performance at region level to the performance of two human observers. One observer was an experienced chest radiologist (CSP, more than 15 years of experience) and one was a chest radiologist in training (PdJ). They were not involved in setting the reference standard for the data in this study. The observers were presented the same set of 52 posterior-anterior chest radiographs. Normal and abnormal images were shuffled and presented in no particular order. The observers reviewed the radiographs using dedicated medical displays (Barco Medical Imaging Systems, Belgium), namely, MFGD 3220D (3MP, 10-bit, 2048×1536 native resolution), comparable to displays they would normally use in their clinical practice.

The values of A_z for both observers and the CAD system are presented in Table 4.3. The results listed in the first three columns were obtained from the evaluation of all eight lung subdivisions. In the last three columns the corresponding performances are shown for the case when the perihilar regions in each radiograph were excluded from evaluation. The statistical analysis described in [90] was applied to compare A_z of the system with that of each of the observers.

The last row of the Table 4.3) demonstrates that both observers and the system performed not significantly different in distinguishing abnormal regions from normal ones when the perihilar regions were excluded from evaluation. It is also shown that the poorer performance on the perihilar regions caused the system to be significantly worse ($P < 0.05$) than the human observers. The humans were significantly better ($P < 0.01$) with slightly abnormal regions throughout lungs, but this difference became smaller when only peripheral regions were considered – the second observer still performed significantly better ($P < 0.05$) than the system, while the difference between the system and the first observer became insignificant. Statistical analysis for very slightly abnormal regions did not show any significant differences in performance. The latter may be attributed to a



(a) *A normal radiograph that the system considered overall normal (mean probability 0.05, 98% percentile value 0.19).*



(b) *A normal radiograph where the system found some suspicious areas (mean probability 0.25, 98% percentile value 0.70).*

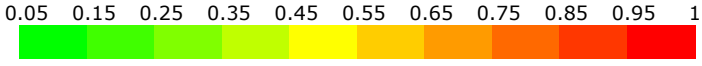
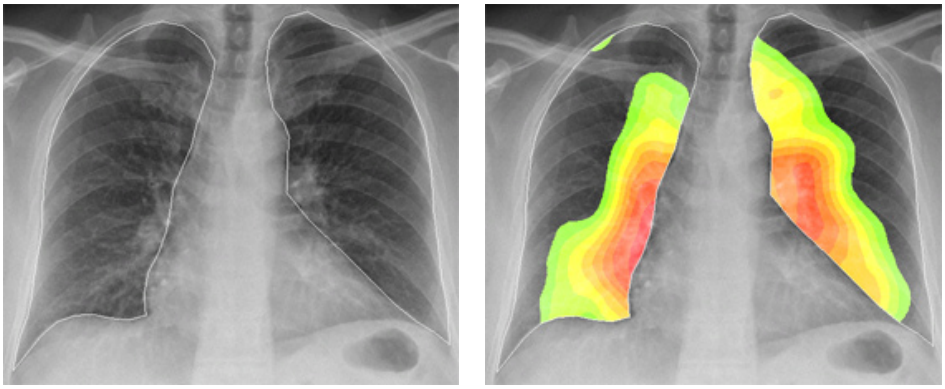
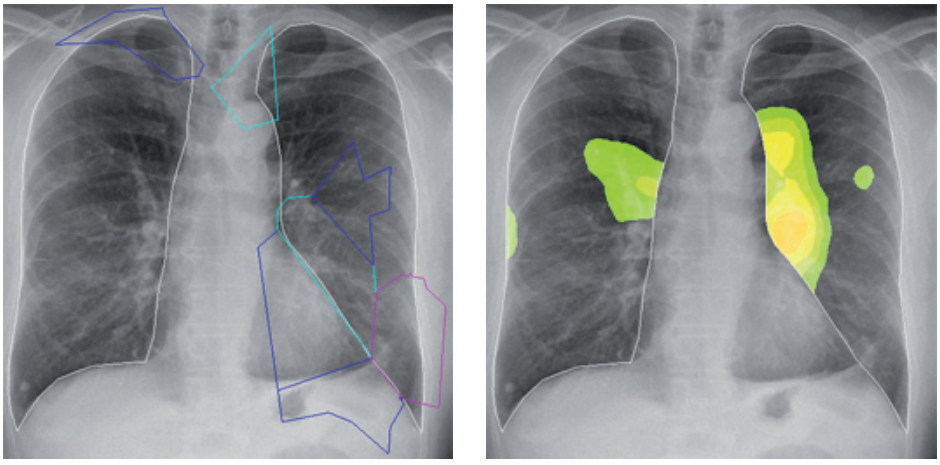


Figure 4.5: *The examples of chest radiographs and corresponding probability maps produced by the CAD system. The left column depicts an original radiograph, with a reference standard for abnormal images. Lung contours are outlined in white. In the probability maps (the right column) only pixels with a posterior probability $p > 0.15$ are shown for convenience. The color bar explains the correspondence between probability ranges and colors.*



(c) *A normal radiograph where the system found more abnormalities and some acutely abnormal areas (mean probability 0.32, 98% percentile value 0.91).*



(d) *An abnormal radiograph where the system did not find any prominent abnormalities (mean probability 0.14, 98% percentile value 0.52).*

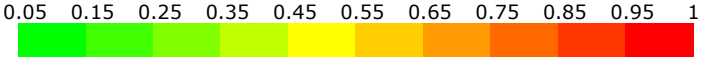
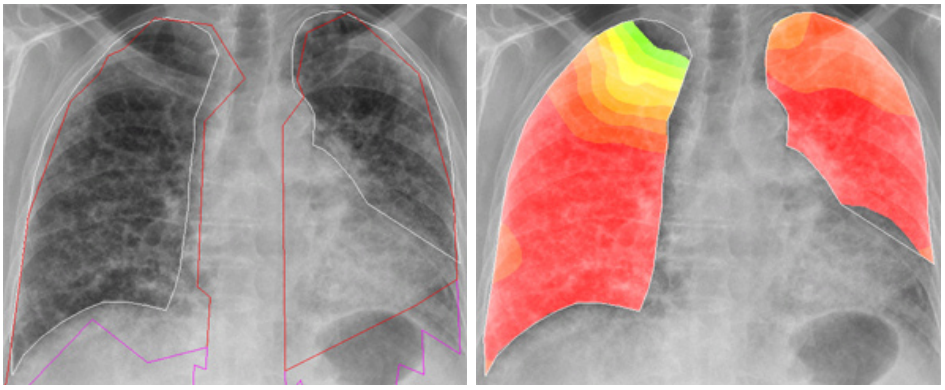
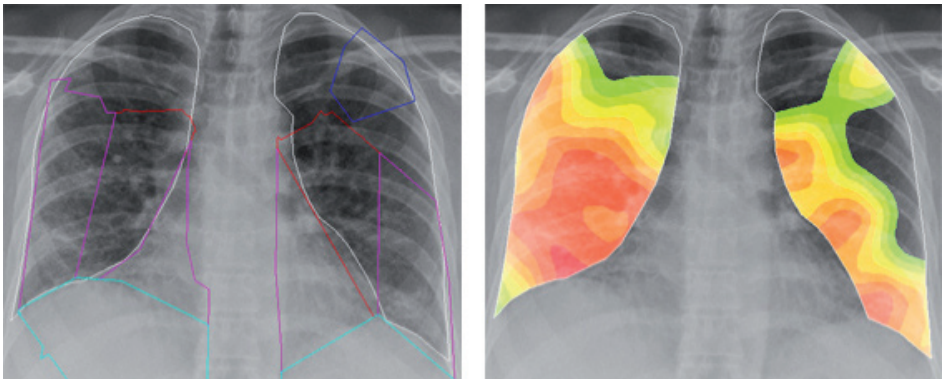


Figure 4.5: *continued.*



(e) *An abnormal radiograph that the system found very abnormal overall (mean probability 0.90, 98% percentile value 0.99).*



(f) *An abnormal radiograph that the system considered partially abnormal, with some spots receiving high probabilities of being abnormal (mean probability 0.51, 98% percentile value 0.94).*

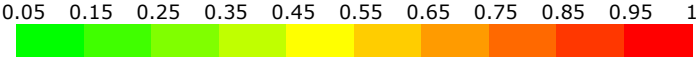


Figure 4.5: *continued.*

Table 4.3: *Performance on regions in terms of the area under the ROC curve. The 1st observer (O1) is an expert chest radiologist, the 2nd observer (O2) is a chest radiologist in training. The performance is estimated separately for different levels of abnormality subtlety vs. normal class, as well as for all abnormality types together vs. normal class. Significantly different human performances are marked with an asterisk (a two-tailed test, at significance level of 5%) or a double asterisk (at significance level of 1%).*

Abnormality subtlety	**All regions**			**Excluding perihilum**		
	CAD	O1	O2	CAD	O1	O2
Obvious	0.92	0.93	0.94	0.96	0.96	0.96
Relatively obvious	0.81	0.85	0.86	0.87	0.87	0.88
Subtle	0.67	0.80**	0.81**	0.73	0.83	0.85*
Very subtle	0.67	0.76	0.71	0.74	0.78	0.69
All categories	0.80	0.86*	0.87*	0.85	0.88	0.88

low sample size of very slightly abnormal regions (see Table 4.1). Moreover, the second observer appeared to perform worse than the system when the perihilar regions were excluded. There were no significant differences in AUC values with obviously and relatively obviously abnormal regions.

4.5 Discussion

We undertook this work in order to provide radiologists with a tool assisting them with the task of finding textural abnormalities in conventional chest radiographs. The complexity of the task of differentiation between normal lung tissue and areas affected by ILD is well illustrated in Figures 4.5 and 4.6.

After training with the annotated data our system assigned a probability to be abnormal to each pixel within the lung fields on the radiograph. The ROC analysis showed that pixel classification results were not reliable for subtle and very subtle areas of abnormality (in Figure 4.4). One of possible reasons for that is the informatively superior nature of our reference standard which was obtained with use of CT known to be a more descriptive modality than conventional radiography. This reason is supported by our pilot experiments mentioned in section 4.3.2, when adding subtly and very subtly abnormal pixels to the training set worsened the system performance.

In future, an additional investigation should be undertaken to identify more power features to cope with subtle textural differences. Furthermore, the use of a larger training database might improve the classification of subtle abnormalities.

However, one should not be discouraged by the low system performance on very subtle abnormalities. The overall classification result was not bad ($A_z = 0.78$) taken into account the superior reference standard. And the pixel probability

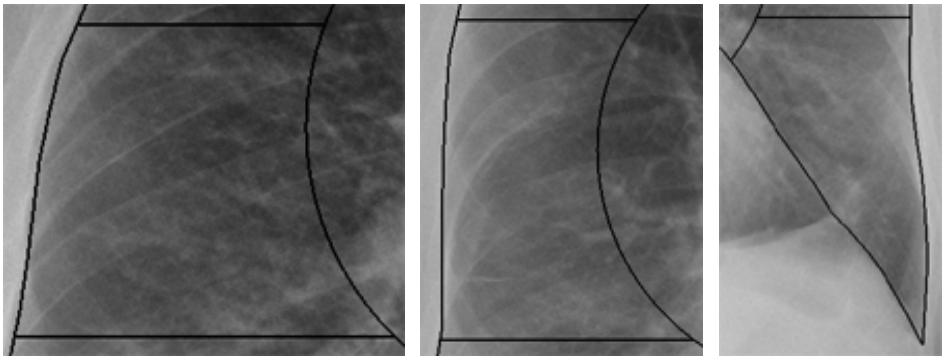


Figure 4.6: *Examples of regions where the CAD system and (one of) the observers disagreed, and either the system or an observer misclassified the region. On the left and right, regions with interstitial abnormalities are shown, the region shown in the middle is normal. The CAD system was correct about the regions on the left and in the middle and missed an abnormality in the region on the right.*

maps are still useful as graphic presentations of the output of the system, even if lesions are only partly found. Such maps give a general idea where, according to the system, abnormalities are situated, and can be conveniently consulted by radiologists as the second opinion.

Averaging posterior probabilities over regions improved the system classification performance (in Table 4.3) and made possible a comparison with human observers. Not only the CAD system but also the observers quite often interpreted radiological findings erroneously as it is seen from Table 4.3. Still, both our observers performed significantly better than CAD on the subtle abnormalities. However, CAD was comparable or even better than the human observers on the very subtle lesions (when the perihilum was excluded), and also on the relatively obvious and obvious lesions. Moreover, the system and the observers relatively often made complementary mistakes, which means some regions were correctly classified by the CAD system and misclassified by one or both observers, and vise versa. This point is illustrated in Figure 4.6, which shows examples of regions where the CAD output and the human opinion disagreed. It implies that even the output from an imperfect system might be used by a radiologist as an advisory vote, as long as the radiologist understands its limitations.

Evaluation of the utility of our probability maps or regions scores in improving the detection performance of humans is a pertinent continuation of this study.

The performance evaluation for different regions and different degrees of abnormality subtlety would be impossible without the local reference standard. Our system uses a new method to obtain a superior reference standard for the estimation of the position and extension of interstitial lesions in the lung fields. It is a semi-automatic method that involves manual segmentations on CT sections.

Although the manual delineation of abnormalities on thin CT sections is more reliable than that on conventional radiographs, it still introduces inherent subjectivity into the pixel-based reference standard, especially near the boundaries of abnormal areas. We hypothesize that differences in segmentation, when performed by different radiologists, might be averaged out to a large extent in the final projection on to the radiograph. However, it is an open question how strong the influence of the manual part of our method could be on the system output. We leave this analysis for future research. One possible extension of our method is an automatic segmentation of lesions on CT sections. The starting point for that could be, for example, a method proposed in [88] that generates regions containing homogeneous texture.

Another potential source of inaccuracies in the reference standard is the mapping function which might introduce errors while transferring abnormality outlines from a CT projection to a corresponding radiograph. We ensured the minimization of such errors by providing an auxiliary tool that enabled a radiologist to test the accuracy of matching between two images preliminary to performing segmentations. Based on the visual correspondence of test shapes, control points could be moved or added until a visually satisfactory matching was obtained. The fact that only a limited number of corrections were made to the obtained outlines suggests that errors possibly introduced by the mapping function were limited.

The system demonstrated in this paper yields promising results but has considerable room for improvement in the perihilar regions of the lungs. This is the region where the bronchi and blood vessels enter the lung. It is a difficult area for texture analysis not only because of its bright and pattern-rich manifestation but also because its normal appearance has large individual variability. Moreover, our database does not contain many images with healthy perihilum since ILD frequently involves this area. Among 44 abnormal images only 5 had a normal perihilum in one or both lungs. Taking into account that only 6 absolutely normal images were present in each training set we might conclude that the training set may not be representative of normal varieties of the perihilum. Extension of the database towards inclusion of more normal representatives of the perihilum might improve the system performance even without further modification of the underlying algorithm. In addition, when a radiologist makes a decision as to whether a perihilum looks normal or not, he or she pays attention to its size and shape features along with the textural signs of abnormality. Perhaps, perihilum classification should become a separate component of an automated system and include perihilar shape and size analysis as well.

4.6 Conclusion

In this paper a computer-aided diagnosis system was presented for the task of the detection and localization of interstitial abnormalities in chest radiographs. The system was built using a supervised pattern recognition approach. As an output,

the system produced a map of posterior probabilities, where each pixel inside the lung fields received a probability of being abnormal.

We collected and annotated a unique database of digital chest radiographs containing ILD abnormalities. A novel method was developed to define a reference standard on the abnormal radiographs. This method utilized a computed tomography scan of the same patient and automatically translated manual delineations of abnormalities made on a subset of thin coronal sections to the corresponding radiograph.

Our CAD system employed local statistical features calculated from filtered images. The filters were the first and second order Gaussian derivatives at multiple scales. A linear discriminant classifier and support vector machine yielded the best classification performances. The evaluation was done by means of ROC analysis for different levels of abnormality subtlety. It was shown that the system was considerably better in distinguishing obviously abnormal pixels from normal ones than in distinguishing between very slightly abnormal and normal pixels. This is likely due to an over-informative nature of the reference standard that we compared our findings with.

The system performance was also compared with that of an expert radiologist and a radiologist in training. The system was shown to perform significantly worse than both observers on slightly abnormal regions and all abnormalities together, with no significant differences in the detection of obviously, relatively obviously and very slightly abnormal regions. Moreover, the system was shown to approach the human performance in the detection of abnormalities when the perihilar regions were excluded from evaluation.

Chapter 5

Estimation of progression of interstitial lung disease in computed tomography images

Y. Arzhaeva, M. Prokop, K. Murphy, E.M. van Rikxoort, P.A. de Jong, H.A. Gietema, M.A. Viergever and B. van Ginneken, "Automated estimation of progression of interstitial lung disease in CT images", *submitted*.

Abstract

A system is presented for automated estimation of progression of interstitial lung disease in serial thoracic CT scans. The system compares corresponding 2D axial sections from baseline and follow-up scans and concludes whether this pair of sections represents regression, progression or unchanged disease status. The correspondence between serial CT scans is achieved by intra-patient volumetric image registration. The system classification function is trained with two different feature sets. Features in the first set represent the intensity distribution of a difference image between the baseline and follow-up CT sections. Features in the second set represent dissimilarities computed between the baseline and follow-up images filtered with a bank of general purpose texture filters. The performance of our system is compared with that of two radiologists. In an experiment on 74 scan pairs, the performance of the system using either feature set is shown not to be significantly different from the performance of either observer using McNemar's test. In terms of accuracy, system performance is 76.1% and 79.5% for the two feature sets, respectively, while the accuracy of the observers is 78.5% and 82%, respectively.

5.1 Introduction

Interstitial lung disease (ILD) is a chronic inflammation of the lung parenchyma, encompassing over 150 specific disorders causing significant morbidity and mortality [1, 80]. In recent years, computed tomography (CT) has obtained a central role in the diagnostics of ILD [78, 91]. ILD manifests itself in CT images as a variety of abnormal patterns in the lung parenchyma. Clinical estimation of disease progression is based on monitoring changes in those patterns, along with the results of physiologic tests. A change in the visual extent of disease over time is an important marker of response to therapy and a predictor of mortality. In this work, we propose an automated system for assessment of interval changes in ILD based on quantitative measurements in serial CT images.

Clinical literature conventionally agrees that an increase in the overall extent of parenchymal abnormalities is associated with disease progression, and a decrease with disease regression. Disease extent is estimated visually by a radiologist. Although highly specialized chest radiologists show moderate agreement, most CT scans with ILD are interpreted by general radiologists who might provide less reproducible and accurate results. With the introduction of multidetector CT that allows one to obtain near volumetric CT scans, the amount of image data a radiologist has to go through in order to compare two scans, has increased considerably. This makes the observational estimation of disease progression a time-consuming task. A recent study [92] showed that the time required to assess ILD changes in a scan pair was on average 123 s and 79 s for a pair of non-aligned and aligned CT scans, respectively. The same study reported a fairly low intra- and interobserver agreement (Fleiss' $\kappa = 0.54$ and 0.58 for non-aligned and aligned pairs, respectively). Therefore, the automation of ILD progression assessment is a valuable clinical application that can offer reproducible and fast estimates of disease changes.

One possible approach to automation would be the automated estimation of the overall extent of parenchymal abnormalities computed as the accumulated extent of different abnormal patterns. Recently, considerable advances have been made by the medical computer vision community in the field of texture classification in lung CT. In high resolution CT, regions of interest from 2D axial sections were automatically classified into several texture categories representative of ILD (see Refs. [88, 93], and Ref. [17] for a review). Automatic classification of 3D volumes of interest (VOI) from multidetector CT scans showed a very good reproducibility of the reference standard set by expert radiologists [94, 95]. There is a gap, however, between classification of independent regions and estimation of disease extent in the whole scan or an axial section.

Not many attempts to bridge this gap have been made so far. Ref. [88] addressed the question whether a given CT section contained a certain abnormal pattern. In [95], four complete lung volumes were manually annotated, and these annotations were compared with the output of an automated classification system that classified every small VOI in the lungs into four texture categories. The

last study showed statistically that the computer system agreed with the experts as well as the experts agreed between themselves in labeling these four subjects. To our knowledge, no work has been published that directly uses the results of automated texture classification in the estimation of change in overall disease extent.

It should be noted, that no reliable reference standard yet exists for training a texture classification system. Present systems are trained on a single expert's annotations, or annotations obtained by the consensus of a panel of experts. Making such annotations is a laborious and time-consuming task and leads to high intra- and interobserver variability [88, 17].

Reproducible quantitative CT measures have been investigated in studies directly related to the estimation of ILD progression in serial CT scans. In [96] and [97] simple statistical features were computed to characterize the distribution of intensities of the lung volume: mean lung attenuation, skewness and kurtosis in [97] and variance, contrast and entropy in [96]. Best *et al* [97] showed that all three features changed significantly in patients with deteriorated idiopathic pulmonary fibrosis (a clinical syndrome often associated with ILD). Sumikawa *et al* [96] demonstrated significant differences in measurements in thirteen cases of non-specific interstitial pneumonia (a common subtype of ILD) before and after treatment.

Our study is the first attempt to automatically assess ILD progression in a patient using quantitative measurements from a pair of CT scans. Disease progression is estimated separately in the lower, middle and upper parts of the lungs. To this end, a pair of corresponding axial sections is analyzed in each part of the baseline and follow-up scans. Alignment of two scans is an important preprocessing step that enables the retrieval of matching sections. Our system, applied to a pair of CT sections, yields an opinion whether the second image in the pair corresponds to a higher, lower, or equivalent extent of disease compared to the first image.

Two sets of quantitative features are investigated for use with an automated analysis. The first set includes statistical features which describe the intensity distribution of the difference image computed between corresponding baseline and follow-up CT sections. For the second set, we derive new dissimilarity features from local texture features that were previously shown to be able to characterize different abnormal patterns associated with ILD [75, 88]. The dissimilarities between individual texture features are used to directly estimate the difference between two images, thereby skipping classification of the lung parenchyma into different abnormal categories. In this way, we avoid the laborious and unreliable step of obtaining manual annotations for training a texture classification system.

This paper is organized as follows. A data set used for training and validation of the system is described in Section 5.2. Section 5.3 gives the system overview and details each part. Section 5.4 describes the experimental setup and observer study. The results are presented in Section 5.5 and discussed in Section 5.6.

5.2 Materials

5.2.1 Data set

Seventy-five pairs of baseline and follow-up thoracic CT scans of patients with histologically proven ILD were collected from daily clinical practices of the University Medical Center Utrecht (21 pairs) and St. Antonius Hospital Nieuwegein (54 pairs), The Netherlands, between 2003 and 2007. Types of ILD included sarcoidosis, idiopathic interstitial pneumonias, and various immune and autoimmune disorders. All patients that underwent more than one CT examination in the given period of time and had a confirmed ILD diagnosis were included in the study except for those whose scans exhibited severe motion artifacts. The time span between the baseline and follow-up scans varied from one month to two years. The data set comprised 40 male and 35 female patients, with a mean age of 53 years (range: 25–77 years) at the time of a baseline scan.

Images were obtained at full inspiration on a multi-detector row scanner (Brilliance-16P, Mx8000 IDT 16, Brilliance-40, or Brilliance-64, Philips Medical Systems, the Netherlands), with standard or low-dose parameters for high-resolution volumetric CT scanning. Collimation varied between 0.625 mm (40- and 64-slice scanners) and 0.75 mm (16-slice). Slices of 0.9 mm thickness (40- and 64-slice) or 1 mm thickness (16-slice) were reconstructed every 0.7 mm at the University Medical Center Utrecht. Slice thickness and spacing were 0.8 mm (16-slice) in the St. Antonius Hospital Nieuwegein. Exposure settings ranged between 15 and 180 mAs, with 120 or 140 kVp. All images had a per-slice resolution of 512×512, with pixel spacing in the X and Y directions varying from 0.3 mm to 0.8 mm.

5.2.2 Reference standard

Three axial sections used in the assessment of ILD progression were manually extracted from the baseline scan at the approximate distance of 2 cm above the carina (the upper part of the lungs), 2 cm below the carina (the mid part), and at 1 cm above the diaphragm (the lower part). The sequential numbers of extracted sections in the whole scan were noted. Then, three sections with the same sequential numbers were taken from the follow-up scan. Due to the previously executed image registration, the sections taken from the follow-up scan were at the same level of anatomy as the sections extracted from the baseline scan.

Corresponding pairs of CT sections were annotated by an expert chest radiologist with more than 15 years of experience. In a side-by-side comparison, the expert classified a change in disease extent in the follow-up sections. There were eight classification categories - massive decrease (disease extent reduction $>$ 50%), moderate decrease (10% to 50% reduction), minor decrease (2% to 10% reduction), stable (any change in the disease extent $\leq$ 2%), minor increase (disease extent increase 2% to 10%), moderate increase (10% to 50% increase), massive

increase (increase > 50%), or the expert could reject a pair altogether if the quality of one or both images was deemed insufficient. During the annotation process the expert had access to full CT scans, before and after registration, as well as to previous pertinent radiological reports. The expert was aware which of the two scans was baseline - its sections were always projected on the left side of the computer display.

The pairs were divided between the categories as follows: massive decrease (27 pairs), moderate decrease (17), minor decrease (11), stable (105), minor increase (24), moderate increase (20), massive increase (1 pair). Twenty pairs of CT sections were discarded by the expert because of motion artifacts, registration misalignments, or a combination of both. Motion artifacts deform the appearance of parenchyma and obscure the difference between normal and abnormal tissue which prevents both human observers and a computerized analysis from correct interpretation. The misalignment of a patient's baseline and follow-up scans makes it impossible to extract corresponding slices in the two scans, which is a critical step in our analysis. Often, motion artifacts or misalignments affected only a part of the lungs and only sections from the affected parts were discarded. Among the 20 discarded pairs, however, 3 pairs were from the same patient, which reduced the final number of participants in the study to 74. In total, there were 205 pairs of CT sections included in the the study.

The "stable" category strongly prevails over any other category in the data set, which is clinically realistic but unsuitable for training a CAD system. Such an inequality in class sizes can make classification biased towards a class that is better represented in the training set. In order to make the classification task more feasible we grouped together massive, moderate and minor categories and defined three classes, "regression", "stable" and "progression". Furthermore, we randomly selected pairs from each category to swap the baseline and follow-up images, and to change the pair label to its opposite (e.g., moderate decrease to moderate increase), until we obtained the uniform distribution of massive, moderate and minor categories in "regression" and "progression" classes. The final distribution between the classes was as follows:

"regression" - included massive, moderate and minor decrease (51 pairs)

"stable" - as previously defined (105 pairs)

"progression" - included massive, moderate and minor increase (49 pairs)

5.3 Methods

5.3.1 System overview

The proposed automated analysis generally follows the typical design of a CAD system. At first, images are preprocessed and useful discriminatory features are computed from them. This is followed by an automated analysis of patterns

described by computed features. The goal of the system is to assign a pattern to the right class. The CAD system operates in two phases. In the training phase, the system, equipped with a classification function, or a classifier, learns the parameters of the classifier from a set of patterns which true classes are known. In the testing phase, the trained classifier is applied to new, previously unseen data. The system outputs either a class label or a probability for a pattern to belong to a certain class.

The preprocessing step of our system included intra-patient registration of thoracic CT scans and subsequent 3D segmentation of the lung fields. Then, three corresponding sections from the upper, mid and lower thirds of the lungs were extracted from the baseline and follow-up scans. We computed two different sets of features from pairs of CT sections. Each feature set characterized a textural change between baseline and follow-up images. The first set of features statistically described the intensity distribution of a difference image obtained by subtracting the baseline image from the follow-up image. Features in the second set are dissimilarities computed between the baseline and follow-up images filtered with a bank of general purpose texture filters.

Two classification strategies were employed in the CAD system. In the first strategy, a classifier used the intensity distribution features to differentiate between three possible categories of change: "regression", "stable", and "progression". In the second strategy, a two-stage classification was employed. In the first stage, image pairs were classified into "changed" and "stable" categories using the dissimilarity-based features. Pairs, that were labeled as "changed", were further classified into "regression" or "progression" categories using the intensity distribution features. The CAD system, using either strategy, was evaluated by means of accuracy and McNemar's statistical test, and its performance was compared to that of two human observers.

5.3.2 Registration

Prior to the extraction of corresponding axial sections, the baseline and follow-up 3D scans were aligned using a two-step registration procedure. The scan to be deformed was selected randomly in each pair. First, the images were roughly aligned using an affine transformation. This was followed by an elastic deformation that allows for non-rigid lung tissue alignment. The elastic deformation was modeled by a B-spline grid [98]. During registration, a similarity between two images is maximized. For this purpose, mutual information was used as the cost function [99] in both steps. An iterative stochastic gradient descent optimizer [100] was applied. To avoid local minima, a multi-resolution approach was adopted. The software package Elastix[1], version 3.90 was used.

Both registration steps involved a multi-resolution strategy using a Gaussian image pyramid. For the initial affine transformation, four resolutions were used,

[1]Elastix can be downloaded from `http://www.isi.uu.l/Elastix`.

and five resolutions were used for the non-rigid deformation. A maximum of 512 optimization iterations were performed in each resolution level during the affine transformation. For the non-rigid deformation, the optimizer performed at most 512 iterations in the first four resolution levels, and 300 iterations in the last resolution level. The B-spline grid spacing used in final resolution level was eight voxels. Registration was performed on images down-sampled by a factor of two in order to reduce the computation time. The acquired transformation was then applied to the full resolution scan. The computation time required to register one image pair was 10 minutes on average on a standard high-end PC, on a single core. After registration, the two scans had the same dimensionality, with comparable anatomy at the same level of sectioning.

5.3.3 Lung segmentation

The segmentation of the lung fields in 3D CT scans was initially performed by an implementation of a conventional lung segmentation algorithm [26, 28]. This fully automatic algorithm exploits a rule-based approach to find the trachea, from which the bronchi and lungs are grown. After the trachea and main stem bronchi have been removed from the grown lung region, the left and right lungs are labeled using another set of rules, whereupon 3D hole filling and morphological closing are applied to each lung field separately.

Although generally reliable and fast, this algorithm has limitations. Relying on the assumption of contrast in attenuation between the lung parenchyma and the surrounding tissue, it tends to under-segment the lung fields in scans containing high density pathology (as often occurs with ILD). Occasionally, the algorithm was not even able to find the trachea to start segmentation with. This happens when the appearance of the trachea does not meet the assumptions made by the algorithm.

Therefore a multi-atlas segmentation-by-registration approach (MAS) was applied to scans where the conventional approach failed. An atlas is an image with a known segmentation, that is registered to a test image. Several studies have shown MAS to be a powerful segmentation tool [101, 102]. MAS assumes that N atlases are registered to an image at hand, resulting in N transformations from an atlas to the target image. These transformations are then used to propagate the atlas segmentation masks to the target image. The final segmentation of the target is obtained by the pixel-wise majority voting, i.e. a voxel is assigned to the final segmentation mask if it belongs to at least $N/2$ deformed atlas segmentations.

In the end, the lung segmentation masks of baseline and follow-up scans were merged by the "union" operation, and the resulting mask was used in the subsequent computation of features.

The lung segmentation was also performed on down-sampled scans in our study. The conventional lung segmentation method, that took on average 55 seconds per scan, failed in 28 cases out of 150. Five scans where the lungs were successfully segmented by the conventional method served as atlases for MAS.

Registration parameters for MAS were the same as for the intra-patient registration. Computation time for MAS was approximately 50 minutes per scan. It should be noted that such time-consuming operations as intra-patient registration and lung segmentation can be performed beforehand, e.g., simultaneously with an image acquisition, thus, not compromising the computation time of the CAD system.

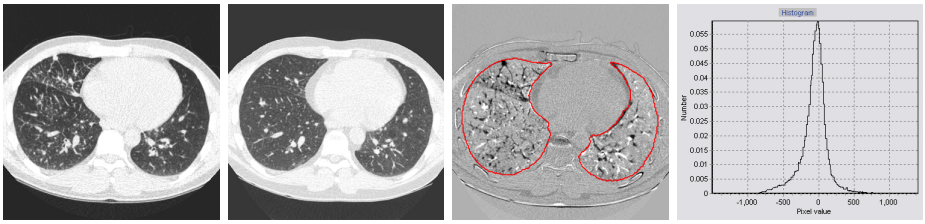
5.3.4 Features from difference image

Progression of ILD is associated with the extension of abnormal patterns in a scan. Abnormal patterns typically increase the opacity of lung parenchyma and therefore lead to higher density values than normal parenchyma. This motivated us to extract discriminatory features to describe interval changes in ILD from the difference image between the aligned follow-up and baseline scans. With no change in the disease state between the two images, this difference image should not exhibit much intensity variation in the lung fields. Ideally, one would expect the intensity histogram of the lung fields of the difference image to have a large symmetrical peak around zero and a small standard deviation. If the disease state has changed over time, we would expect a biased intensity histogram - towards positive numbers in case of disease progression, and towards negative numbers in case of regression. This is illustrated in Figure 5.1. In this figure, the difference images corresponding to regression (Figure 5.1(a)) or progression (Figure 5.1(c)) show more darker or brighter regions, respectively, than the difference image of a stable case (Figure 5.1(b)).

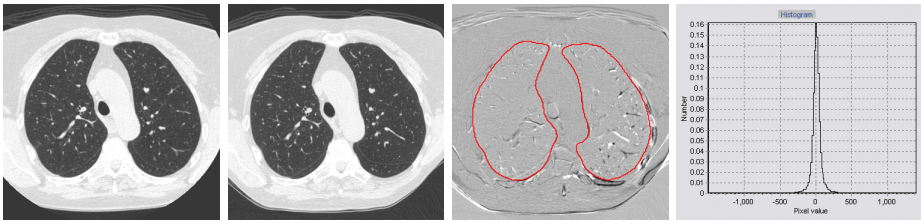
The difference image of the lung fields can be described statistically in a number of ways. We used four statistical features: three quartiles and the mean. The three quartiles were the 25th percentile, the median and the 75th percentile. Prior to computing the features, the difference image was filtered with the Gaussian filter at a scale of 2, in order to decrease spurious registration discrepancies between the two images. The areas outside the lung fields were masked out and were ignored in the computation of the features.

5.3.5 Dissimilarity-based features

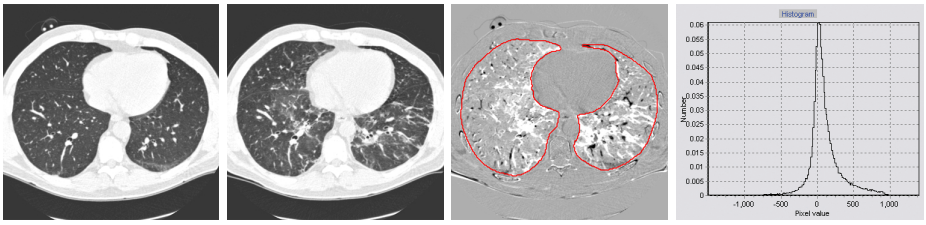
Features that only consider overall parenchyma density changes simplify what happens during development of ILD. Not only the amount but also the type of abnormal patterns can change. For example, a typical sign of deterioration is the substitution of ground glass opacities by fibrotic tissue [7]. Such changes in patterns do not always result in an increase in parenchymal opacity. On the other hand, there may always be density variation in the difference image with causes unrelated to the progression of ILD, for example different levels of inspiration, imperfect registration, different signal-to-noise ratios in the baseline and follow-up scans, etc. As a result, image pairs with subtle changes in abnormal patterns are likely to be confused with “stable” image pairs with common disparities if



(a) *Regression*



(b) *Stable*



(c) *Progression*

Figure 5.1: *Examples of difference images and their intensity histograms. In each row, from left to right: a pair of two corresponding registered CT sections of the same patient taken at different moments of time; difference image obtained by subtracting the first CT section from the second one; normalized histogram of the difference image. The histogram was computed from the lung fields only, which are outlined in the difference image. Prior to computing the histogram, the difference image was smoothed with a Gaussian ($\sigma = 2$). Here the original difference images are shown.*

features extracted from them only characterize the distribution of intensities in the difference image. Figure 5.2 shows three cases with a change in the extent of ILD where the histograms of their difference images are very similar to the typical histogram of a "stable" case. To better illustrate parenchymal changes related to ILD, corresponding texture patches taken from the baseline and follow-up images in Figures 5.2(a) and 5.2(b) are enlarged and presented in Figure 5.3.

Following these considerations, we propose to extract and compare the texture contents of the images in order to describe interval changes in ILD. We start with a number of texture features computed locally throughout the lung fields. Then, the distribution of each feature is described by the histogram, separately in the baseline and follow-up images. Finally, a measure of dissimilarity between the two histograms of each texture feature is computed. In this way, a pair of images is characterized by a vector of dissimilarities in texture. In the next subsection we describe the utilized texture features, and, subsequently, we detail how the dissimilarities were computed.

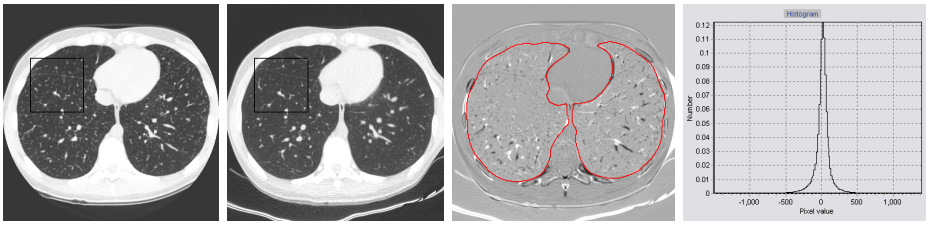
Local texture features

We used a set of general purpose texture features that have been previously applied to the classification of abnormal texture patterns in high resolution thoracic CT [75, 88]. These features were four central statistical moments in eight filtered versions of the original image, calculated on three scales. The eight filters were the Gaussian, the Laplacian, the 1st order derivative of the Gaussian in three orientations between 0 and π, and the 2nd order derivative of the Gaussian in the same orientations. The scales σ were 0.5, 1 and 2 pixels. Prior to filtering the image, pixel values in the lung fields were mirrored outside the lungs symmetrically with respect to the lung borders. This prevented a major distortion in the filter output near the lung borders which is normally caused by a large difference in appearance inside and outside the lungs. The first four moments, i.e., the mean, standard deviation, skewness, and kurtosis, were calculated from multiple regions of interest (ROIs), placed in the lung fields, in order to capture the local texture information. We used an 8 by 8 pixel spacing to define the centers of circular ROIs, each of which had a radius of 16 pixels. On average, there were 870 ROIs per image. In total, 96 features (8 filters $\times$ 3 scales $\times$ 4 moments) were computed for each ROI.

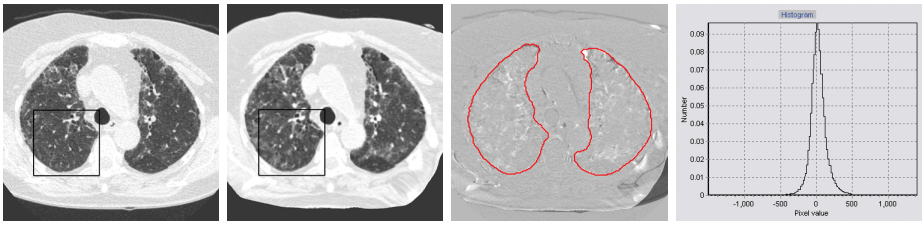
Computation of dissimilarities

Each texture feature distribution was represented by a 64-bin normalized histogram, which resulted in 13-14 entries per bin, on average. The bin partitioning was determined on a set of training images by computing the range of values for each feature and splitting this into 64 equal intervals.

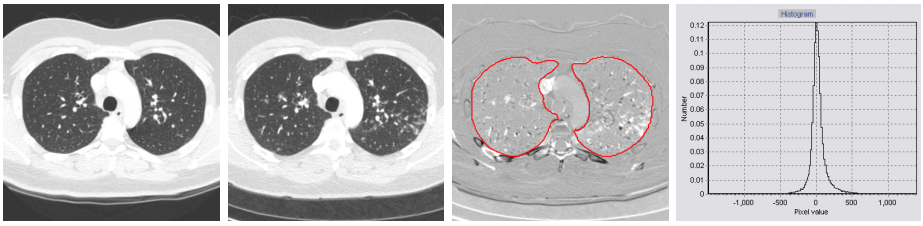
A number of comparison measures between two distributions have been proposed by the image retrieval community (see, for example, Ref. [69] for a review).



(a) *Regression*

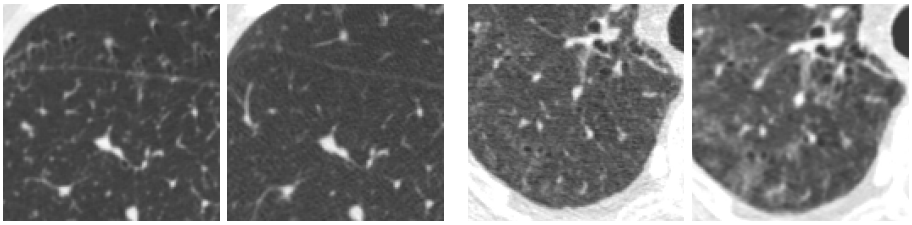


(b) *Progression*



(c) *Progression*

Figure 5.2: *Image pairs with changes in ILD extent whose difference image intensity histograms are similar to that of a typical "stable" case. The corresponding square patches of texture marked in (a) and (b), are enlarged and presented for comparison in Figures 5.3(a) and 5.3(b) respectively.*



(a) *Enlarged regions from Fig. 5.2(a)* **(b)** *Enlarged regions from Fig. 5.2(b)*

Figure 5.3: *Corresponding patches of texture in the baseline and follow-up CT sections. In (a), textural changes associated with ILD regression are shown (left to right), in (b) - textural changes associated with ILD progression. These are examples of subtle changes.*

They are conventionally termed dissimilarity measures, and we adhere to this term in this paper. Let us denote the histograms of feature f, computed from images A and B, as $h_f^A = \{h_f^A(i)\}$ and $h_f^B = \{h_f^B(i)\}$ respectively, i being a bin index. A dissimilarity measure between h_f^A and h_f^B is denoted $d(h_f^A, h_f^B)$. Then, a comparison measure between A and B is obtained as a vector of dissimilarities $D(x, y) = \{d(h_f^A, h_f^B)\}$, where f is a running index through all available features.

For this study, we experimented with the following distance measures: Minkowski distances of order 1 and 2, χ^2 statistics, and Jeffrey divergence. The best results were obtained with the Minkowski distance of order 1, also known as the city block distance. The city block distance is defined as $d(h_f^A, h_f^B) = \sum_i \left|h_f^A(i) - h_f^B(i)\right|$. The dissimilarity vector D of the same dimensionality as the number of local texture features, was computed between the baseline and follow-up images, and this was used as a feature vector entering the second classification strategy, detailed below.

5.3.6 Classification

Strategy I

Our first classification approach exploited the intensity distribution features computed from difference images. A classifier was trained to distinguish between three classes of temporal change: "regression", "stable", and "progression". For convenience, we call a pair of the baseline and follow-up images a "training sample" when it is used by the automated system to train the classifier and its correct class is known to the system. A "test sample" is a pair that is new to the system, and the system attempts to define its class using a classification rule it has learnt. In our system the linear discriminant analysis (LDA) was employed, as the most accurate classifier after preliminary comparison with the k nearest neighbor classifier (k-NN) and support vector machines with linear and exponential kernels. The LDA assumes Gaussian distributions for the samples of each class, with equal

covariance matrices for each distribution.

Strategy II

As noted in Section 5.3.5, the dissimilarity-based features are intended to reflect how "far" from each other two images are, without specifying the "direction" of change. These features only enable the distinction between "changed" and "stable" pairs of images. To be able to perform a 3-class classification with them, we developed the following strategy. First, the dissimilarity-based features were used with a 2-class classifier that distinguishes between "changed" and "stable" pairs of images. Next, we applied another 2-class classifier to those pairs that were labeled as "changed" by the first classifier. The second classifier was trained to discriminate between "regression" and "progression" cases. The same features computed from difference images as used in strategy I were used with the second classifier of Strategy II.

We assume that the introduction of the dissimilarity-based features will improve the separation of "changed" and "stable" pairs, based on the considerations introduced in Section 5.3.5. The intensity distribution features, insufficient on their own for three-class classification, play a role in the combined system because they are deemed distinctive for "regression" and "progression" cases. Regression and progression difference images are opposite to each other in terms of intensity, therefore we expect the intensity distribution features to be able to distinguish between them without the need for more sophisticated features.

The k-NN classifier was employed in first stage of classification, and the LDA in the second stage. The k-NN classifier is a non-parametric classifier. According to the k-NN rule, the test sample is assigned the majority label of the nearest k training samples. The free parameter k has to be chosen experimentally ($k = 15$ in our system). In this work, the fast implementation of the k-NN classifier by Arya and Mount [89] was used.

We found that the classification accuracy in the first stage benefited from applying principal component analysis (PCA) that reduced the number of features entering the classifier. PCA retained 99% of variance in the feature vector. In both classification strategies, feature vectors were normalized to zero mean and unit variance prior to classification. Normalization parameters were estimated in the training data and used on the feature vectors of the test data.

5.4 Experiments

5.4.1 Experimental setup

Classification experiments in this study were performed using the leave-one-out cross validation procedure. Cross validation involves training a classifier n times, each time leaving out one of the n disjoint data subsets from training, and using

only the omitted subset for validation. For leave-one-out cross-validation, n equals the sample size. This technique guarantees the optimal use of the available data.

We divided the data into training and test sets on the basis of scans, not sections. Therefore, 1 to 3 image pairs of the same patient were set aside in each leave-one-out iteration. In this way, the evaluation of the system classification performance was unbiased, because at no time did training and test sets contain samples originating from the same scan pairs. For each strategy I and II the system performed 74 classification. The final accuracy was computed from the outcomes of all classifications.

5.4.2 Observer study

In order to compare the performances of the two classification strategies to that of radiologists, an observer study was conducted. Observers were presented the same set of 205 registered pairs. Prior to that, pairs were randomly shuffled, so that pairs of sections from the the same scans (1 to 3 pairs) did not necessarily follow each other. Additionally, before presenting each pair for a side-by-side comparison, the side of the display on which the baseline scan was projected was randomly chosen. The observer was asked to classify the extent of disease in the image on the right-hand side compared to the image on the left-hand side. There were 3 classification categories - decrease (disease extent reduction $> 2\%$), stable (any change in the disease extent $\leq 2\%$), increase (disease extent increase $> 2\%$). Both observers were chest radiologists in training. They were not involved in setting the reference standard for the data in this study.

5.5 Results

The classification performances of the system and the observers were estimated by means of accuracy and compared to each other using McNemar's statistical test. In the explanation of the results we use the term "rater" to indicate either a computer system or a human observer.

Classification accuracy was calculated as the fraction of correctly classified samples in the test data set. In Table 5.1 accuracies are shown for each system and observer. The corresponding contingency matrices are presented in Table 5.2. In each matrix, rows represent the reference standard and columns represent a rater's opinion. Entries on a matrix diagonal show numbers of pairs correctly classified in each class.

McNemar's test [103] measures whether the total number of misclassified samples of one rater is significantly different from the number of misclassifications of another rater. We computed this test for each pair of raters. The p-values of McNemar's test are given in Table 5.3. According to these results, none of the raters performed significantly different from the others.

Table 5.1: *The performances of two classification strategies and two observers, measured on the same data set by accuracy. The system performances were obtained using leave-one-out cross validation.*

	Strategy I	Strategy II	Observer I	Observer II
Accuracy	0.761	0.795	0.785	0.820

Table 5.2: *Confusion matrices of the two classification strategies and the two observers. Each row represents the reference standard. Each column represents a class obtained by the system or observer. Class names are abbreviated:* R *for "regression",* S *for "stable", and* P *for "progression".*

	Strategy I		
True class	*R*	*S*	*P*
R	27	24	0
S	1	101	3
P	2	19	28

(a)

	Strategy II		
True class	*R*	*S*	*P*
R	33	18	0
S	5	97	3
P	2	14	33

(b)

	Observer I		
True class	*R*	*S*	*P*
R	47	3	1
S	15	66	24
P	1	0	48

(c)

	Observer II		
True class	*R*	*S*	*P*
R	38	13	0
S	8	92	5
P	1	10	38

(d)

The contingency matrices in Table 5.2 show different tendencies in misclassification for different raters. For example, observer I seems more sensitive to small differences in the disease extent than the other raters. Observer I has the fewest misclassifications in "regression" and "progression" categories, but, at the same time, a lot of misclassifications of "stable" cases for cases with change. The tendency of the other three raters is the opposite - they were inclined to erroneously classify pairs with change as "stable".

5.6 Discussion

The results of this study demonstrate that the performance of both computerized classification strategies was not significantly different from that of two human observers in the assessment of ILD progression. Here we discuss the results in more detail.

Table 5.3: p-*values of McNemar's test on the equality of error rates.*

	Strategy I	Strategy II	Observer I
Strategy II	0.27		
Observer I	0.65	0.9	
Observer II	0.09	0.54	0.43

Neither the classification strategies nor the observers had much difficulty in correctly classifying obvious changes. As mentioned in Section 5.2.2, the reference standard was initially given on a 7-point scale, with two extreme points indicating "massive decrease" and "massive increase" respectively. All 28 pairs in these categories were correctly classified as either "regression" or "progression" by both observers. Strategy I correctly classified 25 pairs, making mistakes in three pairs belonging to the same patient. Strategy II made the same errors as strategy I and additionally misclassified one pair. An example of a correctly classified case with massive disease progression is given in Figure 5.1(c). The three pairs misclassified by both strategies exhibited a type of abnormality that was unique in the collected data set. This means that the classification algorithms did not have an opportunity to train on similar patterns. One of these three pairs is shown in Figure 5.2(a). The histogram image in this figure suggests that this pair might have been misclassified by strategy I even in the presence of similar patterns in the training data, because the difference image intensity histogram does not reflect such a subtle change.

Among 37 pairs with "moderate" changes, which meant 10 to 50% decrease or increase in disease extent, both strategy II and observer II made 6 errors, while observer I made no errors, and strategy I made 16 errors. All errors pertained to mislabeling a pair with change as a stable case. Only three image pairs with "moderate" change, belonging to two different patients, were misclassified by both strategies but correctly classified by both observers. Two of these three pairs are shown in Figures 5.4(a) and 5.4(b). An example of a case with moderate disease regression correctly classified by all raters is given in Figure 5.1(a). Some cases were correctly classified by one of the classification strategies but misclassified by observer II: two examples are shown in Figures 5.4(c) and 5.4(d).

Most misclassifications occurred for cases from the "minor decrease" and "minor increase" categories (2 to 10% change), both for the observers and the classification algorithms. Among 35 pairs with "minor" changes, half were misclassified by observer II, and more than half by both classification strategies. However, observer I made only 5 errors with these pairs. It is possible that the threshold of 2% used in our protocol lies within statistical uncertainty and cannot be used reliably. In the literature, thresholds of 10% [104] and 5% [97] have been used to differentiate between "changed" and "stable" disease.

Very few errors were made by confusing "regression" and "progression" cases: both computer strategies and observer I made two errors, while observer II made

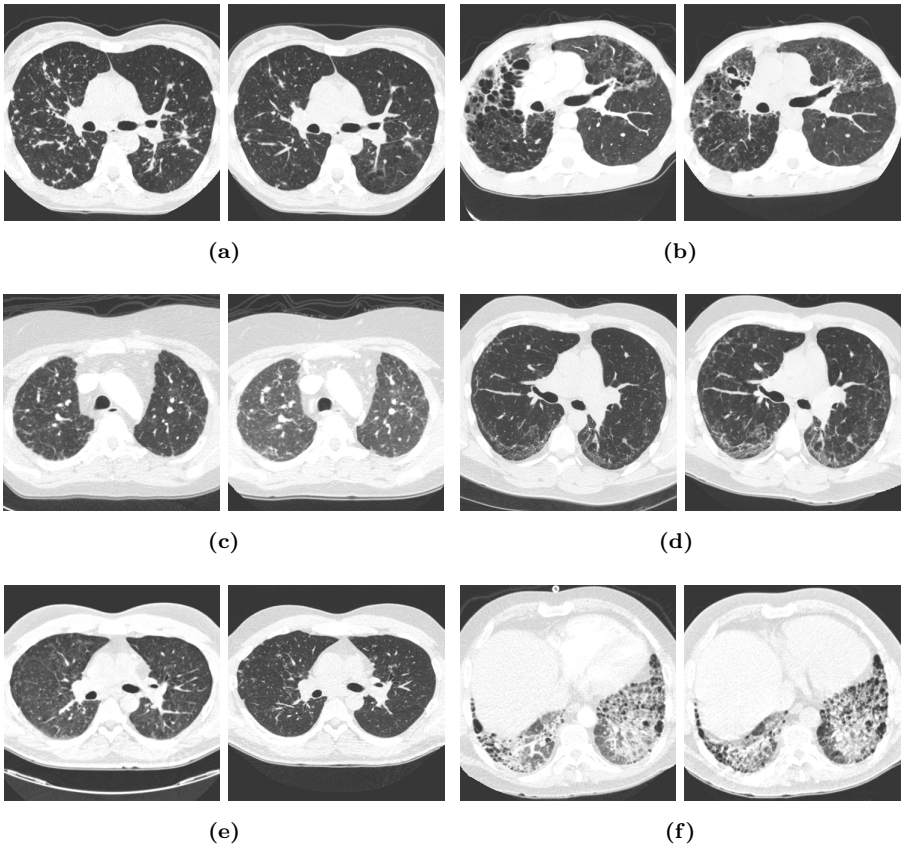


Figure 5.4: *Examples of ILD progression classification in follow-up images. Each example consists of a baseline CT section (left image) and a corresponding follow-up CT section (right image). Pairs (a), (b): each pair exhibits moderate disease regression (between 10% and 50%), and was correctly classified by both human observers but misclassified by both classification strategies. Pairs (c), (d): each pair exhibits moderate disease progression (between 10% and 50%), and was correctly classified by one or both classification strategies, but misclassified by one of the human observers. Pairs (e), (f): each pair exhibits minor disease progression (between 2% and 10%), but was misclassified for disease regression by both classification strategies and one or both radiologists.*

one error. Only three image pairs were misclassified in this way by all raters jointly, two of them being from the same patient. The misclassified pairs always exhibited a minor change in disease extent. Two of the three pairs, taken from different patients, are shown in Figures 5.4(e) and 5.4(f). They clearly represent difficult cases.

The most evident difference between the two strategies is the improvement of the recognition of change by strategy II compared to strategy I, as seen from Tables 5.2(a) and 5.2(b). Figures 5.2(b), 5.2(c), and 5.4(d) demonstrate image pairs with ILD progression that were classified correctly by strategy II but misclassified as "stable" by strategy I. Given the fairly low number of test samples, we could not show, however, that the difference in accuracies of the two strategies was significant. The observed improvement could be attributed to a more appropriate set of features for discrimination between "changed" and "stable" image pairs employed in the first stage of strategy II. Additionally, it should be noted that the training set in the first stage of strategy II was balanced, with approximately the same number of samples in both "changed" and "stable" classes. At the same time, the training set in strategy I contained twice as many samples of the "stable" class as either of the other two classes. This might have negatively influenced the performance of strategy I.

Although the current results obtained by our automated analysis are promising, several possible improvements in the system setup can be identified.

We used a single expert's annotations as the reference standard. As we have already mentioned in Section 5.1, manual annotations are not considered reliable for training a texture classification analysis because of low rating agreement. Our task, however, required the estimation of overall disease extent, which should cause less intra- and interobserver variability between experienced chest radiologists. As opposed to labeling small regions of interest into multiple categories, assessing disease extent is part of the daily clinical routine for radiologists. In clinical studies on ILD prognosis and progression, a single expert's estimate is commonly referred to as the ground truth. For example, in [97], the visual assessment of the disease extent was used as one of three independent criteria to define a patient's disease progression, along with physiologic tests, such as total lung capacity and the resting oxygen saturation level. Combining an expert's estimates, or estimates obtained by consensus, with the results of physiologic tests may improve the reliability and consistency of our reference standard. Another improvement of the training process would be the enlargement of the training data set, so that different abnormal patterns are sufficiently represented.

There is an inherent limitation of the system performance related to the generalizing ability of dissimilarity-based features. Dissimilarities computed over the whole lung fields are likely to neglect some small local changes that would probably be revealed in dissimilarities over smaller areas. A recent study [95] showed a good performance in classification of small VOIs by means of dissimilarities. In that study, a more sophisticated dissimilarity measure than in our study and a different way of histogram binning were used. Experimenting with measures

of dissimilarity between histograms, other than those mentioned in Section 5.3.5 could be beneficial for our system as well. Application of our system to regions smaller than a CT section is also possible provided the ground truth for smaller areas is available.

It is computationally efficient to train and apply automated analysis to two-dimensional CT sections. Modern clinical practice requires, however, the analysis of 3D volumes. In our system, image preprocessing (intra-patient registration and lung segmentation) is already done for full CT scans. The computation of features and classification strategies presented in this paper could be easily extended to 3D. An alternative approach would be to apply our system section-by-section to the 3D scan and subsequently fuse the classification outcomes of individual sections into a decision about the whole lung volume or its part. This approach might be preferable to the direct 3D analysis, because the generalizing quality of dissimilarity features is likely to be enhanced in 3D which will make the system less sensitive to small changes.

5.7 Conclusions

We have developed a classification system that performs estimation of ILD progression in axial sections extracted from serial thoracic CT scans. To achieve this, our system comprises non-rigid intra-patient image registration, multi-atlas lung segmentation, texture feature extraction, and computation and classification of dissimilarities. The accuracy of our system is not significantly different from that of two radiologists.

Chapter 6

Application of dissimilarity-based classification to the automatic detection of chest radiographs suspicious of tuberculosis

Y. Arzhaeva, L. Hogeweg, P. A. de Jong, M. A. Viergever and B. van Ginneken, "Global and local multi-valued dissimilarity-based classification: application to computer-aided detection of tuberculosis", *the International Conference on Medical Image Computing and Computer Assisted Intervention*, 2009, in press.

Abstract

In many applications of computer-aided detection (CAD) it is not possible to precisely localize lesions or affected areas in images that are known to be abnormal. In this chapter a novel approach to computer-aided detection is presented that can deal effectively with such weakly labeled data. Our approach is based on multi-valued dissimilarity measures that retain more information about underlying local image features than single-valued dissimilarities. We show how this approach can be extended by applying it locally as well as globally, and by merging the local and global classification results into an overall opinion about the image to be classified. The framework is applied to the detection of tuberculosis (TB) in chest radiographs. This is the first study to apply a CAD system to a large database of digital chest radiographs obtained from a TB screening program, including normal cases, suspect cases and cases with proven TB. The global dissimilarity approach achieved an area under the ROC curve of 0.81. The combination of local and global classifications increased this value to 0.83.

6.1 Introduction

Pulmonary tuberculosis (TB) is a major cause of death and illness worldwide, with 9.27 million new cases and 1.75 million deaths reported in 2007 [4]. Chest radiography is increasingly important in the fight against TB, especially because the rates of sputum-negative TB are rapidly increasing in populations with a high incidence of HIV/AIDS. On chest radiographs, TB often presents itself through subtle diffuse textural abnormalities. With the advent of digital radiography, computer-aided detection (CAD) systems can be developed that could facilitate mass population TB screening.

However, little research has been done in this area. In [66] texture analysis within the lung fields was used but this required experts to manually delineate abnormal areas, in order to train the system to discern normal regions from abnormal. Although such an approach may lead to a powerful CAD system, obtaining manual delineations of ill-defined diffuse lesions is laborious and likely to produce an unreliable ground truth. This chapter focuses on classification of weakly labeled images, i.e. when the exact locations of abnormalities in training data are unknown and, therefore, local feature-based classifiers cannot be trained. It is based on research described in chapter 3, where the multi-valued dissimilarity-based (MVDB) classification was introduced. We circumvent the absence of local ground truth by using the distances, or dissimilarities, between estimated distributions of local features in the global classification of images. These dissimilarities are estimated per feature and, therefore, a multi-valued dissimilarity-based (MVDB) classification system is built.

The underlying assumption of the MVDB method is that local feature distributions are sufficiently different for normal and abnormal images. However, this assumption is not likely to hold for cases with subtle small abnormalities only. We hypothesize that subdividing the lung fields into smaller parts and applying the MVDB classification to these parts separately, and subsequently combining these local opinions, may improve the sensitivity of the method to such abnormalities and increase overall classification performance. It should be noted that obtaining ground truth labels for fixed large lung subdivisions is easier than obtaining manual delineations of lesions. In this work, we apply the method to a large database of digital radiographs from a TB screening program. The proposed modification of the MVDB classification is general and applicable to other image classification tasks that involve local analysis.

6.2 Methods

6.2.1 Multi-valued dissimilarity-based classification

Dissimilarity-based classification uses dissimilarity representations of objects instead of traditional feature vectors, that is, objects are represented by their pairwise comparisons. This is a natural way to describe a class of similar objects. A

pairwise comparison is done by computing a measure of dissimilarity, or distance, between two objects. In the standard dissimilarity-based classification [60], each training object is represented as a vector of distances to a set of prototype objects. Then, any traditional classifier can be trained on dissimilarity representations of training objects and applied to the dissimilarity representation of a new object. This may not be an efficient strategy for classifying objects characterized by a large set of descriptors, such as numerous local texture features, because it reduces the abundance of local information in two objects to just one dissimilarity value between them.

MVDB classification is built on similar principles but reduces the loss of information compared to standard dissimilarity-based classification. While the standard dissimilarity-based method accumulates the distance over all the object descriptors, the MVBD method is based on computing a distance for every descriptor individually.

Let x and y be two objects characterized by n one- or multi-dimensional descriptors f_i, and $d_i = d(f_i^x, f_i^y)$ be the value of dissimilarity between corresponding descriptors of x and y, where d is a dissimilarity measure. Then, a vector $D(x, y) = (d_1, \ldots, d_n)$ is called the dissimilarity representation of object x with respect to object y. To construct a classifier on such representations, let us consider a training set T, and a set of prototype objects R of size r, $R = \{p_1, \ldots, p_r\}$, where $R \subseteq T$. For each $x \in T$, r different representations $D(x, p_k)$, $1 \leq k \leq r$, can be obtained, and consequently r classifiers can be trained on T using $D(x, p_k)$ as input. A test object, subsequently, can be classified r times using its prototype-bound representations. To obtain a final classification solution for a test object, the outputs of r classifiers must be combined. Combining classifiers benefits from complementary information provided by different dissimilarity representations. In this study we combine the posterior probabilities resulting from different classifiers with the sum rule:

$$P(c|x) = \frac{1}{r} \sum_{k=1}^{r} P_k(c|x)\,, \tag{6.1}$$

where $P(c|x)$ is a posterior probability that the object x belongs to a class c, and $P_k(c|x)$ is a posterior probability yielded by the classifier k.

6.2.2 Application to image classification

Now we show how the MVDB classification can be applied to images by listing the steps for the training and testing phase. A flow chart of the same algorithm can be found in Figure 3.1 of chapter 3. To apply this method to an image classification task that involve local texture analysis, we describe each image by the distributions of its local texture features. These features are extracted at numerous locations inside the image, and their individual distributions are estimated by histograms. From the set of training images, r prototype images are selected, either randomly, or by following a systematic approach. Since the sum rule, used

to combine the results of individual classifiers, is known to be less sensitive than other combiners to the errors of individual classifiers [70], we believe, the random selection of prototypes is a reasonable starting approach. Dissimilarities, computed between corresponding feature histograms of the image and a prototype, constitute a dissimilarity representation of the image.

Training phase:

1. *Selection of r prototype images*
   Input: training images, selection algorithm
   Output: subset of prototypes

2. *Extraction of N local texture features at M locations*
   Input: training images
   Output: M feature vectors of length N for each image

3. *Computation of N feature histograms*
   Input: M values of each feature across an image
   Output: N histograms

4. *Computation of dissimilarities to each prototype*
   Input: N histograms of a training image, N histograms of a prototype, dissimilarity measure
   Output: r N-dimensional dissimilarity representations for each training image

5. *Training r classifiers*
   Input: r N-dimensional dissimilarity representations, classifiers
   Output: r trained classifiers

Testing phase:

1–3. *Same as steps 2–4 of the training phase*
   Input: a test image
   Output: r N-dimensional dissimilarity representations

4. *Classification*
   Input: r N-dimensional dissimilarity representations, r trained classifiers
   Output: r posterior probabilities of the test image of being abnormal

5. *Combining classification results*
   Input: r posterior probabilities, combining rule
   Output: final classification result

6.2.3 Local classification to improve global results

Sometimes image descriptors, such as the local feature histograms, are too generalizing. This is true when an object whose presence we want to detect is too small with respect to the whole image. With the detection of TB, the texture feature histograms computed over the whole lung fields might not be sensitive enough to reflect the presence of subtle localized lesions in the lungs. We assume that the discriminating ability of descriptors will increase if they are computed over smaller image parts. When it is practical to obtain the ground truth for training images on a finer scale, e.g. class labels for a fixed image partition, we propose the following modification of the MVDB classification scheme.

1. Images are partitioned, and the ground truth is obtained for each part.
2. The MVDB scheme is applied to each image part separately, and, optionally, to the whole images too.
3. The classification results are combined to obtain an overall image solution.

Here, the combination rule might be different from the one in step 5 of the testing phase of the original scheme. We use the vote rule for the abnormal class ($c = 1$), and compute the posterior probability of the normal class($c = 0$) such as $P(c = 0|x) = 1 - P(c = 1|x)$, where x is the test object. The vote rule for computing $P(c = 1|x)$ is

$$P(c|x) = \max(P_0(c|x), \max_{l=1}^{L} P_l(c|x_l)), \ c = 1\,, \tag{6.2}$$

where x_l, $1 \leq l \leq L$, are L image subdivisions, P_l is the result of applying the MVDB classification to x_l, and P_0 is the result of applying the MVDB classification to the whole image. The choice for the vote rule for the detection of abnormal images is intuitive because if any part of the image is abnormal then the whole image is abnormal. The use of P_0 in Eq. 6.2 is optional and is not needed if the performances of all regional classifiers are considerably better than that of the global classifier. It should also be noted that, for a certain region, only a fraction of abnormal images will have abnormalities in that particular region. Therefore, a high classification performance on one of the regions is not enough to obtain an equally high performance after combining. In addition to improving the image classification performance, the application of the MVDB to the regions allows one to obtain a prediction on which regions are likely to contain abnormalities.

6.3 Experiments

6.3.1 Materials

All images used in this work were posterior-anterior chest radiographs collected from a TB screening program among a high risk population. Radiographs were acquired with mobile digital thorax units (Delft Imaging Systems, the Netherlands)

developed for cost-effective thorax examination and TB preventive screening. Images have a resolution of 2048×2048 and 12 bits data depth. Each image was read by two radiologists, and a person whose radiograph was considered TB suspect by one of them or both was contacted to undergo further tests. For a subset of the cases, positive microbiological culture tests were available and a definite diagnosis of TB could be established.

We collected all TB suspect and TB proven cases between 2002 and 2005, and a similar amount of randomly selected normal radiographs, excluding radiographs of children. Before collection, radiographs were anonymized. Normal and TB suspect images were re-read by a third radiologist, who classified a part of the cases differently. Re-classified images were excluded from the study. Finally, our database contained 256 normal radiographs (223 males, 33 females, ages 18–70 yrs, median age 41), 178 TB suspect radiographs (155 males, 23 females, ages 16–101 yrs, median age 35), and 37 radiographs with microbiologically proven TB (30 males, 7 females, ages 16–43 yrs, median age 29).

6.3.2 Local feature extraction

For practical considerations, images were downsized to 1024×1024. Prior to feature extraction, lung fields were automatically segmented from the radiographs using multi-resolution pixel classification, with settings as given in [25]. In order to train this segmentation procedure, lung fields were segmented manually from 20 radiographs not used otherwise in this study.

Next, local texture features were extracted from a large number of regions of interest (ROIs). At first, images were filtered with a multiscale filter bank of Gaussian derivatives, and subsequently central moments of histograms were calculated from each ROI in the original and the filtered images. The following parameters were chosen: Gaussian derivatives of orders $0, 1$ and 2 at five scales, $\sigma = 1, 2, 4, 8, 16$ pixels; overlapping circular ROIs with a radius of 32 pixels placed on a grid with 8×8 pixel spacing inside the lung fields; and four central moments, namely, the mean, standard deviation, skewness and kurtosis. Before filtering, pixel values in the lung fields were mirrored outside the lungs symmetrically with respect to the lung borders in order to prevent contamination of extracted features due to strong filter responses at the lung border. Two position features were added that defined x and y coordinates of the ROI centers relative to the center of the mass of a lung field. In total, 126 features were extracted from each ROI, and the number of ROIs per image ranged from 1920 to 8680.

6.3.3 Lung partitioning

In order to perform the MVDB classification on lung subdivisions, each lung field was automatically divided into 4 equal-sized regions, as we did for the study in chapter 4 (see Figure 4.3). The regions around hilum (regions 4 and 8) included lung pixels overlapping with a circle placed at the lungs' center of mass. The

radius of the circle was separately chosen for the left and right lung, such that the overlap covered one quarter of the pixels of that lung. The rest of each lung field was horizontally divided into three equal-sized parts. The third radiologist assessed regions in all the TB suspect and TB proven images, and assigned a region to class 1 if a TB-related abnormality was present in the region, or to class 0 otherwise.

6.3.4 Classification

Training and test images were randomly selected from the available normal and TB suspect data, so that the training and test sets each contained 128 normal and 89 abnormal images. The second test set was formed from the same normal images as in the first test set and all 37 radiographs with proven TB. We randomly selected 10 normal and 10 abnormal radiographs to serve as prototype images. For region classification, the same division into training and test sets was used, but the random selection of prototypes was performed separately for each region, so that 10 normal and 10 abnormal regions were selected each time. Prototypes were always selected from the training images. Normal prototype regions were selected from normal training images only. Normal regions from abnormal images were excluded from the training set during region classifications.

The histograms of each local feature were obtained by a suitable binning of the range of feature values, either across the lung fields, or across a particular region. The range of possible values of each feature was estimated on prototypes and split into equal intervals - 128 for the lung fields, 64 for regions. A dissimilarity between two histograms was computed using χ^2 statistics as a dissimilarity measure

$$d_{\chi^2}(h,k) = \sum_i \frac{\Big(h(i) - m(i)\Big)^2}{m(i)}\,, \tag{6.3}$$

where $h = \{h(i)\}$ and $k = \{k(i)\}$ are two corresponding histograms, i is a bin index, and $m(i) = \frac{h(i)+k(i)}{2}$. Dissimilarity representations were classified by the linear discriminant classifier. Classification was preceded by a principal component analysis (PCA) retaining 99% of variance to the dissimilarity representation, for the purpose of dimensionality reduction.

The MVDB method was compared with a straightforward approach where an image classification was composed of classification of each ROI and subsequent fusion of ROIs' posterior probabilities. In this approach, local feature vectors extracted from ROIs as described in Section 6.3.2 were used as input of the linear discriminant classifier preceded by the PCA. The division into training and test images was the same as for the MVDB experiments. ROIs from training images got the class labels of lung subdivisions they belonged to. An overall image decision was obtained by integrating all ROIs' posterior probabilities using the 95% percentile rule.

Table 6.1: *The performances of the MVDB classification and fusion, in terms of A_z.*

		Regions								
Test set	**Lungs**	1	2	3	4	5	6	7	8	**Vote**
`suspect`	**0.81**	0.79	0.71	0.85	0.66	0.82	0.81	0.72	0.77	**0.83**
`proven`	**0.70**	0.85	0.71	0.95	0.64	0.66	0.43	0.49	0.65	**0.74**

6.4 Results

The classification performance was estimated by means of the area under the receiver operating characteristic (ROC) curve, A_z [29]. A_z values for the two test sets are presented in Table 6.1. The test set `suspect` contained normal and suspect TB images. The test set `proven` contained images with proven TB and the same normal images as `suspect`. The first column of Table 6.1 shows the results of the application of the MVDB classification to the whole lung fields. In the columns titled "1" to "8", A_z values for corresponding lung regions are listed (see Figure 4.3). The final classification performance computed after combining global and local posterior probabilities by voting is given in the last column.

Combining global and local classification decisions slightly improves the overall classification performance compared to the results after applying the MVDB method to the whole lung fields only. To illustrate the gain of using the combination of local and global classifications, an example of a region with a slight diffuse abnormality is shown in Figure 6.1. This region was correctly classified as abnormal by the MVDB method applied locally (posterior probability $p_{c=1} = 0.89$), while an image containing this region was initially misclassified as normal by a global MVDB classifier ($p_{c=1} = 0.24$). After combing global and local results, the image received a probability of 0.89 of being abnormal.

The straightforward classification approach achieved $A_z = 0.77$ on the first test set and $A_z = 0.64$ on the second test set. This demonstrates the advantage of the MVDB method for classification of weakly labeled images.

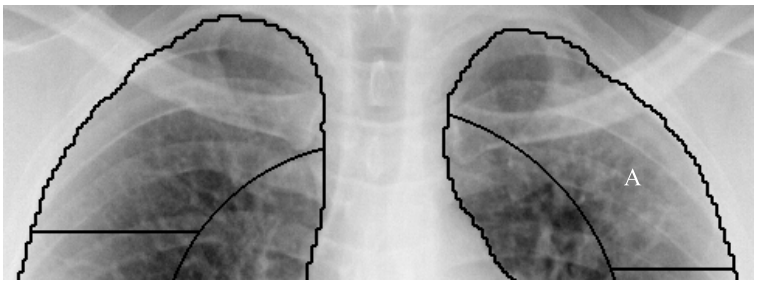


Figure 6.1: *An example of a correctly classified abnormal region (white "A" marks a proven TB lesion). An opposite region in the other lung is normal.*

6.5 Discussion and conclusions

The results presented in Table 6.1 demonstrate that the classification performance on the whole lung fields, as well as the performance of the combining scheme is considerably better on the first test set than on the second one. This observation can be explained in two ways. Firstly, the training set did not contain proven TB cases and so the test set with proven TB cases is expected to perform worse. Adding proven TB cases to the training set is expected to improve the performance of the CAD system. Secondly, there were TB proven images with extremely subtle abnormalities in our collection. Such images are difficult for humans and CAD systems to classify.

The other observation is that the local performances vary greatly for both test sets, from 0.66 to 0.85 on the first set, and from 0.42 to 0.95 on the second set. Such a variation can happen due to low numbers of abnormal samples for some regions in a test set (e.g. the second test set contained only 3 abnormal regions "7" and only 3 abnormal regions "3"). Each misclassification then drastically influences an A_z value for such regions. For some regions, the number of abnormal samples in the training set was also limited, which in general negatively affected the MVDB classification performance on such regions. Future work will include the collection of a much larger data set which we expect to be beneficial for our combination scheme.

In conclusion, we have shown that the multi-valued dissimilarity-based classification is a practical tool that enables a CAD system to deal with weakly labeled images. Combining global and local classification decisions has a potential to improve the overall classification performance. We have been the first to apply such a scheme to the automated detection of tuberculosis in a large database of digital chest radiographs.

Chapter 7

Summary and general discussion

7.1 Summary

The aim of this thesis was to develop automated methods for detection and quantification of interstitial lung disease in conventional chest radiographs and chest CT scans. The focus of the thesis was on pattern classification techniques employed in computer-aided detection systems.

Chapter 2 proposed a classifier that optimized the area under the ROC curve using a linear programming formulation (the AUC-LPC classifier). It appears that the variance of the AUC estimation is smaller than the variance of the classification error, which makes the AUC optimization advantageous for data with highly overlapping or unbalanced classes. The AUC optimization, however, poses a serious computational problem for large data sets, because its complexity is quadratic to the training set size. Subsampling of the training set is conventionally used to decrease the computational complexity. In chapter 2, a new subsampling heuristic was introduced: subsampling of the optimization problem constraints. By randomly subsampling constraints an unbiased constraint distribution was obtained, which led to a better approximation of the AUC than an approximation obtained by subsampling training objects. Furthermore, the constraints that were not used during optimization were used to evaluate the free parameters of the proposed classifier.

The AUC-LPC with the proposed heuristic was used for the automated detection of interstitial abnormalities in chest radiographs, and compared with several standard supervised classifiers on this task. For training purposes, all pixels in radiographs with interstitial abnormalities were considered abnormal, which resulted in highly overlapping normal and abnormal classes in the pixel feature space. A pixel feature vector consisted of local texture features and positional features. A trained classifier, applied to the pixel feature vector, produced a probability of the pixel of being abnormal. Radiographs were classified into normal and abnormal by integrating the classification outputs of individual pixels using a quantile rule. Using the AUC-LPC with constraint subsampling for pixel classification resulted in the best image classification performance, with an AUC of 0.96.

Chapter 3 proposed a new data representation beneficial for classifying weakly

labeled images. An image is considered weakly labeled when only the presence or absence of a disease is recorded but not the precise locations of abnormalities. Without the local ground truth, it is not possible to train a local classifier to discern between normal and abnormal regions or pixels. In chapter 2 we addressed this by assuming that all pixels in abnormal images were abnormal, but such an assumption would be often violated. In this chapter we circumvented the absence of a local ground truth by representing each image via its dissimilarity to a number of normal and abnormal images. A dissimilarity between two images was expressed via dissimilarities between the estimated distributions of individual local texture features. A number of prototype images from each class were randomly selected to serve as bases of comparison. For each image in the data set and each prototype, distances were computed between the histograms of corresponding texture features. A vector of distances to one prototype constituted a dissimilarity representation of the image. Dissimilarity representations of the training images, computed with the same prototype, were used to train a classifier to distinguish between normal and abnormal images. We trained as many classifiers as there were prototypes. Subsequently, the outputs of all classifiers applied to a test image were combined and the final probability of the test image of being abnormal was obtained. Combining classifiers benefited from complementary information provided by different dissimilarity representations of the image.

This classification approach was applied to two sets of chest radiographs: a database from a TB mass screening program, and a database containing images with signs of ILD. The performance of our approach was compared to previously published results obtained on the same data sets with classification systems that used a local ground truth. Our approach was also compared to the standard dissimilarity classification and the region classification that used the same assumption on pixels from abnormal images as in chapter 2. The standard dissimilarity classification differs from our approach in that that it expresses the dissimilarity between two images by just one value and, subsequently, uses a vector of dissimilarities to all the prototypes as an input to a classifier. For all three classification methods, linear discriminant analysis with the same set of local texture features was used. The features were the first four central statistical moments computed from circular ROIs in the filtered versions of the original image. Images were filtered with a multiscale filter bank of Gaussian derivatives up to the 2nd order. The performance of our classification approach was similar to that of previously published methods applied to fully labeled data. Our approach performed similarly or better than the standard dissimilarity classification, and better than the region classification, and achieved an AUC of 0.82 on the TB data set and an AUC of 0.98 on the ILD data set.

Chapter 4 presented a CAD system for the localization of interstitial abnormalities in chest radiographs. In this chapter an emphasis was made on finding a superior local ground truth for training a local classifier. The local ground truth was established by using a chest CT scan of the same patient. Thin 2D CT planes provide excellent contrast resolution, and the lung structures depicted on them

are not superimposed. Therefore, manual delineation of diffuse abnormalities is more feasible in CT slices than in chest radiographs. We estimated the locations of interstitial abnormalities in a radiograph by computing a mapping function between the radiograph and the coronal projection of the corresponding CT scan. Then, this mapping was applied to abnormality outlines delineated in coronal CT slices by an expert. In this way, each pixel within the lung fields in the radiograph could be labeled either normal or affected by ILD, and the local classifier could be trained. Additionally, abnormality subtlety grades were assigned by the expert to abnormal areas in chest radiographs. These grades were used to evaluate the performance of the system on areas that exhibited different levels of visible abnormality. The classifier was trained with the same set of local texture features as the one used in chapter 3. The system output a color-coded probability map that accentuated areas highly probable of being abnormal. It also produced regional scores for eight partitions of the lung fields to enable the comparison of the system performance to that of human observers. The system was shown to perform not significantly different from two radiologists on obviously and relatively obviously abnormal regions (an AUC of 0.92 and 0.81, respectively), but it was significantly worse than humans in detecting subtly abnormal regions (an AUC of 0.67).

In **chapter 5** a system for the automated estimation of ILD progression in serial chest CT scans was proposed. The system compared corresponding 2D axial sections from baseline and follow-up scans and concluded whether this pair of sections represented regression, progression, or unchanged disease status. Alignment of two scans, achieved via non-rigid registration, was an important preprocessing step that enabled the retrieval of matching sections. Two sets of features were investigated for use with the system. The first set comprised statistical features which described the distribution of intensities of the difference image computed between corresponding baseline and follow-up CT sections. The second set consisted of features which characterized textural dissimilarity in a pair of baseline and follow-up images. It contained measures of dissimilarity computed between the estimated distributions of individual local texture features. These features were four central statistical moments computed from circular ROIs in the filtered versions of the original image. Images were filtered with the Gaussian, the Laplacian, and the 1st and 2nd order directional Gaussian derivatives, all computed at three different scales.

The dissimilarity feature set was not able to characterize the direction of change in a pair of CT sections, because a dissimilarity is a symmetric value by definition. And so, it was only used to discern between pairs of CT sections with a change and pairs with stable disease status. We compared the performances of two classification strategies. The first one used the statistical feature set to perform classification of pairs into "regression", "progression" and "stable" classes. The second strategy used the dissimilarity feature set in the first stage, and the statistical feature set in the second stage. Pairs, labeled as "changed" by the first classifier, were further classified into "regression" or "progression" by the second classifier. The performance of the system, in terms of classification

accuracy, was 76.1% and 79.5% with the first and second classification strategies, respectively. The accuracy of the system was not significantly different from that of two radiologists, according to McNemar's test.

In **chapter 6** the classification approach, described in chapter 3, was applied to the detection of TB in digital chest radiographs. A data set contained images with abnormalities suspected of TB, as well as images of the patient whose diagnosis of TB was proven by microbiological tests. Our classification approach was extended to include a simple form of local ground truth. To this end, class labels were obtained for eight fixed partitions of the lung fields. Then, the dissimilarity classification approach was applied to each partition separately, as well as to the whole lung fields. Local and global classification outputs were integrated to obtain an overall image decision. The system was trained with normal images and images with suspect TB. The system validation was performed on two test sets: one containing normal and suspect TB images, and another one containing normal images and images with proven TB. On both test sets, the system that integrated global and local classifications outperformed the system that only used global classification, demonstrating an AUC of 0.83 on the first test set and an AUC of 0.74 on the second test set.

7.2 General discussion

Detection of interstitial lung disease in chest radiographs is a challenging task for radiologists. It was even labeled "a dying art" in one radiological paper [81]. If nothing else, it calls for an experience in this specific area and a systematic approach. Such an expertise is often unavailable in remote hospitals, or for clinical officers who interpret chest x-rays for TB in field hospitals in Africa. Therefore, a dedicated CAD system that detects the presence of ILD or TB in chest radiographs, and, possibly, pinpoint abnormalities, could be very helpful, as a second opinion available to a radiologist. Additionally, it could be used to train the next generation of radiologists.

The conventional design of such a system includes classifying multiple small patches of texture throughout the lung fields (*localization* of abnormalities), and, consequently, integrating local classification outputs into an overall image decision (*detection* of abnormal images). When the localization step works perfectly, the detection step is trivial. In reality, reliable localization of interstitial abnormalities is not a completely solved problem yet, which makes the detection of abnormal images a separate subject of research. This thesis contributed to detection and localization in several ways. We looked into the improvement of local classification by providing a better local reference standard (chapter 4), or by providing a better classifier (chapter 2) when we cannot provide a local reference standard at all. In the latter case, we also investigated how to sidestep local classification altogether and still achieve a good detection performance (chapter 3). Then, we showed that including at least a primitive form of local ground truth in the system design

improves its detection performance (chapter 6). The last contribution of the thesis lay in the kindred area of computed tomography of the lungs. Some of the ideas we applied to the detection of ILD in chest radiographs in the absence of local ground truth proved useful for the estimation of ILD progression in serial CT scans (chapter 5).

In chapters 2 and 3, we assumed that the precise delineations of abnormalities in training data were not available. Therefore, we devised the systems that only used general image labels for training. Although it is a practical and realistic assumption - manual abnormality outlines are not reliable and can be impossible to obtain - the lack of local ground truth for training is expected to cause a CAD system to perform suboptimally. We were able to compare the performances of our systems to that of two systems that used local labeling, either derived from manual abnormality delineations [66] or obtained by applying some empirical rules on measurements extracted from local regions [65]. The CAD systems in chapter 3 and [66] were applied to the same TB data set, while the systems in [65], [66], and chapters 2 and 3 were applied to the same ILD data set. The dissimilarity-based classification approach described in chapter 3 showed no or little difference in performance compared to the systems that used local labeling, which was a very good result. The AUC-LPC classifier described in chapter 2, that used a "naive" assumption that all pixels in an abnormal image were abnormal, in order to circumvent the absence of local ground truth, performed worse than the other three systems on the ILD data set.

Before concluding that the dissimilarity-based classification approach is better in dealing with weakly labeled data than the AUC-LPC, let us have a look at the feature sets that were used with these two methods. Statistical features computed from local neighborhoods were present in both feature sets, but pixel intensities from original and filtered images were excluded in chapter 3. Besides, we used principal component analysis as the means of feature dimensionality reduction in chapter 3. How crucial these differences in local features could be can be observed by comparing the results of the similar "naive" classification approaches in chapters 2 and 3 that both used the LDA for local classification (see Tables 2.7 and 3.3). The LDA showed an average AUC of 0.69 and 0.96 in the experiments in chapters 2 and 3, respectively. Based on these results, we should conclude that local texture features used in chapter 2 were not optimal. We suppose that the performance of the AUC-LPC might have been better than the results reported in chapter 2, had it used a more discriminative feature set, like the one in chapter 3. Nevertheless, the AUC-LPC was able to perform considerably better than the LDA on the same feature set, confirming that the AUC-LPC is a beneficial approach to deal with highly overlapping classes. Because of the "naive" assumption, the negative and positive classes will be highly overlapping in any feature space, which should make the AUC-LPC a more suitable classifier than the LDA regardless of features used.

The fact that the dissimilarity-based classification performed as good or nearly as good as the systems using local labeling, especially on the TB data set which

contained a lot of subtle cases, might implicitly confirm the plausibility of our assumption that manual delineations of interstitial abnormalities are often unreliable and, therefore, can mislead a classifier trained on them. In chapter 4, we presented the system that utilized a superior local reference standard - delineations of interstitial abnormalities obtained using a CT scan of the same patient. The system output a probability of being abnormal for each pixel in the lung fields. Pixel classification outputs could easily be integrated into an image decision, thus, solving the detection task. With this added capability, this system could be another benchmark for the evaluation of methods dealing with weakly labeled data. However, we could not directly apply this system to the ILD data set used in chapters 2 and 3, because the system was trained with a new collection of digital radiographs, while the early ILD data were digitized films. In the future, by adding the detection facility to the system in chapter 4 and applying dissimilarity-based classification to the same digital data set, we can compare performances of the two systems. They utilize the same local texture features, which will make a comparison between them even more instructive.

The classification results presented in chapter 4 provided several interesting observations. First, it showed that localization of interstitial abnormalities in chest radiographs was, indeed, a difficult task for radiologists, especially discerning between normal and subtly abnormal regions. The comparison between the system and the observers revealed that, when perihilum regions were excluded from evaluation, the performance of the system increased more than that of the radiologists (from an AUC of 0.80 to 0.85 for the system and from an AUC of 0.86 and 0.87 to 0.88 for the two observers). It means that classification of the perihilum region posed specific difficulties for our system that was based on the analysis of texture. To our knowledge, radiologists usually pay attention to the size and shape of the perihilum region, and we propose to include equivalent features in a future system.

After the exclusion of perihilum regions from evaluation, the performance of the system on subtly abnormal regions was still considerably lower than that of the observers. Detection of subtle lesions is the biggest challenge in building a CAD system for ILD. We hypothesize that the informatively superior reference standard we used to train the system in chapter 4 might have been misleading in case of subtle lesions, by indicating a lesion where it cannot be perceived in the chest radiograph, and, therefore, cannot be properly described by texture features. But this is just a preliminary hypothesis that requires further investigation. Such an investigation could include the retrospective review of radiographs by an expert radiologist, in order to determine whether small abnormalities found in CT images could also be seen in radiographs. If the expert discarded a lot of subtle abnormalities as invisible, we would know that the reference standard provided by our method was over-informative. On the other hand, it is possible that we have not yet employed the most powerful features able to cope with subtle textural abnormalities characteristic to ILD. In this system general purpose texture features were used. In future, to complete the system, an effort should be made to find the

most discriminative texture features for interstitial abnormalities in chest radiographs. The use of the AUC-LPC with our system for local classification should be investigated too.

In chapters 5 and 6 we presented two more CAD applications that utilized dissimilarities between images. Dissimilarities are attractive features because they provide a natural way of describing a class of similar images. The limitation of using them is the generalization of local information that occurs during computation of dissimilarities. A dissimilarity measure is computed between feature histograms characterizing the distribution of feature values throughout the whole lung fields. Differences in feature values caused by a subtle small abnormality are likely to be indistinct in the resulting histogram and in the corresponding value of dissimilarity. For example, a difference between the distributions of a certain feature, computed from the normal lungs and the lungs with a subtle small abnormality, could be indiscernible from inherent small differences between the distributions of the feature computed from the normal lungs of two different persons.

In chapter 6 we subdivided the lung fields in eight parts and computed feature histograms for each part separately, in order to decrease the level of generalization. Each part as well as the whole lung fields were classified using the dissimilarity-based classification approach as described in chapter 3. Then, the classification outputs were combined to obtain a final image decision. Although combining local and global outputs moderately improved image classification performance, for many lung parts classification performance suffered from unrepresentative training sets. We believe, this prevented a larger improvement of overall classification performance. Expanding the data set, so that it includes more abnormalities in different locations in the lungs, could help. Alternatively, instead of classifying separate lung parts, the dissimilarity features from each part can be concatenated into one vector and used to classify the whole image directly. In this way, no local ground truth is required for training the system. However, for such an approach to be feasible, a smaller set of dedicated local texture features has to be found first, otherwise the size of the feature space will increase enormously.

In chapter 5 we used dissimilarities to describe pairs of images, and a difference between a pair of images was itself a subject of classification. This role of dissimilarities is different from their usage in chapters 3 and 6. There, a vector of dissimilarities to another image comprised the representation of the image at hand, and it was beneficial for classification to have several representations of one image. The utilization of dissimilarities improved the performance of the classification system in chapter 5 compared to its performance when only statistical features were used. However, we believe that the improvement could be more noticeable, had all images in the data set been acquired under the same conditions. The set of CT scans the system was applied to contained images collected from a number of different CT scanners that had, for example, different convolution kernels. The settings for radiation exposure varied from image to image too, which resulted in different levels of noise in the scans. There were also other

differences in scanning protocols. Such a diversity of image acquisition parameters was probably less influential in simple statistical features computed from the difference image. But we believe that for local texture features that were used for computing dissimilarities, image acquisition parameters were likely to play a more important role.

When we devised a system to perform automated estimation of ILD progression, at first we tried to employ the results of computerized texture analysis that segmented an axial CT section into regions of different texture types, as described in [88]. Computing differences between various texture types in baseline and follow-up images would produce an estimate of change in disease extent. A set of previously collected CT sections with manual annotations was available for training a texture analysis algorithm. But a pilot system for the estimation of ILD progression based on texture analysis showed quite unsatisfactory results. Texture analysis, that employed the same local texture features as described in chapter 5, did not perform well on new data acquired with different acquisition characteristics.

Such an inflexibility of texture analysis was one of the reasons why we eventually devised the system presented in chapter 5 that did not require training with previously collected data. Still, we assume that the diversity of image acquisition parameters existing in the current data set might have affected local texture features and caused some of classification errors. It might be advantageous for the system performance if a preprocessing step is introduced that estimates and decreases the level of noise in images without smoothing them. Additional investigation is needed to find out whether different convolution kernels, or other inherent characteristics of CT scanners, influence texture features and, if yes, how to allow for it during preprocessing. A successful implementation of necessary "image equalizing" steps would also revive the applicability of texture analysis, if it could be used without annotating a new training set every time the system is applied to images from a differen scanner or acquired with different settings.

We would like to conclude this general discussion by noting that in spite of not showing perfect or near perfect classification performances as stand-alone systems, the CAD applications presented in this thesis could be valuable instruments in clinical practice. They were intended to support radiologists in making decisions by providing them with a second opinion. For a radiologist, this is equivalent of getting an opinion of a colleague. An obvious next step in our research would be to conduct several observer studies in order to compare the performances of radiologists with and without the assistance of these CAD systems. The systems in this thesis were trained using, directly or indirectly, the reference standard provided by experts. In this way, a trained CAD system can be compared to a student that learnt from his teachers. By performing not significantly different from radiologists in a number of experiments described here, we believe, our "students" showed some intelligence, and there will be more improvement to come.

Samenvatting

Dit proefschrift beschrijft methoden om met behulp van computerprogramma's interstitiële longziekten en tuberculose te detecteren en de progressie ervan te kwantificeren. Het onderzoek richt zich op twee-dimensionale Röntgenfoto's en drie-dimensionale computertomographie (CT) scans. De nadruk ligt op het gebruik van technieken uit de patroonherkenning zoals die toegepast worden in de computer-ondersteunde diagnose van medische beelden. Hoofdstuk 1 bevat een inleiding met achtergrondinformatie over patroonherkenning en medische beeldverwerking, interstiële longziekten, Röntgenfoto's en CT scans.

In Hoofdstuk 2 wordt een nieuwe classificatietechniek voorgesteld die niet, zoals gebruikelijk, het aantal correcte classificaties probeert te optimaliseren maar een andere maat, de oppervlakte onder de ROC curve. Deze oppervlakte is een manier om te meten in hoeverre de normale en abnormale samples gescheiden zijn wanneer ze gesorteerd gerangschikt worden. Deze techniek is geschikt voor problemen waarbij er een onbalans is in de trainingsdata en waarbij de klassen overlappen. De optimalisatie is echter computationeel complex omdat deze kwadratisch toeneemt met het aantal trainingssamples. Een nieuwe optimalisatie wordt voorgesteld waarbij een selectie gemaakt wordt van de randvoorwaarden in plaats van het aantal trainingssamples en dit leidt tot betere resultaten. De methode wordt toegepast op de classificatie van Röntgenfoto's van patiënten met en zonder interstitiële ziekte. Om het systeem te trainen worden alle pixels uit de longvelden van de (ab)normale beelden worden als (ab)normaal beschouwd en hierdoor overlappen de klassen sterk. Na classificatie van alle pixels worden deze gecombineerd tot een score voor het hele beeld met een quantielregel.

Hoofdstuk 3 stelt een nieuwe manier voor om data te representeren die voordelig uitpakt voor zwakgelabelde data. Een beeld wordt beschouwd als zwakgelabeld als alleen bekend is dat het hele beeld normaal of abnormaal is, maar niet bekend is waar de abnormaliteiten zich precies bevinden in de abnormale beelden. Dit is een veel voorkomende situatie in de praktijk omdat een radioloog niet in staat is precies aan te geven waar in een Röntgenfoto het normale gebied ophoudt en het abnormale gebied begint. Als niet bekend is welke pixels abnormaal zijn kunnen we in principe geen pixelgebaseerd classificatiesysteem trainen. In hoofdstuk 2 werd dit probleem omzeild door simpelweg aan te nemen dat alle pixels in een abnormaal beeld abnormaal zijn, maar deze aanname is niet altijd correct. In dit hoofdstuk benaderen we het probleem anders: we verlaten de pixelgebaseerde aanpak en vergelijken beelden met elkaar door afstanden te berekenen tussen de

verdeling van verschillende lokale textuurmaten. Door voor elk beeld de afstand tot een aantal prototype beelden te berekenen kunnen we de beelden classificeren voor elke textuurmaat, en deze classificaties combineren tot een eindresultaat. De methode wordt toegepast op de classificatie van Röntgenfoto's van patiënten met en zonder interstitiële ziekte en op Röntgenfoto's van patiënten met en zonder tuberculose. We tonen aan dat deze methode vergelijkbaar presteert als methoden die wel getraind zijn met locale labels.

In Hoofdstuk 4 gebruiken we driedimensionale CT scans, geanalyseerd door een radioloog, om veel preciezer te kunnen bepalen waar in tweedimensionale Röntgenfoto's de abnormaliteiten zich bevinden van patiënten met en zonder interstitiële longziekte. De afwijkingen die zijn aangegeven in de CT data worden automatisch overgebracht naar het Röntgenbeeld. Een computersysteem wordt getraind dat voor elk pixel in de longvelden bepaalt wat de waarschijnlijkheid is dat dit pixel in een abnormaal gebied ligt. Dit wordt door middel van een kleurenoverlay weergegeven aan de gebruiker. Om het systeem te kunnen vergelijken met radiologen worden de longvelden elk automatisch in vier gebieden verdeeld. Radiologen en het computersysteem scoren vergelijkbaar wanneer ze van deze gebieden moeten aangeven of ze abnormaal zijn, zolang de abnormaliteiten duidelijk of subtiel zijn. Alleen bij zeer subtiele afwijkingen en afwijkingen in de regio van de hilus scoren de radiologen beter.

Hoofdstuk 5 beschrijft een nieuwe toepassing van computer-ondersteunde diagnose. In CT scans van patiënten met interstitiële longziekte wordt de progressie van de ziekte gemeten. Dit gebeurt door de follow-up beelden te vervormen zodat ze per plak vergelijkbaar worden. Het systeem werkt daarna in twee stappen. Eerst wordt bepaald of er sprake is van een stabiele situatie of van verandering. In het laatste geval wordt vervolgens bepaald of er progressie of regressie van de ziekte is waar te nemen. Twee verschillende sets van karakteristieken om deze beslissing te kunnen nemen worden vergeleken: een die naar de verdeling van densiteiten in de longvelden kijkt, en een die gebaseerd is op textuurmaten en die de methode van afstanden tussen beelden uit hoofdstuk 3 gebruikt. De computersystemen worden vergeleken met twee radiologen. Mens en computer blijken niet significant verschillend te presteren.

In Hoofdstuk 6 wordt de methode uit hoofdstuk 3 toegepast op een grote database van een tuberculosescreening met digitale Röntgenbeelden. De methode wordt uitgebreid door de analyse ook uit te voeren per longgebied, en hiervoor wordt de onderverdeling van de longen in vier gebieden uit hoofdstuk 4 gebruikt. De analyse van het hele beeld en de analyses per longgebied geven negen oordelen die worden gecombineerd tot een eindoordeel waarbij de computer de waarschijnlijkheid schat dat het beeld normaal dan wel verdacht of abnormaal is. Hoewel het systeem is getraind met normale beelden en beelden die in de screening als verdacht zijn afgegeven, blijkt het niet veel slechter te werken wanneer het getest wordt op een database met normalen en gevallen van bewezen tuberculose.

Het proefschrift sluit af met een samenvatting en algemene discussie in Hoofdstuk 7.

Publications

Publications related to this thesis

Papers in international journals

D.M.J. Tax, Y. Arzhaeva, R.P.W. Duin and B. van Ginneken, "AUC optimization by subsampling constraints," *in preparation.*

Y. Arzhaeva, M. Prokop, K. Murphy, E.M. van Rikxoort, P.A. de Jong, H.A. Gietema, M.A. Viergever and B. van Ginneken, "Automated estimation of progression of interstitial lung disease in CT images", *submitted*, 2009.

Y. Arzhaeva, D.M.J. Tax and B. van Ginneken, "Dissimilarity-based classification in the absence of local ground truth: Application to the diagnostic interpretation of chest radiographs," *Pattern Recognition*, *Pattern Recognition*, vol. 42, no. 9, pp. 1768–1776, 2009.

Y. Arzhaeva, M. Prokop, D.M.J. Tax, P.A. de Jong, C.M. Schaefer-Prokop and B. van Ginneken, "Computer-aided detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography," *Medical Physics*, vol. 34, no. 12, pp. 4798–4809, 2007.

Papers in conference proceedings

Y. Arzhaeva, L. Hogeweg, P.A. de Jong, M. A. Viergever and B. van Ginneken, "Global and local multi-valued dissimilarity-based classification: application to computer-aided detection of tuberculosis", in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2009.

Y. Arzhaeva, D.M.J. Tax and B. van Ginneken, "Improving computer-aided diagnosis of interstitial disease in chest radiographs by combining one-class and two-class classifiers," in *SPIE Medical Imaging*, vol. 6144, 2006.

Y. Arzhaeva, B. van Ginneken and D. M. J. Tax, "Image classification from generalized image distance features: application to detection of interstitial disease in chest radiographs", in *International Conference on Pattern Recognition*, 2006.

D.M.J. Tax, R.P.W. Duin, Y. Arzhaeva, "Linear model combining by optimizing the Area under the ROC curve", in *International Conference on Pattern Recognition*, 2006.

Abstracts in conference proceedings

L. Hogeweg, Y. Arzhaeva, P.A. de Jong, M. Prokop and B. van Ginneken, "Computer-aided detection of tuberculosis from chest radiographs using a dissimilarity approach," *submitted*, 2009.

K. Murphy, M. Prokop, C.M. Schaefer-Prokop, H.A. Gietema, G.D. Nossent, Y. Arzhaeva, B. van Ginneken and J.P.W. Pluim, "Improved efficiency of assessment of interstitial lung disease progression in HRCT of the chest by visualization of automatically-registered image pair", in *Radiological Society of North America*, 94th Annual Meeting, 2008.

Y. Arzhaeva, K. Murphy, M. Prokop, C.M. Schaefer-Prokop and B. van Ginneken, "Application of computerized texture analysis of CT lung images for estimation of interstitial lung disease progression", in *Radiological Society of North America*, 93th Annual Meeting, 2007.

Y. Arzhaeva, M. Prokop, C.M. Schaefer-Prokop, P.A. de Jong and B. van Ginneken, "Computer-Aided Detection of Interstitial Abnormalities in Chest Radiographs", in *Radiological Society of North America*, 93th Annual Meeting, 2007.

Other publications

T. Heimann, B. van Ginneken, M. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. Cashman, Y. Chi, A. Cordova, B. Dawant, M. Fidrich, J. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmuller, R. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H.-P. Meinzer, G. Nemeth, D. Raicu, A.-M. Rau, E. M. van Rikxoort, M. Rousson, L. Rusko, K. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. Waite, A. Wimmer and I. Wolf, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Transactions on Medical Imaging*, Epub ahead of print, Feb 10, 2009.

E.M. van Rikxoort, I. Isgum, Y. Arzhaeva, M. Staring, S. Klein, M.A. Viergever, J.P.W. Pluim and B. van Ginneken, "Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus," *submitted*, 2008.

E.M. van Rikxoort, Y. Arzhaeva and B. van Ginneken, "A multi-atlas approach to automatic segmentation of the caudate nucleus in MR brain images," in *3D Segmentation in the Clinic: A Grand Challenge*, pp. 29–36, 2007.

Y. Arzhaeva, E.M. van Rikxoort and B. van Ginneken, "Automated segmentation of caudate nucleus in MR brain images with voxel classification," in *3D Segmentation in the Clinic: A Grand Challenge*, pp. 65–72, 2007.

E.M. van Rikxoort, Y. Arzhaeva and B. van Ginneken, "Automatic segmentation of the liver in computed tomography scans with voxel classification and atlas matching," in *3D Segmentation in the Clinic: A Grand Challenge*, pp. 101–108, 2007.

Acknowledgments

I started this part of the book so many times in my mind. There are so many things I am grateful for, and now there is an opportunity to rave about it as much as I like. I'll start with the country and then downscale. Dear Netherlands (may I call you like this?), it was more than a pleasure to be your guest for over five years. I felt right, I felt liked and accepted. Maybe, it was so because of an unobtrusive Dutch hospitality (a.k.a. gezelligheid) or because I never mastered the language well enough to understand how the society worked and to feel its pressure. Or, maybe, because I spent a lot of time in academic surroundings. But I prefer to think about it that a special affinity exists between me and the Dutch ways of being.

I have been always fascinated by hospitals, and it was one of the reasons why I applied for this PhD position. The UMCU immediately seemed like a place where I wanted to work. I liked how it looked - big, light and beautiful, with flowers in a flower shop and fish in a small pond. And some alien-looking guy in a green outfit wished me good luck with my interview when I was propping the wall in front of Max's office. It was a bit of luck indeed that this position was eventually offered to me - I was a runner-up in the list of candidates, according to Bram. I felt lucky then, when I got in, and I continue feeling fortunate that this whole experience happened to me. It wouldn't be true to say that my employment with the UMCU was pure joy day in, day out. Everybody who has been through PhD studies knows how frustrating it can be at times. But I like my difficulties too, as a part of the whole package.

"Image Sciences Institute, University Medical Center Utrecht" - that's what I proudly put in my papers. I am indeed very proud that I belonged to this research institution. The ISI has been the best place for me in professional and personal sense. Dear Max, thank you for making and keeping the ISI such a special place.

Max Viergever and Mathias Prokop have been my promotors. Max, thank you for being caring and understanding, for having time for me in spite of your busy schedule and various responsibilities, for believing in me when I left to join my husband in Australia, and for being a meticulous reviewer.

Mathias, it was such a joy to collaborate with you. Your experience as a chest and heart radiologist was indispensable for our group. Your energy, your passion for CAD projects and inspirational ideas were so motivating, and your charm, humor and kindness never failed to lift my spirits. Our "pulmo" meetings have been my all time favorite. Mathias, thank you for all the observer studies you did

for me and for your significant contributions to our papers.

Bram, you have been a terrific supervisor. I am so happy I have known you and worked closely with you for five years. You have truly been my teacher in medical image analysis and CAD, and more. When you talked about the role and opportunities of our science for medical imaging, it was extremely motivating for me. I admire your scientific intuition, the mastery of our field of knowledge and your leadership qualities. You are so good at critical thinking, debating, and asking difficult questions that make one think. I hope I have learnt something from you. Bram, thank you for always being there for me, for making a long-distance supervision a real thing, for emailing and skyping incessantly despite a time difference. Thank you for being open and kind to me and for keeping our student-supervisor relationship so joyfully informal.

For the last year of the five years that I needed to accomplish my thesis, I lived and worked full-time in Australia. I should say a million thanks to my CAD group fellows - the biggest source of my motivation during this year of overwhelmingly long hours and non-existent weekends. Guys, you might not have known that but you gave me a lot of energy to go on. Poetically speaking, you were my light at the end of a tunnel. Bram, Eva, Meindert, Ivana, Keelin, Adriënne, Arnold, Ingrid and Marco - every time I sat to work on my thesis, here on the other side of the globe, the images of you filled my mind. At my time with the ISI I took all the pleasures for granted - our enjoyable coffee breaks and lunches, jokes and chats, CAD dinners, ice skating nights and all the fun at conferences. Now I know it was a blessing. To have a team at work that feels like a family does not happen that often.

Ingrid, you might not believe it, but here I am, the person who has read your thesis from cover to cover more than once. I enjoyed your clear and concise writing and learnt how to write scientific papers by your example. Apart from professional matters, I admire your keen mind, fighting spirit and amusing talk.

Marco, thanks for all the spice and fun, you are simply unique. And we had the most interesting BVD meetings when you were around.

Ivana, thank you for giving me a heartfelt support and understanding when I needed it most, for your frankness and warmth, and for sharing your experiences as a fellow expat.

Keelin, thank you for introducing me to a Sunday soccer in Westerpark in Amsterdam, for your help with image registration issues, for reading and correcting my English essays and journal papers. I enjoyed our conversations and your wonderful emails a lot.

My office mates of different times, Arnold, Meindert, Eva and Adriënne, you were perfect mates for me. Thank you for all the joy of an easy-going communication and for help and support you gave me on different occasions.

Arnold, once in Australia I mistook some person for you. During few seconds that passed before I understood my mistake I was violently attacked by sudden emotions of joy and happiness. You were an awesome colleague for me.

Adriënne, with three girls in the office, it was the most stimulating working

environment I ever had. Every hour of hard work was rewarded. It was a pleasure to share the office with you.

Meindert, thank you for being a good friend for me, for good times we had together in and out of the office, and for many interesting conversations we had on all sorts of topics. You provided me with lots of insights in Dutch culture and character. Apart from that, your help as my "paranimf" is much needed and appreciated.

Eva, in those four years we have been together through it all. I feel that I can rely on you in everything, from parties to deadlines. We were writing up our theses on different continents but side by side, communicating daily about our frustrations and with words of support. It helped me a lot in those crazy weeks. You are such a great friend and a wonderful joyous person.

Our software, IX, has had many contributors, and I am grateful to all of them. Bram and Joes, I mastered and came to love object-oriented programming mostly by uncovering your code. Thanks for being my "gurus" in these matters. Joes, I also appreciate your being such a faithful alumnus of the CAD group and always showing up in the ISI-CAD blog.

I would like to thank all my co-authors and collaborators for their valuable contributions to various chapters of this thesis. David Tax, thank you for all the help with pattern recognition theory and practice, for enlightening discussions and fun chats. Your scientific enthusiasm was contagious, and your teaching abilities inspired my admiration. Pim de Jong, Hester Gietema and Cornelia Schaefer-Prokop, thank you very much for carrying out the observer studies for me, without them this research would not be valid. Laurens Hogeweg, I know you only by voice but I am looking forward to meeting you in the Netherlands. Thank you for implementation and setting up the TB observer study.

The data set used in the sixth chapter of my thesis has been obtained from the GGD Hart van Brabant in Tilburg. I am extremely grateful to Marcel Berkel, Ton Froklage and Walid Haddad for cooperation and help with searching and evaluating the data and for burning CD disks for me. The personnel of the GGD made me feel welcome on the days I spent with them.

Many people from the ISI contributed to my research in some ways - by discussions, teaching, or just creating a great social atmosphere. In particular, I would like to thank Mirela Tanase, Marleen de Bruijne, Josien Pluim, Marius Staring, Stefan Klein, Michiel Schaap, Rashindra Manniesing, Wiro Niessen, Everine van de Kraats and Ewoud Smit. I am very grateful to Gerard van Hoorn for a prompt technical support, both on-site and remote. Dear Marjan, Jacqueline and Renee, you were always most helpful whenever I had a general inquiry or problem. I am especially grateful for your help when I was in a housing crisis. Wilbert Bartels, Koen Vincken and the others, thanks for organizing such an unforgettable Xmas party in the castle.

I am grateful to my friend and "paranimf" Natalia Stash for hospitality, entertainment and help. Besides, you were an example for me how to finish a thesis while working full-time. Finally, I thank all my friends and family. Especially,

my mother and Igor for their inexhaustible emotional support.

Bibliography

[1] British Thoracic Society, "BTS guidelines on the diagnosis, assessment and treatment of diffuse parenchymal lung disease in adults," *Thorax*, vol. 54, no. Supplement 1, pp. S24–S30, 1999.

[2] A. U. Wells and N. Hirani, "Interstitial lung disease guideline: the British Thoracic Society in collaboration with the Thoracic Society of Australia and New Zealand and the Irish Thoracic Society," *Thorax*, vol. 63, no. Suppl V, pp. v1–v58, 2008.

[3] American Thoracic Society, "American Thoracic Society/European Respiratory Society international multidisciplinary consensus classification of the idiopathic interstitial pneumonias," *American Journal of Respiratory and Critical Care Medicine*, vol. 165, pp. 277–304, 2002.

[4] World Health Organization, "WHO Report 2009: Global tuberculosis control, Epidemiology, Strategy, financing," 2009.

[5] World Health Organization, "The Stop TB Strategy. building on and enhancing DOTS to meet TB-related millennium development goals," 2006.

[6] I. A. Campbell and O. Bah-Sow, "Pulmonary tuberculosis: diagnosis and treatment," *British Medical Journal*, vol. 332, pp. 1194–1197, 2006.

[7] D. A. Lynch, W. D. Travis, N. L. Müller, J. R. Galvin, D. M. Hansell, P. A. Grenier, and T. King Jr, "Idiopathic interstitial pneumonias: CT features," *Radiology*, vol. 236, pp. 10–21, 2005.

[8] W. C. Röntgen, "Über eine neue Art von Strahlen," *Sitzungsberichte der Physikalisch-Medicinisch Gesellschaft zu Würzburg*, pp. 132–141, 1895.

[9] H. P. McAdams, E. Samei, J. Dobbins III, G. D. Tourassi, and C. E. Ravin, "Recent advances in chest radiography," *Radiology*, vol. 241, no. 3, pp. 663–683, 2006.

[10] M. B. Gotway, "Interstitial lung diseases: imaging evaluation," *Applied Radiology*, vol. 29, no. 9, pp. 31–46, 2000.

[11] G. Hounsfield, "Computerized transverse axial scanning (tomography): Part I. Description of system," *British Journal of Radiology*, vol. 46, pp. 1016–1022, 1973.

[12] J. Radon, "On the determination of function from their integrals along certain manifolds," *Ber. Saechs. Akad. Wiss. Leipzig Math. Phys. Kl.*, vol. 69, pp. 262–277, 1917.

[13] M. Prokop and M. Galanski, *Spiral and multislice computed tomography of the body*. Stuttgart, Germany: Thieme, 2003.

[14] M. L. Giger, H.-P. Chan, and J. Boone, "Anniversary paper: history and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM," *Medical Physics*, vol. 35, no. 12, pp. 5799–5820, 2008.

[15] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, vol. 31, no. 4-5, pp. 198–211, 2007. PMID: 17349778.

[16] B. van Ginneken, B. M. ter Haar Romeny, and M. A. Viergever, "Computer-aided diagnosis in chest radiography: a survey," *IEEE Transactions on Medical Imaging*, vol. 20, no. 12, pp. 1228–1241, 2001.

[17] I. C. Sluimer, A. M. R. Schilham, M. Prokop, and B. van Ginneken, "Computer analysis of computed tomography scans of the lung: a survey," *IEEE Transactions on Medical Imaging*, vol. 25, no. 4, pp. 385–405, 2006.

[18] H.-P. Chan, L. Hadjiiski, C. Zhou, and B. Sahiner, "Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography-a review," *Academic Radiology*, vol. 15, no. 5, pp. 535–555, 2008.

[19] H. Abe, H. Macmahon, R. Engelmann, Q. Li, J. Shiraishi, S. Katsuragawa, M. Aoyama, T. Ishida, K. Ashizawa, C. e. Metz, and K. Doi, "Computer-aided diagnosis in chest radiography: Results of large-scale observer tests at the 1996–2001 RSNA scientific assemblies," *Radiographics*, vol. 23, pp. 255–265, 2003.

[20] J. Shiraishi, H. Abe, F. Li, R. Engelmann, H. MacMahon, and K. Doi, "Computer-aided diagnosis for the detection and classification of lung cancers on chest radiographs ROC analysis of radiologists' performance," *Academic Radiology*, vol. 13, no. 8, pp. 995–1003, 2006. PMID: 16843852.

[21] M. S. Brown, J. G. Goldin, S. Rogers, H. J. Kim, R. D. Suh, M. F. McNitt-Gray, S. K. Shah, D. Truong, K. Brown, J. W. Sayre, D. W. Gjertson, P. Batra, and D. R. Aberle, "Computer-aided lung nodule detection in CT: Results of large-scale observer test," *Academic Radiology*, pp. 681–686, 2005.

[22] K. Awai, K. Murao, A. Ozawa, Y. Nakayama, T. Nakaura, D. Liu, K. Kawanaka, Y. Funama, S. Morishita, and Y. Yamashita, "Pulmonary nodules: estimation of malignancy at thin-section helical CT–effect of computer-aided diagnosis on performance of radiologists," *Radiology*, vol. 239, no. 1, pp. 276–284, 2006.

[23] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. New York: John Wiley and Sons, 2nd ed., 2001.

[24] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[25] B. van Ginneken, M. B. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Medical Image Analysis*, vol. 10, no. 1, pp. 19–40, 2006.

[26] S. Hu, E. Hoffman, and J. Reinhardt, "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images," *IEEE Trans Med Imaging*, vol. 20, pp. 490–498, 2001.

[27] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer Jr., "Quo vadis, atlas-based segmentation?," in *The Handbook of Medical Image Analysis – Volume III: Registration Models*, (New York, NY), pp. 435–486, Kluwer Academic / Plenum Publishers, 2005.

[28] I. Sluimer, M. Prokop, and B. van Ginneken, "Towards automated segmentation of the pathological lung in CT," *IEEE Trans Med Imaging*, vol. 24, no. 8, pp. 1025–1038, 2005.

[29] C. E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology*, vol. 21, no. 9, pp. 720–733, 1986.

[30] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[31] R. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proceedings of the International Conference on Machine Learning*, pp. 445–453, 1998.

[32] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[33] C. Ling, J. Huang, and H. Zhang, "AUC, a better measure than accuracy in comparing learning algorithms," in *Advances in Artificial Intelligence*, vol. 2671 of *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, 2003.

[34] Y. Freund, R. Iyer, R. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, pp. 933–969, 2003.

[35] C. Ferri, P. Flach, and J. Hernández-Orallo, "Learning decision trees using the area under the ROC curve," in *Proceedings of the International Conference on Machine Learning*, 2002.

[36] A. Rakotomamonjy, "Optimizing AUC with support vector machine," in *European Conference on Artificial Intelligence, Workshop on ROC Curve and AI*, 2004.

[37] D. M. J. Tax, R. P. W. Duin, and Y. Arzhaeva, "Linear model combining by optimizing the Area under the ROC curve," in *Proceedings of the International Conference on Pattern Recognition*, 2006.

[38] K. Ataman, W. N. Street, and Y. Zhang, "Learning to rank by maximizing AUC with linear programming," in *Proceedings of IEEE International Joint Conference on Neural Networks*, 2006.

[39] A. Rakotomamonjy, "SVMs and area under ROC curve," tech. rep., 2004.

[40] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in *Advances in Neural Information Processing Systems 16*, pp. 313–320, 2003.

[41] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[42] A. Asuncion and D. J. Newman, "UCI machine learning repository." University of California, Irvine, School of Information and Computer Sciences, 2007. Available at `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[43] C. Bhattacharyya, L. Grate, A. Rizki, D. Radisky, F. Molina, M. Jordan, M. Bissel, and I. Mian, "Simultaneous classification and relevant feature identification in high-dimensional spaces: Application to molecular profiling data," *Signal Processing*, vol. 83, pp. 729–743, 2003.

[44] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization Methods and Software*, vol. 1, pp. 23–24, 1992.

[45] J. T. Kwok and I. W. Tsang, "Learning with idealized kernels," in *Proceedings of the International Conference on Machine Learning*, 2003.

[46] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[47] U. Brefeld and T. Scheffer, "AUC maximizing support vector learning," in *International Conference on Machine Learning, Workshop on ROC analysis in Machine Learning*, 2005.

[48] O. L. Mangasarian, "Linear and nonlinear separation of patterns by linear programming," *Operations Research*, vol. 13, pp. 444–452, 1965.

[49] C. J. Veenman and D. M. J. Tax, "LESS: a model-based classifier for sparse subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 9, pp. 1496–1500, 2005.

[50] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[51] 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, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[52] R. P. W. Duin, "On the choice of the smoothing parameters for Parzen estimators of probability density functions," *IEEE Transactions on Computers*, vol. C-25, no. 11, pp. 1175–1179, 1976.

[53] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[54] A. Makhorin, "GNU linear programming kit," 2000. Software available at `http://www.gnu.org/software/glpk`.

[55] S. Katsuragawa, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography: classification of normal and abnormal lungs with interstitial lung disease in chest radiographs," *Medical Physics*, vol. 16, no. 1, pp. 38–44, 1989.

[56] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham, "Active shape models – their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[57] D. Ruprecht and H. Müller, "Image warping with scattered data interpolation," *Computer Graphics and Applications*, vol. 15, no. 2, pp. 37–43, 1995.

[58] M. Loog and B. van Ginneken, "Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification," *IEEE Transactions on Medical Imaging*, vol. 25, pp. 602–611, 2006.

[59] C. M. Bishop and I. Ulusoy, "Object recognition via local patch labelling," in *Workshop on Machine Learning* (J. Winkler, N. Lawrence, and M. Niranjan, eds.), p. 1, 2004.

[60] E. Pekalska, *Dissimilarity representations in pattern recognition.* PhD thesis, Delft University, the Netherlands, 2005.

[61] E. Pekalska and R. P. W. Duin, "Dissimilarity representations allow for building good classifiers," *Pattern Recognition Letters*, vol. 23, pp. 943–956, 2002.

[62] J. Puzicha, T. Hofmann, and J. M. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *Proceedings of Computer Vision and Pattern Recognition*, p. 267, IEEE Computer Society, 1997.

[63] D. S. Guru and B. B. Kiranagu, "Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic patterns," *Pattern Recognition*, vol. 38, pp. 151–156, 2005.

[64] T. Ishida, S. Katsuragawa, K. Ashizawa, H. MacMahon, and K. Doi, "Artificial neural networks in chest radiographs: detection and characterization of interstitial lung disease," in *Proceedings of the SPIE*, vol. 3034, pp. 931–937, 1997.

[65] T. Ishida, S. Katsuragawa, K. Ashizawa, H. MacMahon, and K. Doi, "Application of artificial neural networks for quantitative analysis of image data in chest radiographs for detection of interstitial lung disease," *Journal of Digital Imaging*, vol. 11, no. 4, pp. 182–192, 1998.

[66] B. van Ginneken, S. Katsuragawa, B. M. ter Haar Romeny, K. Doi, and M. A. Viergever, "Automatic detection of abnormalities in chest radiographs using local texture analysis," *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 139–149, 2002.

[67] E. Pekalska, R. P. W. Duin, and P. Paclik, "Prototype selection for dissimilarity-based classifiers," *Pattern Recognition*, vol. 39, pp. 189–208, 2006.

[68] "Mathworld." `http://mathworld.wolfram.com/`.

[69] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[70] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[71] S. Katsuragawa, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography: detection and characterization of interstitial lung disease in digital chest radiographs," *Medical Physics*, vol. 15, no. 3, pp. 311–319, 1988.

[72] M. Unser and M. Eden, "Multi-resolution feature extraction and selection for texture segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 717–728, 1989.

[73] B. van Ginneken and B. M. ter Haar Romeny, "Multi-scale texture classification from generalized locally orderless images," *Pattern Recognition*, vol. 36, pp. 899–911, 2002.

[74] Y. Arzhaeva, M. Prokop, D. M. J. Tax, P. A. de Jong, C. M. Schaefer-Prokop, and B. van Ginneken, "Computer-aided detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography," *Medical Physics*, vol. 34, no. 12, pp. 4798–4809, 2007.

[75] I. C. Sluimer, P. F. van Waes, M. A. Viergever, and B. van Ginneken, "Computer-aided diagnosis in high-resolution CT of the lungs," *Medical Physics*, vol. 30, no. 12, pp. 3081–3090, 2003.

[76] Y. Arzhaeva, D. M. J. Tax, and B. van Ginneken, "Improving computer-aided diagnosis of interstitial disease in chest radiographs by combining one-class and two-class classifiers," in *Proceedings of the SPIE* (J. M. Reinhardt and J. P. W. Pluim, eds.), vol. 6144, p. 614458, 2006.

[77] M. Loog, B. van Ginneken, and M. Nielsen, "Detection of interstitial lung disease in PA chest radiographs," in *SPIE Medical Imaging: Physics of Medical Imaging* (M. Jaffe and M. Flynn, eds.), vol. 5368, pp. 848–855, 2004.

[78] E. A. Kazerooni, "High-resolution CT of the lungs," *American Journal of Roentgenology*, vol. 177, no. 3, pp. 501–519, 2001.

[79] H. Abe, K. Ashizawa, F. Li, N. Matsuyama, A. Fukushima, J. Shiraishi, H. MacMahon, and K. Doi, "Artificial neural networks (ANNs) for differential diagnosis of interstitial lung disease: results of a simulation test with actual clinical cases," *Academic Radiology*, vol. 11, no. 1, pp. 29–37, 2004.

[80] G. Raghu and K. K. Brown, "Interstitial lung disease: clinical evaluation and keys to an accurate diagnosis," *Clinics in Chest Medicine*, vol. 25, pp. 409–419, 2004.

[81] W. T. Miller Jr., "Chest radiographic evaluation of diffuse infiltrative lung disease: review of a dying art," *European Journal of Radiology*, vol. 44, no. 3, pp. 182–197, 2002.

[82] S. P. Padley, D. M. Hansell, C. D. Flower, and J. P., "Comparative accuracy of high resolution computed tomography and chest radiography in the diagnosis of chronic diffuse infiltrative lung disease," *Clinical Radiology*, vol. 44, no. 4, pp. 222–226, 1991.

[83] T. Ishida, S. Katsuragawa, T. Kobeyashi, H. MacMahon, and K. Doi, "Computerized analysis of interstitial disease in chest radiographs: improvement of geometric-pattern feature analysis," *Medical Physics*, vol. 24, no. 6, pp. 915–924, 1997.

[84] S. Katsuragawa, K. Doi, H. MacMahon, L. Monnier-Cholley, T. Ishida, and T. Kobayashi, "Classification of normal and abnormal lungs with interstitial diseases by rule-based method and artificial neural networks," *Journal of Digital Imaging*, vol. 10, no. 3, pp. 108–114, 1997.

[85] S. Kido, S. Tamura, N. Nakamura, and C. Kuroda, "Interstitial lung disease: evaluation of the performance of a computerized analysis systems versus observers," *Computerized Medical Imaging and Graphics*, vol. 23, pp. 103–110, 1999.

[86] T. Ishida, S. Katsuragawa, K. Nakamura, K. Ashizawa, H. MacMahon, and K. Doi, "Computerized analysis of interstitial lung diseases on chest radiographs based on lung texture, geometric pattern features and artificial neural networks," in *Proceedings of the SPIE*, vol. 4684, pp. 1331–1338, 2002.

[87] W. R. Webb, N. L. Müller, and D. P. Naidich, *High resolution CT of the lung*. Philidelphia, PA: Lippincott Williams & Wilkins, 3rd ed., 2001.

[88] I. C. Sluimer, M. Prokop, I. Hartmann, and B. van Ginneken, "Automated classification of hyperlucency, fibrosis, ground glass, solid and focal lesions in high resolution CT of the lung," *Medical Physics*, vol. 33, no. 7, pp. 2610–2620, 2006.

[89] 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*, vol. 45, no. 6, pp. 891–923, 1998.

[90] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.

[91] C. Schaefer-Prokop, M. Prokop, D. Fleischmann, and C. Herold, "High-resolution CT of diffuse interstitial lung disease: key findings in common disorders," *European Radiology*, vol. 11, pp. 373–392, 2001.

[92] K. Murphy, M. Prokop, C. M. Schaefer-Prokop, H. Gietema, G. D. Nossent, B. van Ginneken, J. P. W. Pluim, and Y. Arzhaeva, "Improved efficiency of assessment of interstitial lung disease progression in CT of the chest by visualisation of automatically-registered image pairs," in *Radiological Society of North America*, vol. 94th Annual Meeting, 2008.

[93] H. J. Kim, G. Li, D. Gjertson, R. Elashoff, S. K. Shah, R. Ochs, F. Vasunilashorn, F. Abtin, M. S. Brown, and J. G. Goldin, "Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study," *Academic Radiology*, vol. 15, no. 8, pp. 1004–1016, 2008.

[94] Y. Xu, E. J. R. van Beek, Y. Hwanjo, J. Guo, G. McLennan, and E. A. Hoffman, "Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM)," *Academic Radiology*, vol. 13, no. 8, pp. 969–978, 2006.

[95] V. A. Zavaletta, B. J. Bartholmai, and R. A. Robb, "High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis," *Academic Radiology*, vol. 14, no. 7, pp. 772–787, 2007.

[96] H. Sumikawa, T. Johkoh, S. Yamamoto, K. Takahei, T. Ueguchi, Y. Ogata, M. Matsumoto, Y. Fujita, J. Natsag, A. Inoue, M. Tsubamoto, N. Mihara, O. Honda, N. Tomiyama, S. Hamada, and H. Nakamura, "Quantitative analysis for computed tomography findings of various diffuse lung diseases using volume histogram analysis," *Journal of Computer Assisted Tomography*, vol. 30, no. 2, pp. 244–249, 2006.

[97] A. C. Best, J. Meng, A. M. Lynch, C. M. Bozic, D. Miller, G. K. Grunwald, and D. A. Lynch, "Idiopathic pulmonary fibrosis: Physiologic tests, quantitative CT indexes, and CT visual scores as predictors of mortality," *Radiology*, vol. 246, no. 3, pp. 935–940, 2008.

[98] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans Med Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[99] P. Thévenaz, T. Blu, and M. Unser, "Image interpolation and resampling," in *Handbook of Medical Imaging, Processing and Analysis*, pp. 393–420, San Diego, CA: Academic Press, 2000.

[100] S. Klein, M. Staring, and J. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," *IEEE Trans Image Process*, vol. 16, pp. 2879–2890, 2007.

[101] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer Jr., "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," *NeuroImage*, vol. 21, no. 4, pp. 1428–1442, 2004.

[102] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–26, 2006.

[103] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, pp. 1895–1923, 1998.

[104] N. J. Screaton, M. P. Hiorns, K. Lee, T. Franquet, T. Johkoh, K. Fujimoto, K. abd Ichikado, T. V. Colby, and N. L. Müller, "Serial high resolution CT in non-specific interstitial pneumonia: prognostic value of the initial pattern," *Clinical Radiology*, vol. 60, no. 1, pp. 96–104, 2005.