Unsupervised Discovery of Emphysema Subtypes in a Large Clinical Cohort

by

Polina Binder

B.S, Computer Science and Mathematics, 2013

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science
in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

June 2016

Signature of Author: _____

Polina Binder
Department of Electrical Engineering and Computer Science
THESIS SUBMISSION DATE

Certified by: _____

Polina Golland
Associate Professor of Elictrical Engineering and Computer Science
Thesis Supervisor

Accepted by: _____

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Committee for Graduate Students

Unsupervised Discovery of Emphysema Subtypes in a Large Clinical Cohort

by Polina Binder

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Science

Abstract

Emphysema is one of the hallmarks of Chronic Obstructive Pulmonary Disease (COPD), a devastating lung disease often caused by smoking. Emphysema appears on Computed Tomography (CT) scans as a variety of textures that correlate with the disease subtypes. It has been shown that the disease subtypes and the lung texture are linked to physiological indicators and prognosis, although neither is well characterized clinically. Most previous computational approaches to modeling emphysema imaging data have focused on supervised classification of lung textures in patches of CT scans. In this work, we describe a generative model that jointly captures heterogeneity of disease subtypes and of the patient population. We also derive a corresponding inference algorithm that simultaneously discovers disease subtypes and population structure in an unsupervised manner. This approach enables us to create image-based descriptors of emphysema beyond those that can be identified through manual labeling of currently defined phenotypes. By applying the resulting algorithm to a large data set, we identify groups of patients and disease subtypes that correlate with distinct physiological indicators.

Thesis Supervisor: Polina Golland
Title: Associate Professor

Contents

List of Figures

Chapter 1

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a chronic lung disease characterized by poor airflow. One of the hallmarks of COPD is emphysema, (i.e., destruction of structures supporting lung alveoli and permanent enlargement of airspaces) [9]. Several subtypes of emphysema have been identified by radiologists. Patients with emphysema exhibit a mixture of disease subtypes. This aspect of emphysema differentiates it from most other diseases, in which patients only exhibit a single disease subtype. These subtypes are used for diagnosis and predicting patient prognosis [23]. The subtypes have also been shown to correlate with genetic data and biological markers [21]. Emphysema manifests on Computed Tomography (CT) scans as a variety of textures, which are associated with clinically defined emphysema disease subtypes. Figure 1.1 illustrates normal lung tissue, along with patches of several clinically defined emphysema subtypes.

There is substantial intra-reader and inter-reader variability when identifying subtypes in CT images [27]. Computational approaches to the classification of textures in CT scans promise to identify subtle textural differences beyond those that are visible

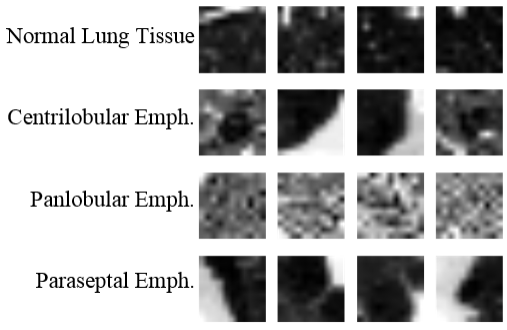


Figure 1.1: Image patches showing clinically defined emphysema subtypes

to human readers. This nuanced information can be harnessed to produce well-defined, reproducible disease subtypes. Beyond fully 3D texture analysis, the additional benefits of computational approaches include the possibility of providing novel insights into the disease once the heterogeneity of the patient population is properly characterized.

Our approach departs from the majority of prior research that has focused on supervised classification of patches extracted from CT scans based on examples labelled by clinical experts [4, 17, 18]. A notable exception is a recently demonstrated method for joint modelling of imaging and genetic data in the same clinical population [1]. Our work models only the imaging data, but we explicitly detect and characterize homogeneous sub-populations defined based on the phenotypic similarities, which opens interesting directions for future analysis.

1.1 Contributions

In this thesis, we address the challenge of modelling heterogeneity in the disease subtypes and in the patient population in the context of an unusually large medical imaging data set consisting of 2457 thoracic CT scans.

Our primary contribution is a method that simultaneously detects distinct patient clusters and disease subtypes. The algorithm is based on a generative model that captures the underlying assumptions about population structure and distributions of disease subtypes. Specifically, we assume that each cluster of patients is associated with a distinct distribution of disease subtypes. We derive an inference algorithm that is based on variational Expectation-Maximization [2]. We apply the algorithm to our data set and observe notable associations between physiological indicators and patient clusters and disease subtypes identified by the method. Further, we examine associations in simplified models that omit either patient clusters or disease subtypes to demonstrate the clinical advantage of the fully hierarchical model that includes both patient clusters and disease subtypes.

We also examine the choice of an appropriate texture descriptor that is used to differentiate textures in the scans that appear in our data set. We choose these texture descriptors based on their classification accuracy on a labeled portion of our data set. These descriptors serve as the observed data in our generative model.

Figure 1.2 shows CT scans that belong to different patient clusters identified by the algorithm presented in this thesis. The colors overlaying the lungs correspond to disease subtypes identified by our algorithm. Each of the lungs exhibits a mixture of disease subtypes.

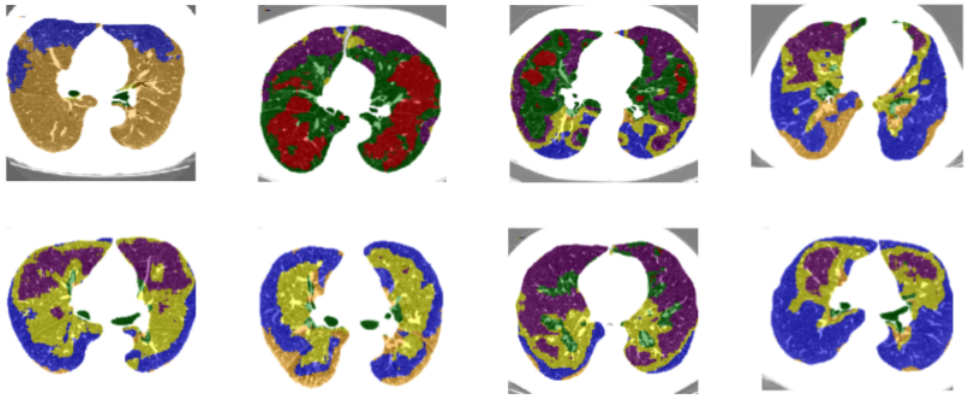


Figure 1.2: Example CT scans from each of the eight patient clusters identified by our algorithm. Colors correspond to disease subtypes identified by our algorithm.

■ 1.2 Thesis Outline

This thesis is organized as follows. In Chapter 2 we review background relevant to the method development in this thesis, and we also place the proposed methods in the context of previous work. In Chapter 3 we discuss our choice of texture descriptors. In Chapter 4, we describe a generative model that we employed to identify disease subtypes and patient clusters. In Chapter 5, we present the data and the empirical evaluation procedure and discuss the experimental results. In the last chapter, we summarize and examine directions for future work.

Chapter 2

Background and Previous Work

In this section, we describe clinical background relevant to the understanding of emphysema and COPD. We discuss the basics of CT imaging, and methods for texture classification in medical image analysis. We then place this work in the context of previous medical imaging research that aimed to classify CT scans of patients with COPD and related diseases. To the best of our knowledge, ours is the first study that has successfully identified emphysema subtypes in a fully unsupervised manner.

2.1 Background on Emphysema and COPD

COPD is the third leading cause of death in the United States, affecting approximately 15 million people each year [11]. It is a highly heterogeneous disease. The disease's subtypes and causality are not well characterized [12]. Except for smoking, the risk factors associated with COPD and those influencing its prognosis are poorly understood [21]. A few genetic variants that correlate with COPD risk have recently been identified, along with certain environmental factors [12]. Currently, COPD is diagnosed based on a ratio of volume of air that can be exhaled in one second and the total amount of air that can be exhaled in one breath. If the ratio is less than 70%, COPD diagnosis is established [19]. Biologically, COPD manifests as a combination of chronic bronchitis and emphysema. To distinguish between and within these contributions to COPD, radiological characterizations, generally based on CT scans, are employed [15]. Emphysema presents as various patterns of physical lung tissue destruction, which can be observed as texture in CT scans. It has been shown that texture patterns found in CT scans correlate strongly with histopathological findings [16].

Three common emphysema subtypes have been established in the medical practice: centrilobular, panlobular, and paraseptal emphysema. Further, radiologists may utilize a variety of terminology including "honeycombing" and "ground-glass texture" to describe patterns of lung destruction seen in emphysema. A patient may exhibit a

combination of these subtypes and textures to varying degrees, along with healthy lung tissue [9]. Emphysema subtypes have been shown to strongly correlate with clinical prognosis [23]. However, there are no uniform clinical, pathological, or texture-based standards for identifying these subtypes or textures. This also leads to high degrees of intra-reader and inter-reader variability when interpreting CT scans [27]. Additionally the emphysema textures are inherently three-dimensional, so humans cannot fully visualize them. Improved understanding of emphysema subtypes would not only improve the biological understanding of the disease, but also enable better tailored treatments and more accurate prognosis. Moreover, it promises to help classify the subtypes of the disease as linked to genetic components or environmental factors.

2.2 CT Imaging

CT imaging is used for diagnosis and imaging of structural changes in organs including the brain, lungs, heart, extremities, and abdomen [5]. It has been an important diagnostic tool for emphysema and COPD for two decades [16]. CT imaging is a non-invasive imaging technique that uses X-rays to produce virtual slices, or tomographs of a given scanned object. These are processed to produce a three-dimensional representation of the scanned area [17].

Texture is observed in CT scans as spatial intensity variation in the image, created when X-rays are scattered by tissues with varying physical properties [17]. Although the texture is created by different underlying physical structures, in this work we will not attempt to reconstruct the underlying physical properties of the tissue. Instead, we will analyze the texture features that are extracted from CT scans, and employ these as markers to differentiate the underlying tissues.

2.3 Texture Definition

There is no single definition of texture. It is generally understood as the spatial distribution of voxel or pixel intensity in an area of interest. Three dimensional textures exist in filled objects and are generated by volumetric data acquisition devices. Three dimensional textures cannot be characterized in terms of reflectivity and surface properties, but instead represent volumetric properties of the materials or tissues. Additionally, three-dimensional textures cannot be fully visualized by humans, so it is inherently only possible to model them algorithmically [5].

2.4 Texture Descriptors

Here we survey texture descriptors that have been used to model lung textures and which we will employ in the proposed work. Each of these descriptors is defined for a patch centered around a particular voxel. Several important properties differentiate among these texture descriptors, including sensitivity to the underlying parameters and rotational invariance [5]. In this work, we utilize the first three texture descriptors.

2.4.1 Histograms

Histograms describe the discretized distribution of intensities within a patch. Mendoza et al. [14] employed histogram texture descriptors along with kernel density estimation to perform supervised classification of emphysema subtypes, demonstrating superior performance to that of many commonly used complex descriptors. Histograms are rotationally invariant but are sensitive to the patch and bin size [5]. It is necessary to empirically determine the values of the bin size.

2.4.2 Grey Level Co-Occurrence Matrices

Grey Level Co-Occurrence Matrices (GLCMs) represent the joint probability distribution of intensity values of pixel pairs in a given patch [18]. To construct this descriptor, the image is discretized into a given number of grey levels, often eight or 16. Pixel pairs are examined at a given offset. Generally, the distance is set to be fairly small (between one and three voxels). The value of the entry at position (i, j) in the GLCM captures the proportion of pixel pairs at the offset where one voxel has intensity i, and the other has an intensity j. This descriptor effectively extends histograms to pairwise marginal distributions.

A common approach to obtain a degree of rotational invariance is to average the GLCMs over some number of uniformly distributed directions in three dimensions. The feature vector corresponding to a given voxel is a collection of features that can be extracted from GLCMs - including the entropy, maximal probability, homogeneity, and others [10]. This descriptor is sensitive to patch size, number of levels, and offset used to compute the histogram.

2.4.3 Fourier Analysis and Discrete Cosine Transformation

Fourier transforms are equivalent to convolution of the patch with sine and cosine functions. They are defined over functions with infinite support. To obtain a local texture representation, the basis functions are typically bounded to a given region of

interest, and the boundary conditions are specified. The discrete cosine transform uses only the real coefficients. The feature vector is generally constructed from the largest coefficients. These descriptors can be modified to be rotationally invariant [5].

2.4.4 Difference of Gaussians and Gabor Filters

Let G_σ be the Gaussian kernel with standard deviation σ. Radially symmetric receptive fields correspond to the difference of Gaussians and are modeled by $F_{rad} = G_{\sigma_1} - G_{\sigma_2}$. These are rotationally invariant. The two-dimensional Gabor filter bank can be extended to three dimensions, and maintain rotational invariance. These filter banks are constructed across various octaves to cover the scale space of the scan or patch, and so are generally not highly sensitive to underlying parameters. The feature vector of the voxel can be defined as the convolution of the patch with the filter bank at a the voxel of interest [13]. Alternatively, the feature vector can be a histogram defined over the convolution of the filter bank with the patch [17].

2.4.5 Riesz Features and Wavelets

Riesz features were specifically proposed by Depeursinge et al. [4] for lung texture classification. They are wavelets, which are filter banks that cover the entire spatial spectrum of the image. The descriptor was originally defined for two dimensions, but can be extended to 3D. The Riesz transform maps a function to its harmonic conjugate, and can be thought of as a generalization of the Hilbert transform for Euclidean spaces of dimension greater than one. Riesz transforms are convolved with the Laplacian of a Gaussian at various scales to obtain rotationally-covariant basis functions. These convolutions create a steerable filter bank, which makes it possible to analytically obtain the filter coefficients at any orientation as a linear combination of basis filters. This allows for the orientation of the filters at each voxel in such a way that they produce a maximal response [22]. The feature vector for a voxel can be defined as the convolution of the patch with the filter bank at a voxel, or as the energy of this convolution [17].

2.5 Previous Work on CT Classification

Here, we survey previous research that aimed to classify CT scans of patients with COPD and related diseases.

Comparing classifications across prior methods is challenging for a number of reasons. Most previous work employed distinct clinical patient cohorts, largely due to the

fact that few public sets of COPD scans are available. These cohorts have contained from 18 [25] to 342 CT scans [14]. Additionally, a large part of the related work is based on patient cohorts affected by diseases related to but not COPD. Most previous work has also employed the results from a single scanner or scanning protocol, which limits transfer to new clinical cohorts as textures may manifest differently with varying scanners and protocols. There has been some recent work [14] on newly available multi-site patient cohorts [21, 23]. In previous work, 2D and 3D neighborhoods of varying sizes were used for feature extraction to train classifiers. Typically, these were square or cubic patches of a fixed size [14]; however, manually annotated regions of variable shape have also been employed [17]. Additionally different types of class labels have been investigated across different studies. Some studies focused solely on identifying emphysema subtypes [6, 14, 24], while others treated emphysema as a single texture class among other textures [4, 17, 18]. Additionally, the emphysema subtypes and other textures present in images were not defined consistently across different studies [17].

The majority of previous work focused on supervised learning for identifying clinically defined emphysema subtypes, generally by classifying image patches. A broad variety of modeling approaches have been employed, including Random Forests [17], SVMs [3], and K-Nearest Neighbors [14]. Additionally, classification of lung disease subtypes has been demonstrated for content-based image retrieval, which seeks to retrieve earlier images similar to the input example [20].

A similar approach to ours was proposed by Dy et al. [6]. This work introduces a partially supervised approach for lung texture classification, within the framework of content-based retrieval. The authors used a collection of thoracic CT scans of patients suffering from a variety of diseases related to and including emphysema. However, they employed only two-dimensional regions for characterizing lung texture. Additionally, they used supervised approaches to distinguish between emphysema subtypes, and then perform unsupervised classification within these subtypes, which prevents discovery of truly novel subtypes.

The most similar work to ours was proposed by Batmanghelich et al. [1]. This work constructed a generative model that discovered disease subtypes based on imaging and genetic data. In contrast, we discover emphysema subtypes in a strictly unsupervised manner, by modelling both the heterogeneity of our patient population and the distribution of emphysema subtypes within groups of patients, based only on imaging data.

Chapter 3

Choice of Texture Descriptors

In this section we discuss our data set and choice of texture descriptors. A variety of texture descriptors are described in the previous chapter, some of which we analyze in this section. As discussed in Section 2.5, COPD data sets differ greatly across their size and choice of regions in which to classify and identify textures. Thus we must identify the specific texture descriptors that are suitable to our cohort and problem. We employ a supervised approach for feature selection, but not for training the generative model discussed later in this thesis.

3.1 Data

We will investigate the proposed methods in the context of an imaging study that includes CT scans of 2457 patients' lungs. COPDGene is a multicenter study that acquired CT scans, genetic data, and physiological indicators such as spirometry measures, six-minute walking distance, height, weight, and blood pressure in COPD patients who are smokers [21]. The study's goal is to understand COPD subtypes, pathology, and genetics. The data was collected by 21 sites across the United States, using different CT scanners. The volumetric CT scans were obtained at full inhalation and at relaxed exhalation. Image reconstruction produces sub-millimeter slice thickness, and employs edge and smoothness enhancing filtering [21]. The images are then resampled to obtain $1.5mm$ slice thickness. In addition, we have 1525 patches from the CT scans of 267 patients from this cohort which were manually labeled by a clinical expert [14]. The data was made available to us by our collaborators at Brigham and Women's Hospital. This is an unusually large patient cohort, which promises to provide new, powerful insights into the effects of emphysema and COPD on lungs.

3.2 Identifying Texture Descriptors

For each voxel in the image we seek to construct a feature vector whose entries correspond to values of texture features extracted from a volumetric patch around the voxel. Emphysema has been described at the level of the secondary pulmonary lobule [14], therefore we selected patches large enough to encapsulate an entire secondary lobule, but not too large as to blur the boundaries between regions. We chose to utilize 11x11x11 patches around each voxel. On our CT scans, these correspond to patches of size approximately $24 \times 24 \times 24mm^3$, which is the approximate size of secondary pulmonary lobules.

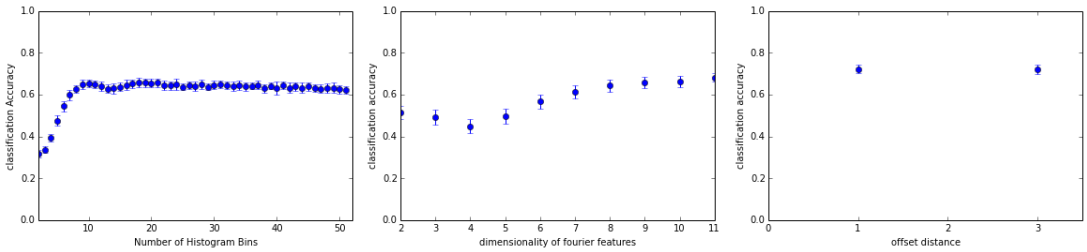
We choose the appropriate texture descriptors by examining their accuracy in classifying patches that have been labeled by clinicians. Although a big motivator for our unsupervised algorithm is that we wish to discover structure beyond that which is available from clinician's labels, they still contain a degree of information that can be harnessed to select the proper texture descriptors. For the feature selection, we used the 1525 labeled patches from the CT scans of 267 patients. Each of these patches was identified by one of four labels: centrilobular emphysema, panlobular emphysema, paraseptal emphysema, and normal lung tissue.

To evaluate the classification accuracy, we performed repeated random sub-sampling [2] 100 times on a balanced portion of the data set which was split in half each time between testing and training data. We then trained a Support Vector Machine (SVM) classifier on the training set and evaluated its accuracy on the testing set within each split.

We examined three types of texture descriptors: histograms, the discrete cosine transform, and GLCMs, which are described in Section 2.4. Additionally, after a primary exploration of the data we found that the vertical distance from the top of the lung (normalized by lung size) correlates with the emphysema subtype, so we also experimented with appending this value to the feature vector.

Initially, we examined the optimal number of bins for classification with histograms. As shown in Figure 3.1, feature vectors consisting of 10 bins lead to as high a classification accuracy as those with higher bin counts. These feature vectors produce a classification accuracy of 0.657.

We then applied the discrete cosine transform to the image patches and repeated the classification experiment. As shown in Figure 3.1, we obtain a maximal classification accuracy of 0.675 when using the first 11 Fourier coefficients in each direction. However, this leads to an 11^3, or 1331-dimensional feature vector. Further, we only obtain



(a) Accuracy with Histograms (b) Accuracy with the Discrete Cosine Transform (c) Accuracy with GLCMs

Figure 3.1: Comparison of classification accuracies for different feature descriptors.

GLCM Feature	Formula
Energy	$\sum_i \sum_j M_{ij}^2$
Contrast	$\sum_i \sum_j (i-j)^2 M_{ij}$
Correlation	$\frac{1}{\sigma_x \sigma_y} \sum_i \sum_j ij \cdot M_{ij} - \mu_x \mu_y$
Maximal Probability	$\max\{M_{ij}\}$
Dissimilarity	$\sum_i \sum_j \lvert i-j \rvert M_{ij}$
Local Homogeneity	$\sum_i \sum_j \frac{1}{1+(i)} M_{ij}$
Entropy	$-\sum_i \sum_j M_{ij} \log(M_{ij})$
Cluster Shade	$\sum_i \sum_j (i+j-\mu_x-\mu_y)^3 \cdot M_{ij}$
Cluster Prominence	$\sum_i \sum_j (i+j-\mu_x-\mu_y)^4 \cdot M_{ij}$

Table 3.1: Table describing features that we extracted from GLCMs.

accuracy comparable to that of the histogram feature vectors when using the first 8 Fourier coefficients, which corresponds to a 512-dimensional feature vector.

We then repeated the classification experiment with feature vectors that consist of features extracted from GLCMs. To construct the GLCMs, we first discretized the patches into eight image intensity levels. We examined rotationally invariant features, since lung texture features do not appear to exhibit a direction. These were produced by summing the GLCMs over uniformly distributed directions in three dimensions and extracting features from this new matrix. We examined offsets of distance one and three and used nine common feature descriptors. These are listed in Table 3.1. In this table, M_{ij} corresponds to the entry in row i and column j in the GLCM. Additionally, $\mu_x = \sum_i i \sum_j M_{ij}$ and $\mu_y = \sum_j j \sum_i M_{ij}$. Similarly $\sigma_x^2 = \sum_i (i-\mu_x)^2 \sum_j M_{ij}$ and $\sigma_y^2 = \sum_j (j-\mu_j)^2 \sum_i M_{ij}$.

We found that we obtained a classification accuracy of 0.721 when using an offset distance of 1, and an accuracy 0.719 with an offset distance of 3, as shown in Figure 3.1.

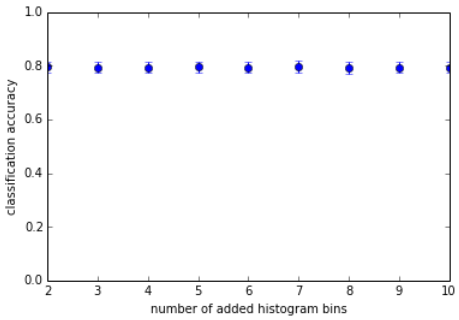


Figure 3.2: Comparison of classification accuracies of GLCMs with a variable number of histogram bins appended.

These three types of texture descriptors - histogram, the discrete cosine transform, and GLCMs capture different aspects of texture within the patch. Histograms capture the intensity distribution, while Fourier components capture the strengths of various frequencies within a patch. GLCMs capture patterns of intensity variation between the voxels.

It is slightly surprising that texture descriptors as simple as histograms prove so accurate at differentiating the emphysema subtypes. Previous work [14] has demonstrated that histograms are more accurate than several more complex approaches at predicting emphysema subtypes. A possible explanation for the good performance of histograms is that image intensities account for a large fraction of differences in emphysema subtypes.

We then experimented with combining GLCMs and histograms to produce texture descriptors. The motivation behind this is that both feature descriptors produce low-dimensional representations and each captures different aspects of the texture. The feature vectors were constructed by appending various numbers of histogram bins to a feature vector consisting of GLCM descriptors. The classification accuracy is shown in Figure 3.2. As can be seen in the Figure, there is very little improvement from appending more than 2 histogram bins. With such feature vectors, we achieve a classification accuracy of 0.792. We also appended the distance from the top of the lung to our feature vector, but the classification results remained virtually identical.

Thus our feature vectors are 11-dimensional, where the first nine values correspond to GLCM features, and the next two values correspond to histogram bins from the patch around the voxel. This combination of descriptors captures different aspects of texture, which creates powerful feature vectors.

Chapter 4

Generative Model

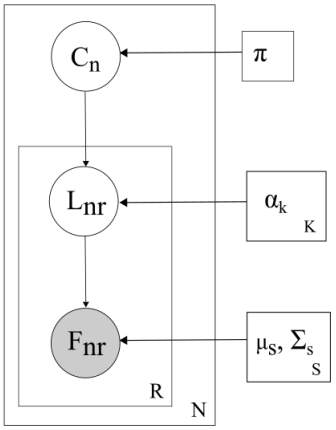
In this chapter, we present a probabilistic generative model that captures assumptions about the population structure of our cohort. We then derive a corresponding inference algorithm. The generative model assumes that each underlying patient cluster shares a common distribution of disease subtypes. This is an assumption supported by the clinical understanding that different disease subtypes and combinations of subtypes correlate with distinct clinical prognoses [23]. The evaluation of the identified patient clusters and disease subtypes will be described in Chapter 5.

4.1 Formulation

Our generative model relies on the assumption that there are K underlying patient clusters, each characterized by a different distribution of disease subtypes. We use N to denote the total number of CT scans in the study. When processed, each scan is represented by R non-overlapping patches. Let S_{nr} be the patch around voxel r in patient n. Patches are entirely contained within a lung. We apply a chosen feature extraction method to S_{nr} to construct a feature vector $F_{nr} \in \mathbf{R}^d$. The feature vectors $\{F_{nr}\}$ serve as the input into our algorithm. The images are not spatially aligned, as it is challenging to find spatial correspondences between lungs of different individuals [15]. In the experiments presented in the next chapter of this thesis, we use a combination of Grey Level Co-Occurrence Matrix (GLCM) [18] features and intensity histograms as feature descriptors; the modeling approach readily accepts a broad range of descriptors.

The full generative model and a summary table of the parameters and variables is shown in Figure 4.1.

The distribution of cluster assignments for any patient in the study is parameterized by π and is represented by a vector C_n for patient n. $C_{nk} = 1$ if patient n belongs to cluster k; $C_{nk} = 0$ otherwise. For all patients in cluster k the distribution of disease subtypes is parameterized by α_k and is represented by L_{nr} for patch r in patient



Parameter/Variable	Description
N	Number of patients
S	Number of subtypes
R	Number of regions in each patient
K	Number of patient clusters
π	Frequency of each patient cluster
α_k	Frequencies of emphysema subtypes in cluster k
μ_s	Mean of subtype s
Σ_s	Variance of subtype s
C_n	Cluster label of patient n
L_{nr}	Subtype of patch r in patient n
F_{nr}	Feature vector of patch r in patient n

Figure 4.1: Graphical representation and summary of variables and parameters of the generative model.

n. Each patch belongs to one of S disease subtypes. $L_{nrs} = 1$ if the patch belongs to subtype s; $L_{nrs} = 0$ otherwise. We use a Gaussian distribution $\mathcal{N}(\cdot; \mu, \Sigma)$ with mean μ_s and covariance Σ_s to model feature vectors in the disease subtype s.

The generative model can be summarized as follows:

$$C_n \sim \prod_{k=1}^{K} \pi_k^{C_{nk}}, \tag{4.1}$$

$$L_n|C_n \sim \prod_{k=1}^{K} \prod_{r=1}^{R} \prod_{s=1}^{S} (\alpha_{ks})^{L_{nrs}C_{nk}}, \tag{4.2}$$

$$F_n|L_n \sim \prod_{s=1}^{R} \prod_{r=1}^{S} \mathcal{N}(F_{nr}; \mu_s, \Sigma_s)^{L_{nrs}}. \tag{4.3}$$

Each subject is viewed as an independent and identically distributed sample from this distribution, giving rise to the full likelihood model:

$$p(F, C, L; \alpha, \pi, \mu, \Sigma) = \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{r=1}^{R} \prod_{s=1}^{S} \left(\pi_k {\alpha_{ks}}^{L_{nrs}}\right)^{C_{nk}} \mathcal{N}(F_{nr}; \mu_s, \Sigma_s)^{L_{nrs}}. \tag{4.4}$$

We set the number of patient clusters K and the number of disease subtypes S. The observed data consists of feature vectors $\{F_{nr}\}$ of N patients for whom we extracted features from R patches each. We aim to infer the most likely subtype L_{nr} for each patch r in patient n and the most likely cluster C_n for each patient n. Additionally, we estimate the parameters: the mixing proportions of the patient clusters π, the mixing proportions of the disease subtypes $\{\alpha_k\}$ for each patient cluster, and the means and variances $\{\mu_s, \Sigma_s\}$ of the image features for each disease subtype.

■ 4.2 Inference with the Expectation-Maximization Algorithm

We perform inference on the model via an algorithm based on the variational Expectation-Maximization (EM) algorithm [2], which approximates the exact EM algorithm. In the exact EM algorithm, we seek to maximize the marginal log-likelihood $\ln p(F; \alpha, \pi, \mu, \Sigma)$ over the observed variables by iterative coordinate ascent [2]. To describe the exact EM algorithm, we re-write the marginal log-likelihood $\ln p(F; \alpha, \pi, \mu, \Sigma)$ by choosing an arbitrary distribution q over the latent variables C and L. We then obtain:

$$
\begin{aligned}
&\ln p(F;\alpha,\pi,\mu,\Sigma) \\
&= E_q\left[\ln p(F,C,L;\alpha,\pi,\mu,\Sigma)\right] - E_q\left[\ln p(C,L|F;\alpha,\pi,\mu,\Sigma)\right] \\
&= E_q\left[\ln \frac{p(F,C,L;\alpha,\pi,\mu,\Sigma)}{q(C,L)}\right] - E_q\left[\ln \frac{p(C,L|F;\alpha,\pi,\mu,\Sigma)}{q(C,L)}\right] \\
&= L_q(F,C,L;\alpha,\pi,\mu,\Sigma) + \mathrm{KL}(q(C,L)\|p(C,L|F;\alpha,\pi,\mu,\Sigma)),
\end{aligned} \tag{4.5}
$$

where

$$
\begin{aligned}
&L_q(F,C,L;\alpha,\pi,\mu,\Sigma) = E_q\left[\ln \frac{p(F,C,L;\alpha,\pi,\mu,\Sigma)}{q(C,L)}\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \ln \pi_k E_q[C_{nk}|F,\alpha,\pi,\mu,\Sigma] + \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{s=1}^{S} \ln \alpha_{ks} E_q[L_{nrs}C_{nk}|F,\alpha,\pi,\mu,\Sigma] \\
&+ \sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{s=1}^{S} E_q[L_{nrs}|F,\alpha,\pi,\mu,\Sigma] \cdot \ln \mathcal{N}(F_{nr};\mu_s,\Sigma_s) + H(q(C,L)),
\end{aligned} \tag{4.6}
$$

where $H(q(C,L))$ is the entropy of $q(C,L)$, and

$$
\mathrm{KL}(q(C,L\|p(C,L|F;\alpha,\pi,\mu,\Sigma)) = E_q\left[\ln \frac{p(C,L|F;\alpha,\pi,\mu,\Sigma)}{q(C,L)}\right] \tag{4.7}
$$

is the Kullback-Liebler (KL) divergence between $q(C,L)$ and $p(C,L|F;\alpha,\pi,\mu,\Sigma)$. Since the KL-divergence is non-negative, $L_q(F,C,L;\alpha,\pi,\mu,\Sigma)$ is a lower bound for $\ln p(F;\alpha,\pi,\mu,\Sigma)$.

The exact EM algorithm then iteratively maximizes $\ln p(F;\alpha,\pi,\mu,\Sigma)$ by maximizing the lower bound $L_q(F,C,L;\alpha,\pi,\mu,\Sigma)$. In this algorithm, we randomly initialize the parameters α, π, μ, and Σ. Then the algorithm iterates between two steps until convergence criteria are met: the expectation step (E-step), and the maximization step (M-step).

In the E-step, we hold the model parameters fixed and find the parameters of the approximating distribution, q, that maximize $L_q(F,C,L;\alpha,\pi,\mu,\Sigma)$. We then compute the values of expectations seen in equation 4.6 for the current estimates.

The KL-divergence is non-negative, so we can see by inspecting equation 4.5 that

$L_q(F, C, L; \alpha, \pi, \mu, \Sigma)$ will be maximized when

$$\mathrm{KL}(q(C, L) \| p(C, L|F; \alpha, \pi, \mu, \Sigma) = 0. \tag{4.8}$$

For this to hold, we must set

$$q(C, L) = p(L, C|F, \alpha, \pi, \mu, \Sigma), \tag{4.9}$$

the full posterior distribution. At the end of the E-step,

$$L_q(F, C, L, \alpha, \pi, \mu, \Sigma) = \ln p(F, \alpha, \pi, \mu, \Sigma). \tag{4.10}$$

In the M-step, we hold the parameters of $q(C, L)$ fixed and maximize $L_q(F, C, L; \alpha, \pi, \mu, \Sigma)$ with respect to model parameters α, π, μ, and Σ in $L_q(F, C, L; \alpha, \pi, \mu, \Sigma)$. The values of the expectations evaluated in the E-step are necessary to perform these calculations. Maximizing the lower bound in the M-step causes the marginal log-likelihood of the data to increase at every step.

■ 4.3 Variational Expectation Maximization

In the exact EM algorithm presented above we must compute the expectations in equation 4.6 with respect to the full posterior distribution in the E-step. This is intractable due to coupling between the latent variables C and L. Thus, we employ a variational EM algorithm [2]. The difference from the exact EM algorithm is that we constrain the distribution $q(C, L)$ in a way that will simplify our derivations in the E-step [2]. We choose $q(C, L)$ to approximate the full posterior distribution with a product of two categorical distributions:

$$q(C, L; \psi, \theta) = q_C(C; \psi) \cdot q_L(L; \theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \psi_{nk}^{C_{nk}} \prod_{r=1}^{R} \prod_{s=1}^{S} \theta_{nrs}^{L_{nrs}}, \tag{4.11}$$

where ψ and θ are variational parameters. In this case the expectations in the E-step become:

$$\begin{aligned} E_{q(C,L;\psi,\theta)}[L_{nrs}C_{nk}|F, \alpha, \pi, \mu, \Sigma]] &= \psi_{nk} \cdot \theta_{nrs}, \\ E_{q(C,L;\psi,\theta)}[C_{nk}|F, \alpha, \pi, \mu, \Sigma] &= \psi_{nk}, \\ E_{q(C,L;\psi,\theta)}[L_{nrs}|F, \alpha, \pi, \mu, \Sigma] &= \theta_{nrs}. \end{aligned} \tag{4.12}$$

This enables us to re-write the lower bound $L_q(F, C, L; \alpha, \pi, \mu, \Sigma)$ in equation 4.6 as follows:

$$
\begin{aligned}
&L_q(F, C, L; \alpha, \pi, \mu, \Sigma) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \ln \pi_k E_{q(C)}[C_{nk}] + \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{s=1}^{S} \ln \alpha_{ks} E_{q(L)}[L_{nrs}] \cdot E_{q(C)}[C_{nk}] \\
&+ \sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{s=1}^{S} E_{q(L)}[L_{nrs}] \cdot \ln \mathcal{N}(F_{nr}; \mu_s, \Sigma_s) + H(q_C(C; \psi)) + H(q_L(L; \theta)) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \ln(\pi_k)\psi_{nk} + \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{s=1}^{S} \ln \alpha_{ks} \psi_{nk} \cdot \theta_{nrs} \\
&- \sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{s=1}^{S} \frac{\theta_{nrs}}{2} \cdot \left(d \cdot \ln(2\pi) + \ln|\Sigma_s| + (F_{nr} - \mu_s)^T \Sigma_s^{-1} (F_{nr} - \mu_s)^T\right) \\
&- \sum_{n=1}^{N}\sum_{k=1}^{K} \psi_{nk} \ln \psi_{nk} - \sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{s=1}^{S} \theta_{nrs} \ln \theta_{nrs} \\
&= L^{var}_{(C,L;\psi,\theta)}(q(F, C, L; \alpha, \pi, \mu, \Sigma)).
\end{aligned} \tag{4.13}
$$

In the variational algorithm, we iteratively optimize this variational lower bound for $\ln p(F; \alpha, \pi, \mu, \Sigma)$ with respect to the parameters $\{\pi_k, \alpha_{ks}, \mu_s, \Sigma_s, \psi_{nk}, \theta_{nrs}\}$. We randomly initialize the parameters and then iterate between the E-step and M-step until convergence.

In the E-step, we hold the model parameters π, α, μ, and Σ fixed and estimate the variational parameters ψ and θ to maximize the lower bound in equation 4.13. Unlike the exact EM algorithm, we can no longer find $q(C, L)$ such that $KL(q(C, L; \theta, \psi) \| p(C, L|F; \alpha, \pi, \mu, \Sigma)) = 0$, so the lower bound is no longer equal to the marginal log-likelihood at the end of the E-step. The M-step proceeds as in the exact EM-case.

Once the parameter estimation process is complete, we determine the cluster labels C_n and the disease subtype labels L_{nr} by maximizing the approximate posterior distributions $q_C(C_n; \psi_n)$ and $q_L(L_{nr}; \theta_{nr})$ respectively.

This algorithm is highly similar to the EM algorithm described in the previous section. However, in the variational algorithm, we are maximizing the lower bound $L^{var}_{q(C,L;\psi,\theta)}(F, C, L; \alpha, \pi, \mu, \Sigma)$ in equation (4.13), instead of the general lower bound $L_q(F, C, L; \alpha, \pi, \mu, \Sigma)$. At the end of every E-step in the exact EM algorithm the lower bound equals $\ln p(F, \alpha, \pi, \mu, \Sigma)$, which is not true in the variational algorithm,

Algorithm 1: Variational EM Algorithm for Patch and Patient Classification

1. Select R patches from each of N patients to obtain $\{S_{nr}\}$.
2. Extract features from each patch, $F_{nr} = F(S_{nr})$.
3. Randomly initialize parameters $\alpha,\ \pi,\ \mu$ and Σ.
4. **E-Step**: Determine $\theta^*, \psi^* = \text{argmax}_{\theta,\psi}\{L^{var}_{q(\theta,\psi)}(F, C, L; \alpha, \pi, \mu, \Sigma)\}$:

$$\psi_{nk} \propto \prod_{s=1}^{S}\prod_{r=1}^{R} \alpha_{ks}^{\theta_{nrs}}, \quad \text{s.t.} \quad \sum_{k=1}^{K} \psi_{nk} = 1,$$
$$\theta_{nrs} \propto \prod_{k=1}^{K} \alpha_{ks}^{\psi_{nk}}, \quad \text{s.t.} \quad \sum_{s=1}^{S} \theta_{nrs} = 1.$$

5. **M-step**: Determine
$\alpha^*, \pi^*, \mu^*, \Sigma^* = \text{argmax}_{\alpha,\pi,\mu,\Sigma}\{L^{var}_{q(C,L;\theta,\psi)}(F, C, L; \alpha, \pi, \mu, \Sigma)\}$:

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N} \psi_{nk},$$
$$\alpha_{ks} \propto \sum_{n=1}^{N} \psi_{nk} \sum_{r=1}^{R} \theta_{nrs}, \quad \text{s.t.} \quad \sum_{s=1}^{S} \alpha_{ks} = 1,$$
$$\mu_s = \frac{1}{N_s}\sum_{n=1}^{N}\sum_{r=1}^{R} \theta_{nrs} \cdot F_{nr}, \quad \text{where} \quad N_s = \sum_{n=1}^{N}\sum_{r=1}^{R} \theta_{nrs},$$
$$\Sigma_s = \frac{1}{N_s}\sum_{n=1}^{N}\sum_{r=1}^{R} \theta_{nrs} \cdot (F_{nr} - \mu_s) \cdot (F_{nr} - \mu_s)^T.$$

6. Repeat steps 2 or 3 until convergence criteria are met.
7. For each (n, k), set $C_{nk} = \begin{cases} 1 & \text{if } \psi_{nk} = \max_k\{q_C(C_{nk}; \psi_k)\} \\ 0 & \text{otherwise} \end{cases}$
8. For each (n, r, s), set $L_{nrs} = \begin{cases} 1 & \text{if } \theta_{nrs} = \max_s\{q_L(L_{nrs}; \theta_{nr})\} \\ 0 & \text{otherwise} \end{cases}$

so the variational algorithm is not guaranteed to maximize the log-likelihood at every step. However, the variational approximation enables us to implement our algorithm. In practice, these types of variational algorithms are highly effective and converge to good results.

■ 4.4 Deriving the E-Step

In the E-step, we keep the parameters of the full likelihood model fixed and seek to calculate the parameters of q that maximize $L^{var}_{q(C,L;\psi,\theta)}(F, C, L; \alpha, \pi, \mu, \Sigma)$:

We find

$$\theta^*, \psi^* = \text{argmax}_{\theta,\psi}\{L^{var}_{q(C,L;\psi,\theta)}(F, C, L; \alpha, \pi, \mu, \Sigma)\}. \tag{4.14}$$

Equation 4.14 remains challenging to optimize simultaneously with respect to both θ and ψ, since for a given value of n, ψ_{nk} and θ_{nrs} are coupled. Instead, we iteratively optimize $L^{var}_{q(C,L;\psi,\theta)}(F, C, L; \alpha, \pi, \mu, \Sigma)$ with respect to the ψ and the θ parameters separately. Once we hold the θ parameters fixed, the ψ parameters are decoupled, so we maximize each value of ψ_{nk} and θ_{nrs} independently.

With respect to a given ψ_{nk}, the expectation is convex. We can find the maximum by taking the derivative of $L^{var}_{q(C,L;\psi,\theta)}(F, C, L; \alpha, \pi, \mu, \Sigma)$ with respect to ψ_{nk} and setting it to 0. We have that for all n, $\sum_{k=1}^{K} \psi_{nk} = 1$, so we add a Lagrange multiplier before taking the derivative. The terms of (4.13) that contain a given ψ_{nk}, along with the Lagrange multiplier, are

$$\psi_{nk} \sum_{r=1}^{R} \sum_{s=1}^{S} \ln \alpha_{ks} \theta_{nrs} - \psi_{nk} \ln \psi_{nk} + \alpha_n (\sum_{k=1}^{K} \psi_{nk} - 1). \tag{4.15}$$

By taking the derivative with respect to ψ_{nk} and setting it to 0, we obtain

$$\sum_{r=1}^{R} \sum_{s=1}^{S} \ln \alpha_{ks} \theta_{nrs} - \ln \psi_{nk} - 1 + \alpha_n = 0. \tag{4.16}$$

With a bit of algebra, and setting $\beta_n = \exp(1 - \alpha_n)$. we obtain:

$$\frac{1}{\beta_n} \prod_{r=1}^{R} \prod_{s=1}^{S} \alpha_{ks}^{\theta_{nrs}} = \psi_{nk}. \tag{4.17}$$

By summing over all k, we obtain:

$$\sum_{k=1}^{K} \psi_{nk} = 1 = \frac{1}{\beta_n} \cdot \sum_{k=1}^{K} \prod_{r=1}^{R} \prod_{s=1}^{S} \alpha_{ks}^{\theta_{nrs}}. \tag{4.18}$$

Hence we have the update rule for ψ_{nk}:

$$\psi_{nk}^{*} = \frac{1}{\beta_n} \cdot \prod_{r=1}^{R} \prod_{s=1}^{S} \alpha_{ks}^{\theta_{nrs}}, \text{ where } \beta_n = \sum_{k=1}^{K} \prod_{r=1}^{R} \prod_{s=1}^{S} \alpha_{ks}^{\theta_{nrs}}. \tag{4.19}$$

Similarly, we can derive the update rules for θ_{nrs}:

$$\theta_{nrs}^{*} = \frac{1}{\gamma_{nr}} \prod_{k=1}^{K} \alpha_{ks}^{\psi_{nk}}, \text{ where } \gamma_{nr} = \sum_{s=1}^{S} \prod_{k=1}^{K} \alpha_{ks}^{\psi_{nk}}. \tag{4.20}$$

■ 4.5 Deriving the M-Step

In the M-step, we determine the values of the parameters of the full likelihood model that maximize $L^{var}_{q(C,L;\psi,\theta)}(F, C, L; \alpha, \pi, \mu, \Sigma)$ while keeping the parameters of q fixed. In other words, we find:

$$\alpha^{*}, \pi^{*}, \mu^{*}, \Sigma^{*} = \text{argmax}_{\alpha,\pi,\mu,\Sigma} \{L^{var}_{q(C,L;\psi,\theta)}(F, C, L; \alpha, \pi, \mu, \Sigma)\}. \tag{4.21}$$

■ 4.5.1 Deriving the update rule for π

We derive the update rule for π as in the case of the standard EM algorithm. We have that $\sum \pi_k = 1$. So we add a Lagrange multiplier to the terms of (4.13) that contain π_k:

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \ln \pi_k \psi_{nk} + \eta \left(\sum_{k=1}^{K} \pi_k - 1 \right). \tag{4.22}$$

Taking the partial derivative with respect to π_j produces

$$\frac{1}{\pi_j} \sum_{n=1}^{N} \psi_{nj} - \eta. \tag{4.23}$$

We set this to 0 to find the optimal setting of π_j. Thus $\sum_{n=1}^{N} \psi_{nj} = \eta \cdot \pi_j$. Summing over all such k, we derive

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \psi_{nk} = \eta \sum_{k=1}^{K} \pi_k = \eta, \tag{4.24}$$

based on the constraint that $\sum_{k=1}^{K} \pi_k = 1$. Thus

$$\eta = \sum_{n=1}^{N}\sum_{k=1}^{K} \psi_{nk} = N. \tag{4.25}$$

Hence, we have that

$$\pi_j^* = \frac{1}{N}\sum_{n=1}^{N} \psi_{nj}. \tag{4.26}$$

■ 4.5.2 Deriving the update rule for α

We have that for all k, $\sum_{s=1}^{S} \alpha_{ks} = 1$. Hence for all k, we add a Lagrange multiplier to the terms of (4.13) that contain α to obtain:

$$\sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{s=1}^{S} \ln \alpha_{ks}\theta_{nrs}\psi_{nk} + \nu_k(\sum_{s=1}^{S} \alpha_{ks} - 1). \tag{4.27}$$

Taking the derivative with respect to a given α_{kj}, we calculate:

$$\sum_{n=1}^{N}\sum_{r=1}^{R} \theta_{nrs}\psi_{nk} - \nu_k \cdot \alpha_{ij} = 0. \tag{4.28}$$

By summing over all s and re-arranging, we have that:

$$\nu_k = \sum_{n=1}^{N} \psi_{nk} \sum_{r=1}^{R}\sum_{s=1}^{S} \theta_{nrs} = R \cdot \sum_{n=1}^{N} \psi_{nk}. \tag{4.29}$$

Thus, we obtain:

$$\alpha_{kj}^* = \frac{1}{\nu_k}\sum_{n=1}^{N}\sum_{r=1}^{R} \psi_{nk}\theta_{njr}, \text{ where } \nu_k = R \cdot \sum_{n=1}^{N} \psi_{nk}. \tag{4.30}$$

■ 4.5.3 Deriving the update rules for μ and Σ

The terms of $L^{var}_{q(C,L;\psi,\theta)}(F, C, L; \alpha, \pi, \mu, \Sigma)$ that contain μ_s are

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{r=1}^{R}\theta_{nrs}\cdot(F_{nrs}-\mu_t)^T\Sigma_s^{-1}(F_{nrs}-\mu_s). \tag{4.31}$$

Taking the derivative with respect to μ_s and setting it to 0, we obtain:

$$\sum_{n=1}^{N}\sum_{r=1}^{R}\theta_{nrs}\left(\Sigma_s^{-1}I_{ns}-\Sigma_s^{-1}\mu_s\right)=0. \tag{4.32}$$

It follows that:

$$\mu_s^*=\frac{1}{N_s}\cdot\sum_{n=1}^{N}\sum_{r=1}^{R}\theta_{nrs}\cdot I_{nr},\text{ where }N_s=\sum_{n=1}^{N}\sum_{r=1}^{R}\theta_{nrs}. \tag{4.33}$$

We similarly find the update rule for Σ_s, which is

$$\Sigma_s^*=\frac{1}{N_s}\sum_{n=1}^{N}\sum_{r=1}^{R}\theta_{nrs}\cdot(F_{nr}-\mu_s)\cdot(F_{nr}-\mu_s)^T. \tag{4.34}$$

Chapter 5

Analysis of the Generative Model and Discussion of Results

In this chapter, we discuss the implementation and performance of the algorithm described in the previous chapter. We present the methods used determine the number of patient clusters and the number of disease subtypes and examine parameters estimated by our model. To ensure that our algorithm's results are meaningful, we analyze the spatial contiguity of the disease subtypes and the model's stability. We conclude by discussing the clinical relevance of our results.

5.1 Parameter Selection

The algorithm was run on 2457 patients with 1000 non-overlapping patches randomly chosen from each patient. The patches are $11 \times 11 \times 11$ and the feature vectors are 11-dimensional where the first 9 values consist of GLCM features, and the last two consist of histogram bins, as described in Chapter 3.

The algorithm was run on a range of the number of patient clusters K and disease subtypes S. We chose to examine the model with eight patient clusters and six disease subtypes, as this was the largest number of disease subtypes and patient clusters for which each patient cluster and disease subtype received at least five percent probability. The rest of this chapter proceeds with a discussion of the algorithm's performance with eight patient clusters and six disease subtypes.

5.2 Disease Subtypes

Patches belonging to each of our disease subtypes are shown in Figure 5.1. Subtype 1 is the one that most closely corresponds to normal lung tissue.

We compared the disease subtypes identified by our model to clinically identified

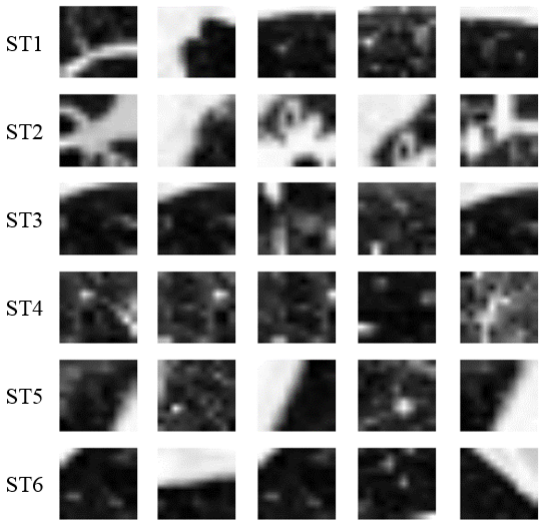


Figure 5.1: Patches showing different disease subtypes identified by our model.

Clinical Label	ST 1	ST 2	ST 3	ST 4	ST 5	ST 6
Normal Lung Tissue	339	0	1	103	7	61
Panlobular Emph.	1	146	9	0	0	0
Paraseptal Emph.	16	53	100	48	20	6
Mild Centrilobular Emph.	96	3	11	68	3	30
Moderate Centrilobular Emph.	69	74	112	28	4	2
Severe Centrilobular Emph.	8	57	49	0	0	0

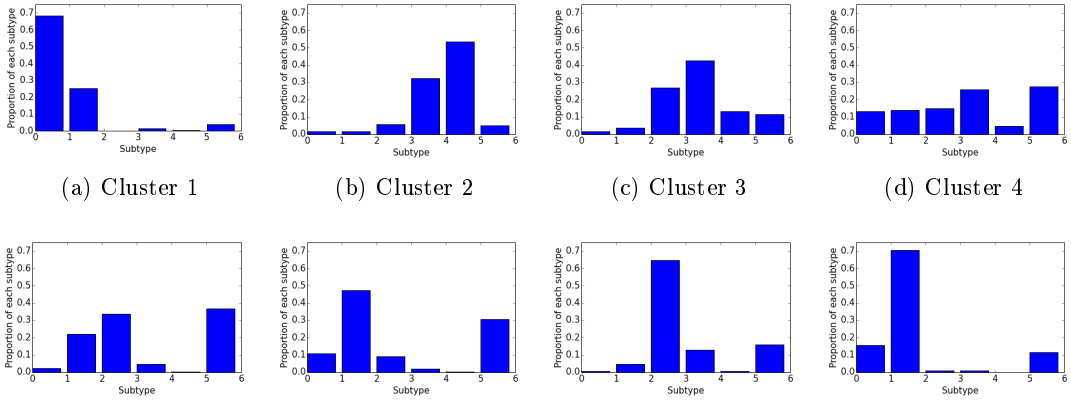
Table 5.1: Confusion matrix between clinically defined subtypes and automatically detected subtypes. The values in the table correspond to the number of patches with the corresponding clinical label and detected subtype.

ones. To this end, we used the labelled patches described in Chapter 3, though we employ different clinical labels from the one used in Chapter 3 for feature selection. Here, we have six clinical labels for our patches: normal lung tissue, panlobular emphysema, and paraseptal emphysema, along with mild, moderate, and severe centrilobular emphysema.

A confusion matrix between the disease subtypes and the clinical labels is shown in

Patient Cluster	1	2	3	4	5	6	7	8
Proportion of Patients (π)	0.258	0.219	0.146	0.123	0.092	0.059	0.053	0.051

Table 5.2: Settings for π: mixing proportions of the patient clusters



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3 (d) Cluster 4

(e) Cluster 5 (f) Cluster 6 (g) Cluster 7 (h) Cluster 8

Figure 5.2: Expected distribution of subtypes in each patient cluster. The graph for cluster k corresponds to the values of α_k.

Table 5.1. Subtype 1 most closely corresponds to normal lung tissue. On the labelled portion of our data set, we found that 67% of patches that were labelled as clinically normal were placed in the same disease subtype by our algorithm, and clinically normal patches represent 64% of all labelled patches within this disease subtype. Panlobular and paraseptal emphysema correspond to disease subtype 2 and subtype 3 respectively. Our results suggest that centrilobular emphysema is a mixture of identified disease subtypes 1, 2, 3 and 4.

■ 5.3 Patient Clusters

The values for π, i.e. the proportion of patients in each cluster is reported in Table 5.2. The values of the expected proportion of subtypes α in each patient category is displayed in Figure 5.2. In this figure, the plot for cluster k corresponds to the values

of α_k - the proportion of subtypes in cluster k. We observe that the distributions of the subtypes is quite different for each patient cluster, which shows that we have successfully identified distinct patient clusters.

Figure 5.3 shows labelled lungs of patients in different patient clusters. In these images, each color corresponds to an emphysema subtype or normal tissue, which is closest to the blue label.

5.4 Spatial Contiguity

Emphysema clusters spatially in the lungs, as do the disease subtypes our algorithm identifies, as can be seen in Figure 5.3. We evaluated spatial contiguity by permutation testing [8]. For each voxel labelled by our algorithm we compute the proportion of neighboring voxels that belong to the same disease subtype. We then average this value over the entire lung to obtain a spatial contiguity score. To obtain a distribution of the score under the null hypothesis we assigned voxels within the lungs to random disease subtypes 1000 times for each scan while maintaining the proportion of disease subtypes for each lung that was identified by our algorithm. We found that across all CT scans, the spatial contiguity scores produced by our algorithm are greater than the maximal values in the corresponding null distribution. This corresponds to rejecting the null hypothesis with $p < 0.001$. Spatial contiguity is an important result, as we have not imposed this constraint on our model, and instead it organically arose out of the data.

5.5 Model Stability

We analyze the model's stability, using a method motivated by Levine, et al. [7]. We ran our algorithm on a randomly selected half of the scans and labelled the remaining scans based on the estimated model parameters. In particular, we assigned each patient to the most likely cluster, and we assigned each voxel to the most likely subtype. We repeated this process 10 times. Hence, we obtain 10 assignments of patients to clusters, and 10 assignment of voxels to subtypes. We only compare the assignments of voxels to subtypes in 100 patients, since it would be too cost-intensive to compute for all of the patients. We calculate the adjusted mutual information between each pair of assignment of patients to clusters, and average these values. Similarly, we calculate the adjusted mutual information between each pair of assignment of voxels to subtypes, and average these scores.

This adjusted mutual information score between two cluster assignments X and Y

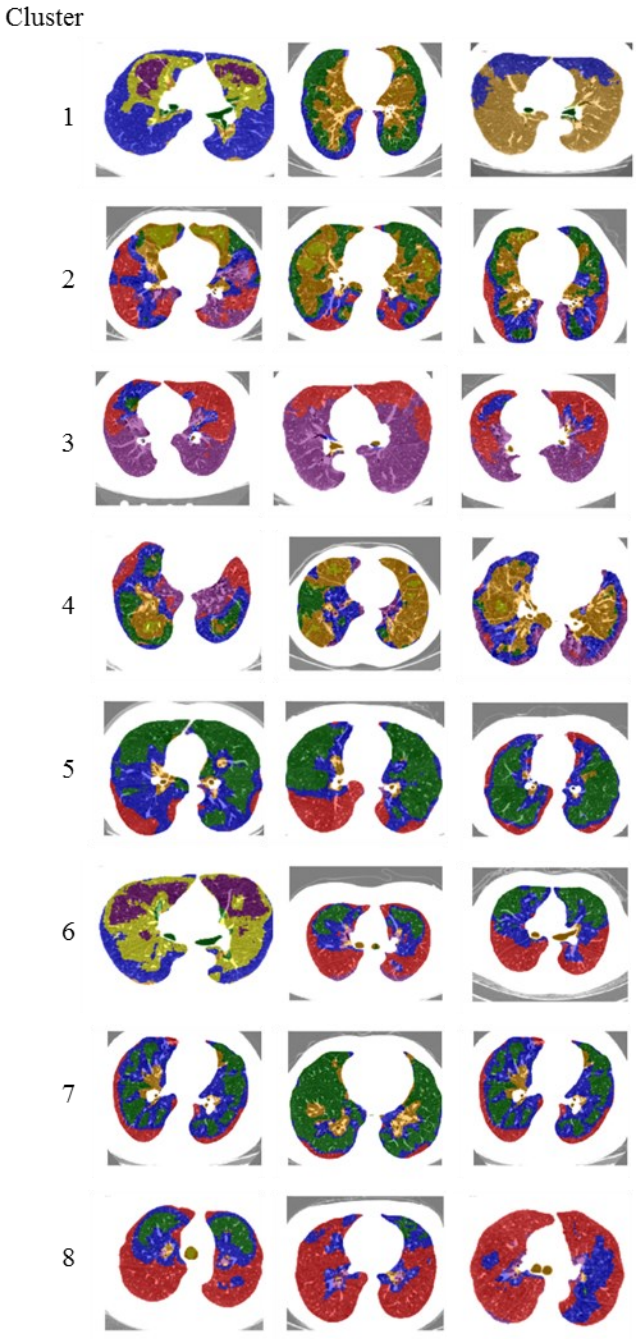


Figure 5.3: Slices from example CT scans from each of the eight patient clusters identified by our algorithm. Colors correspond to disease subtypes identified by our algorithm. Blue most closely corresponds to normal lung tissue.

is defined as

$$\frac{I(X,Y) - E[I(X,Y)]}{\max(H(X), H(Y)) - E[I(X,Y)]}. \tag{5.1}$$

The score takes on values between 0, when the mutual information between two cluster assignments equals its expected value, and 1, when two cluster assignments are identical [26]. Here, $E[I(X,Y)]$ is the expected mutual information in the case that X and Y have the same proportion of elements in each cluster, but the two cluster assignments are independent. $I(X,Y)$ is bounded by $\max(H(X), H(Y))$, so $\max(H(X), H(Y)) - E[I(X,Y)] \geq I(X,Y) - E[I(X,Y)]$. This will be equal precisely when $I(X,Y)$ is maximized, that is when X and Y are identical, producing a score of 1. When X and Y are independent, $I(X,Y) = E[I(X,Y)]$, so the score is 0.

The averaged score across assignments to patient clusters is 0.60 - which shows some stability in these labelings. The averaged score across assignments of voxels to disease subtypes is 0.79. This suggests that the identities of the disease subtypes are more stable than the identities of the patient clusters, though both are consistent across running the algorithm on different subsets of the data. This is likely due to the fact that the disease subtypes are more directly linked to the data, while the patient clusters are linked to the data only through the disease subtypes.

■ 5.6 Associations with Physiological Indicators

To evaluate the clinical relevance of our model, we quantify the associations between the structure detected by our method and the physiological indicators relevant to COPD: six minute walking distance, body mass index (BMI), forced vital capacity (FVC), forced expiratory volume (FEV), change in FVC value from treatment, and the ratio between the FEV and FVC values. We ran our algorithm on a randomly selected half of the scans and labelled the remaining scans based on the estimated model parameters. In particular, we assigned each patient to the most likely cluster and constructed an empirical distribution of disease subtypes for the patient based on the image patches. We repeated this procedure 100 times to estimate variability in the results.

We constructed three baseline models by eliminating patient clusters ($K = 1$) or disease subtypes ($S = 1$) or both ($K = 1,\ S = 1$). In the last case, we extract feature vectors from patches in each patient, and then average and normalize the feature vectors in each patient to produce a single patient-specific feature vector.

5.6.1 Methods for Quantifying Association

The association between patient clusters and physiological indicators is quantified via the normalized mutual information score [26].The normalized mutual information score of two random variables X and Y is defined as

$$\frac{I(X,Y)}{\sqrt{H(X)\cdot H(Y)}} \tag{5.2}$$

where $I(X,Y)$ is the mutual information between X and Y, and $H(X)$ is the entropy of X. This score takes on values between 0 (no association), and 1 (perfect dependency).

To quantify the associations between distributions of disease subtypes or the averaged normalized feature vector for a patient and a physiological indicator we perform linear regression. We have a physiological indicator c, and a_i is the proportion of subtype i in a given patient or the i-th entry in a feature vector. The linear regression finds the optimal settings for $\{\beta\}_{i=1}^{t}$ to best approximate $c = \sum_{i=1}^{t} \beta_i a_i$ across all patients. We can quantify the strength of this correlation with the R^2 score, which is the percentage of the variance in c that is explained by the linear regression. The R^2 score is defined as

$$R^2 \triangleq 1 - \frac{\sum_{j=1}^{N}(c_j - f_j)^2}{\sum_{j=1}^{N}(c_j - \bar{c})^2} \tag{5.3}$$

where $f_j = \sum_{i=1}^{6} \beta_i a_{ij}$, and $\bar{c} = \frac{1}{N}\sum_{j=1}^{N} c_j$. R^2 takes on values between 0 (no linear correlation), and 1 (perfect linear correlation).

Different metrics are used to quantify the associations between patient clusters and proportions of disease subtypes or feature vectors, since the former is a discrete label while the last two are continuous quantities.

5.6.2 Discussion of Identified Associations

Fig. 5.4 reports the associations for all models. These results demonstrate the advantage of modelling both patient clusters and disease subtypes. We observe that there is a stronger association between physiological indicators and patient clusters in the full model than in the model with only clusters. For all physiological indicators, there is a higher association with the distributions of disease subtypes in the full model than in the model with only disease subtypes. This demonstrates that modelling patient clusters produces more clinically relevant distributions of disease subtypes in each patient. The model without patient clusters or disease subtypes exhibits even weaker

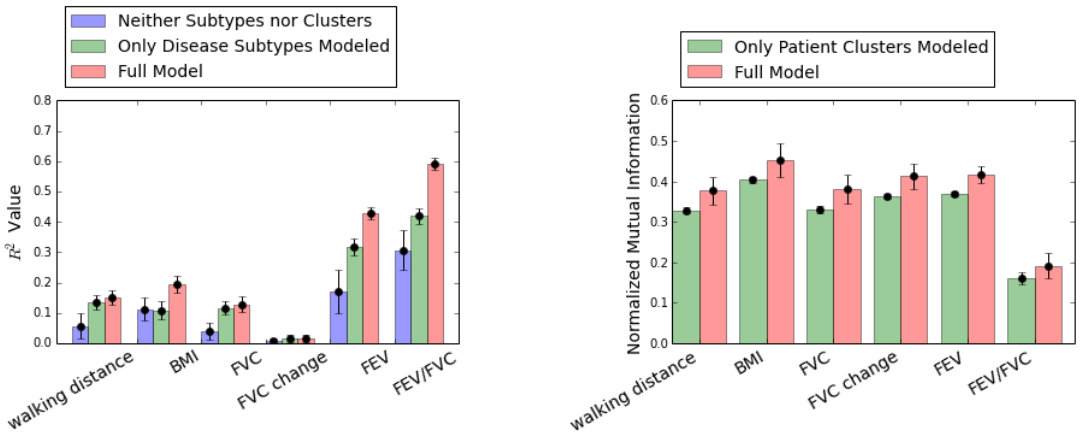


Figure 5.4: Left: R^2 value between the distributions of disease subtypes or feature vectors and physiological indicators. Right: Normalized Mutual Information between patient clusters and physiological indicators.

associations than a model with only disease subtypes.

■ 5.7 Discussion

We have shown that our method produces spatially contiguous clusters - which is an important verification of our results since emphysema patterns tend to cluster spatially. We have also shown that our method is stable across runs on different subsets of the data.

The clinical relevance of our model is demonstrated by the associations between both patient clusters and distributions of subtypes and a variety of physiological indicators. Additionally, there are certain physiological indicators that correlate strongly with patient clusters but not with distributions of disease subtypes, showing the importance of the patient clusters. It appears that some clinical information is present in the distribution of subtypes but not in the patient clusters, suggesting that the patient clusters may not capture all of the necessary clinical information. We have shown that our model has small but significant advantages over a model in which only subtypes or clusters are modeled, and even larger advantages over a model with neither subtypes nor clusters.

Chapter 6

Conclusions and Future Work

The work presented in this thesis enables us to model population structure across a large cohort of patients and to differentiate groups of patients that exhibit the same distributions of emphysema disease subtypes and consequently may have the same clinical prognoses and manifestations of the disease. Additionally, our method enables us to distinguish three-dimensional textures in CT scans of lungs affected by COPD, which correspond to distinct disease subtypes.

6.1 Contributions

In this thesis, we presented an unsupervised framework for the discovery of both patient clusters and of disease subtypes. Specifically, we construct a generative model that parameterizes the assignment of voxels in CT scans to subtypes and the assignment of patients to clusters. The observed data for our algorithm consists of texture descriptors of patches extracted from CT scans of patients with COPD. Our model performs inference using a variational expectation-maximization approach.

Our model enables us to harness the information available in our data set of 2457 CT scans and identify disease subtypes in the context of population structure. We examine the performance of our model and demonstrate that the patient clusters and disease subtypes that our model produces are clinically relevant.

6.2 Extensions and Future Work

Our work could be extended by incorporating several clinical markers into the generative model. In this work, we compare our clusters to these markers but do not model them directly. Many clinical markers correspond to patient prognosis, so their inclusion could cause patients with similar disease prognosis and disease phenotype to be assigned to the same cluster.

Another extension is to incorporate clinically defined subtype labels in a semi-supervised manner. In this framework, our algorithm would generally proceed in an unsupervised manner, but it would attempt to group regions of the lung that clinicians assign the same label to into the same subtype. This model would likely produce different results than the model that we describe in this thesis. It would then be possible to explore how clinician's labelings of emphysema subtypes change the patient clusters.

Further, the patient clusters that our model produces merit further exploration. It would be worthwhile to examine their correlation to genetic markers. An additional extension is to directly examine whether different patient clusters exhibit distinct clinical prognoses or respond to different clinical interventions.

Bibliography

[1] Nematollah K Batmanghelich, Ardavan Saeedi, Michael H Cho, Raúl San Jose Estépar, and Polina Golland, *Generative method to discover genetically driven image biomarkers*, Information processing in medical imaging : proceedings of the conferenceI **24** (2015), 30–42.

[2] Christopher M. Bishop, *Pattern recognition and machine learning (information science and statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] A. Depeursinge, A. Foncubierta-Rodriguez, A. Vargas, D. Van De Ville, A. Platon, P.-A. Poletti, and H. Muller, *Rotation-covariant texture analysis of 4d dual-energy ct as an indicator of local pulmonary perfusion*, (2013), 145–148.

[4] Adrien Depeursinge, Antonio Foncubierta-Rodriguez, Dimitri Van de Ville, and Henning Müller, *Lung texture classification for locally-oriented riesz components*, Medical Image Computing and Computer-Assisted Intervention (2011), 231–238.

[5] Adrien Depeursinge, Antonio Foncubierta-Rodriguez, Dimitri Van De Ville, and Henning Muller, *Three-dimensional solid texture analysis in biomedical imaging: Review and opportunities*, Medical Image analysis **18** (2014), no. 1, 176 – 196.

[6] Jennifer G. Dy, Carla E. Brodley, Avi Kak, Lynn S. Broderick, and Alex M. Aisen, *Unsupervised feature selection applied to content-based retrieval of lung images*, IEEE Transactions on Pattern Analysis and Machine Intelligence **25** (2003), no. 3, 373–378.

[7] Levine E and Domany E, *Resampling method for unsupervised estimation of cluster validity. neural computation*, Neural Computation **13** (2001), no. 11, 2573 – 2593.

[8] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993.

[9] David M. Hansell, Alexander A. Bankier, Heber MacMahon, Theresa C. McLoud, Nestor L. Muller, and Jacques Remy, *Fleischner society: Glossary of terms for thoracic imaging*, Radiology **246** (2008), no. 3, 697–722.

[10] Robert M. Haralick, K Shanmugam, and Its'Hak Dinstein, *Textural features for image classification*, IEEE Transactions on Systems, Man and Cybernetics **SMC-3** (1973), no. 6, 610–621.

[11] Donna L. Hoyert and Jiaquan Xu, *Deaths: Preliminary data for 2011*, National Vital Statistics Report (2012).

[12] Decramer M., Janssens A., and Miravitlles M., *Chronic obstructive pulmonary disease*, The Lancet **379** (2012), no. 9823, 1341 – 1351.

[13] Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi, *Contour and texture analysis for image segmentation*, International Journal of Computer Vision **43** (2001), no. 1, 7–27.

[14] C.S. Mendoza, G.R. Washko, J.C. Ross, A.A. Diaz, D.A. Lynch, J.D. Crapo, E.K. Silverman, B. Acha, C. Serrano, and R.S.J. Estepar, *Emphysema quantification in a multi-scanner HRCT cohort using local intensity distributions*, Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on, May 2012, pp. 474–477.

[15] O. Minai, J. Benditt, and F. Martinez, *Natural history of emphysema*, Proceedings of the American Thoracic Society **5** (2008), no. 4.

[16] C. Mueller-Mang, C. Grosse, and L. Stiebellehner andA. Bankier K. Schmid, *What every radiologist should know about idiopathic interstitial pneumonias*, RadioGraphics **27** (2007), no. 3, 595–615.

[17] Joachim Ofner, *Computational texture analysis in idiopathic interstitial pneumonia*, Master's Thesis at the Vienna University of Technology (2010).

[18] Mithun Prasad, Arcot Sowmya, and Peter Wilson, *Multi-level classification of emphysema in HRCT lung images*, Pattern Analysis Applications **11** (2006), no. 1.

[19] Amir Qaseem, Timothy J. Wilt, Steven E. Weinberger, Nicola A. Hanania, Gerard Criner, Thys van der Molen, Darcy D. Marciniuk, Tom Denberg, Holger Schunemann, Wisia Wedzicha, Roderick MacDonald, and Paul Shekelle, *Diagnosis and*

management of stable chronic obstructive pulmonary disease: A clinical practice guideline update from the american college of physicians, american college of chest physicians, american thoracic society, and european respiratory society, Annals of Internal Medicine **155** (2011), no. 3, 179–191.

[20] Jose Ramos, Thessa Kockelkorn, Isabel Ramos, Rui Ramos, Jan Grutters, Max A. Viergever, Bram van Ginneken, and Aurelio Campilho, *Content based image retrieval by metric learning from radiology reports: Application to interstitial lung diseases.*, IEEE Journal of Biomedical and Health Informatics **PP** (2014), no. 99, 1–11.

[21] Elizabeth A. Regan, John E. Hokanson, James R. Murphy, Barry Make, David A. Lynch, Terri H. Beaty, Douglas Curran-Everett, Edwin K. Silverman, and James D. Crapo, *Genetic epidemiology of COPD (COPDGene) study design*, COPD **7** (2010), no. 1, 32–43.

[22] E.P. Simoncelli and W.T. Freeman, *The steerable pyramid: a flexible architecture for multi-scale derivative computation*, **3** (1995), 444–447 vol.3.

[23] B. M Smith and J. H. M. Austin, *Pulmonary emphysema subtypes on computed tomography in smokers*, The American Journal of Medicine **127** (2014), no. 1.

[24] L. Sorensen, S.B. Shaker, and M. de Bruijne, *Quantitative analysis of pulmonary emphysema using local binary patterns*, IEEE Transactions on Medical Imaging **29** (2010), no. 2, 559–569.

[25] Zavaletta VA, Bartholmai BJ, and Robb RA, *High resolution multi-detector CT aided tissue analysis and quantification of lung fibrosis*, Academic radiology **14** (2007), no. 7, 772–787.

[26] Nguyen Xuan Vinh, Julien Epps, and James Bailey, *Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance*, Journal of Machine Learning Research **11** (2010), 2537 – 2854.

[27] Aziz Z, Wells A, and Hansell Dl, *HRCT diagnosis of diffuse parenchymal lung disease: inter-observer variation*, Thorax **59** (2004), no. 6, 506–511.