Generative Interpretability: Application in Disease Subtyping

Payman Yadollahpour, Ardavan Saeedi, Sumedha Singla, Frank C. Sciurba, Kayhan Batmanghelich

***Abstract*—We present a probabilistic approach to characterize heterogeneous disease in a way that is reflective of disease severity. In many diseases, multiple subtypes of disease present simultaneously in each patient. Generative models provide a flexible and readily explainable framework to discover disease subtypes from imaging data. However, discovering local image descriptors of each subtype in a fully unsupervised way is an ill-posed problem and may result in loss of valuable information about disease severity. Although supervised approaches, and more recently deep learning methods, have achieved state-of-the-art performance for predicting clinical variables relevant to diagnosis, interpreting those models is a crucial yet challenging task. In this paper, we propose a method that aims to achieve the best of both worlds, namely we maintain the predictive power of supervised methods and the interpretability of probabilistic methods. Taking advantage of recent progress in deep learning, we propose to incorporate the discriminative information extracted by the predictive model into the posterior distribution over the latent variables of the generative model. Hence, one can view the generative model as a *template* for interpretation of a discriminative method in a clinically meaningful way. We illustrate an application of this method on a large-scale lung CT study of Chronic Obstructive Pulmonary Disease (COPD), which is a highly heterogeneous disease. As our experiments show, our interpretable model does not compromise the prediction of the relevant clinical variables, unlike purely unsupervised methods. We also show that some of the discovered subtypes are correlated with genetic measurements suggesting that the discovered subtypes characterize the underlying etiology of the disease.**

***Index Terms*—Interpretable Models, Disease Subtyping, Variational Inference, Neural Network, Generative Model, Discriminative Model, Probabilistic Graphical Model, Chronic Obstructive Pulmonary Disease, COPD.**

I. Introduction

Characterizing the heterogeneity of diseases is essential in understanding their etiology [1], improving prediction of patient survival [2], and guiding patient treatment [3], [4]. However, it is a challenging problem since there are many sources of variation at the patient and population-level. It is essential to define heterogeneity objectively such that it is reflective of disease severity. In this paper, we propose to build a generative model to explain variations in the patient and population levels. To ensure the explanation is predictive of disease severity, we train a predictive model that interacts with the generative model in a novel way. We apply our approach in the context of Chronic Obstructive Pulmonary Disease (COPD), which is a highly heterogeneous disease [5], [6].

COPD, which is characterized by inflammation of the airway and destruction of the air sacs (emphysema) [7], is the leading cause of death worldwide [8], [9]. There are differences between risk factors of different COPD subtypes [10], and hence understanding subtypes is important. Respirometry measurement is used for the diagnosis of COPD; however, it cannot identify the underlying process of COPD. Hence, computed tomography (CT) imaging, which allows direct qualitative and quantitative evaluation of tissue destruction, is routinely requested for COPD patients. For example, phenotypic abnormality of emphysema is evident from CT images [11], [12]. Although there has been significant work on defining *visual* subtypes of emphysema [12]–[19] from CT images, there is significant intra-reader and inter-reader variability of visual subtypes [20], [21]. In this paper, we aim at the discovery of visual subtypes in a data-driven way so that they are reflective of disease severity.

Various unsupervised subtype discovery methods have been proposed. Image-based phenotype discovery in CT images via spatial texture patterns have been explored in emphysema [14], [15]. Ross *et al.* [12] propose a generative graphical model that incorporates patient trajectories to identify disease subtypes for COPD. Binder *et al.* [20] present a generative model for unsupervised discovery of visual subtypes for COPD along with inferring population structure. Their method identifies sub-populations and clusters of image pattern simultaneously. One of the underlying assumptions of these methods is that the patient population can be divided into sub-populations, which is disputed for COPD [22]. Furthermore, these methods are unsupervised – solving a highly ill-posed problem – hence, the resulting subtypes may not reflect disease severity.

On the other hand, many supervised methods have been proposed to characterize the severity of lung diseases from CT images [11], [16]–[19]. These methods study local descriptors such as local binary pattern (LBP) [17], wavelet and gray-level features [18] as well as various predictive methods ranging from $k-$nearest neighbor classifier [17] to Support Vector

Maunscript received August 1, 2019. This work was supported in part by the NIH under Grant R01HL141813-01, and the NSF under Grant 1839332 TRIPODS+X.

P. Yadollahpour is with the Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15206 USA (e-mail: payman@pitt.edu). He is also with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: pyadolla@csail.mit.edu).

A. Saeedi is with Butterfly Network Inc., New York, NY 10010 USA (e-mail: asaeedi@butterflynetwork.com).

S. Singla is with the Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15206 USA (e-mail: sumedha.singla@pitt.edu).

F. C. Sciurba is with the Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15261 USA (e-mail: sciurbafc@upmc.edu).

K. Batmanghelich is with the Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15206 USA (e-mail: kayhan@pitt.edu).

Machine (SVM) [11]. However, it is not clear how these methods can inform subtype discovery. Furthermore, thanks to advances in deep learning, the field is shifting toward less generic and more task-specific local descriptors [23], [24] which are more challenging to incorporate with subtyping approaches.

Our proposed approach is different from the previous works in two directions: (1) Rather than modeling the disease cohort into sub-populations, we view it as a continuum. We aim at discovering sub-processes across the disease cohort; each patient is a mixture of these sub-processes. We assume that these sub-processes are manifested in the CT images. We use a probabilistic generative approach for modeling, where the image signature of the subtypes and the patient-specific mixture are latent variables. (2) To ensure that discovered sub-processes are related to the disease severity and not just anatomical variation, we proposed a novel approach to combine the generative model with a predictive model. The predictive model extracts discriminative information from the images. We propose a novel way to incorporate this information into the posterior distribution of the latent variables. Alternatively, our generative model can be viewed as a *template* for interpretation of the discriminative method in a clinically meaningful way. It is a generic framework that is applicable for any choice of predictive model, for example, a deep learning-based method.

This paper makes the following contributions:

- We develop a general framework that allows an explanation template (via a generative model) for a discriminative model. Our approach does not compromise the predictive power of the discriminative model.
- Our framework enables us to incorporate prior knowledge into the explanation. The choice of template is problem dependent. Given that COPD is a heterogeneous disease, a topic model is a natural template for the explanation.
- We propose an efficient algorithm for approximate inference of the posterior distribution over the latent variables, including the image signature of the subtypes and ensuring the discovered subtypes are disease relevant.
- We apply our method on a large scale COPD study showing good predictive performance and clinically interpretable subtypes. Three of the subtypes are shown to have significant genetic heritability.

II. Method

A. Overview of the Model

We develop a framework to *explain* a discriminative model. The discriminative model predicts the severity of lung disease from CT images. Our framework allows a user to provide a *template* for the explainer model. The template is provided in the form of a Probabilistic Graphical Model (PGM). Although our approach is general, we aim at a specific way of explaining the discriminative model in this paper where the lung region of a patient is divided into K different tissue subtypes. The *Explainer* models the data by fitting K typical reoccurring imaging patterns across the population. We dub the typical pattern a tissue subtype. Such a specific way of explaining data results in the so-called topic model [25] as a template of the graphical model; the topics are tissue subtypes. Hence, we use "subtype" and "topic" interchangeably. As a result, our *Explainer* model can be viewed as a subtyping method that incorporates the discriminative information. Subtyping, using topic modeling, can be done in a fully unsupervised fashion [12], [20]; however, (1) the predictive performance of the generative model is reduced, and (2) there is no guarantee that the derived topics are related to the abnormality. Our model addresses these issues. Our method consists of three building blocks:

- *Predictive Model:* A forward discriminative model that accepts images (I_s) as input and produces the disease severity "y_s" as output. It is also produces a subject-level representation ($\boldsymbol{t}_s$) that summarizes the discriminative information of I_s to predict "y_s". The discriminative model can be a deep neural network or a complicated pipeline of functions leading to a prediction.
- *Explainer Template:* A probabilistic graphical model (PGM) specifying a template for the explanation. A PGM is a general framework to represent dependencies between observed and hidden random variables [26]. One can view the explainer as a decoder that maps the latent variables to an observation [27], [28]. In this paper, topic modeling is used as the decoder where the latent variables correspond to image patterns of the subtypes across the population, the proportion of the subtypes in each patient, and the corresponding spatial distribution over the lung region for each patient.
- *Posterior Explainer:* An encoder that uses the subject-level representation from the discriminative model ($\boldsymbol{t}_s$) along with imaging data to produce the posterior distribution over the latent variables of the template model.

The general idea of the paper is shown in Fig.1.

In this paper, we adopt the Bag of Words (BOW) model [29]. It represents a subject s with a *set*, $\mathcal{X}_s$, containing features extracted from N_s regions covering the lung region of subject s. This modeling choice allows us to accommodate lungs of different sizes; the number of elements in $\mathcal{X}_s$ can vary depending on the size of the lungs. The BOW model assumes that features of every subject, $\mathbf{x}_{sn} \in \mathcal{X}_s$, are drawn from subject-specific probability distributions, *i.e.,* $\mathbf{x}_{sn} \sim p_s$. Those regions can be patches or supervoxels covering the lung area; in this paper we use supervoxels. The $\boldsymbol{x}_{sn} \in \mathbb{R}^D$ represents a D-dimensional descriptor centered at spatial location n in the image of subject s. Effectively,

1) Our discriminative model maps subject probability densities, p_s, to their corresponding disease severity values, y_s's, without directly modeling p_s.
2) The explainer template provides a parametric form for p_s.
3) Finally, the posterior model uses the subject-level representation and $\mathcal{X}_s$ to estimate the parameters of the template model.

In the following section, we present our discriminative model (Section II-B) followed by the *Explainer* (Section II-C). We discuss our specific choice for the *Explainer* model and show how the prediction and *Explainer* models can interact.

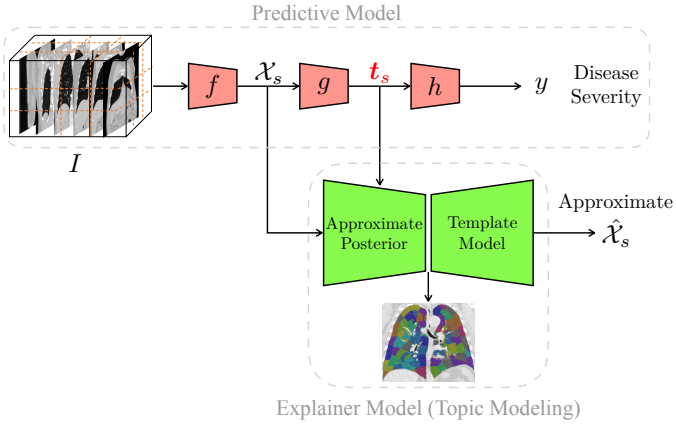


Fig. 1: Outline of the approach consisting of separate *explainer* and *predictor* models. The predictor model is composed of feature extractor f computing bag of features $\mathcal{X}_s$; aggregator g producing the subject-level representation, $\boldsymbol{t}_s$, and regressor h predicting y. The $\boldsymbol{t}_s$ provide supervision to the explainer model to ensure that the patient specific subtype proportions are relevant to the disease severity. The explainer model consists of a *template* for an explanation (can be viewed as a *decoder*) and a posterior estimator (can be viewed as an *encoder*).

B. Predictive Model

Consider a discriminative model for predicting disease severity y_s from a subject's lung CT image I_s. We define this model as a composition of two functions: (1) $f(\cdot)$ which is a function that extracts local descriptors from image I_s, hence $\mathcal{X}_s = f(I_s)$, and (2) an aggregation function, $g(\cdot)$ which we use to construct subject-level features relating the subject to the rest of the population. We minimize

$$\ell(y_s; h(\overbrace{g(f(I_s))}^{t_s})), \tag{1}$$

where h is a regressor or a classifier, depending on y being continuous or discrete and $\ell(\cdot;\cdot)$ is a loss function that is chosen accordingly. We define $\mathbf{t}_s \triangleq g(\mathcal{X}_s)$ to be the features relating the subject to the rest of the population. Each of the functions can either be hand engineered or learned; for example $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ can consist of different layers of a CNN, or a combination of hand engineered feature functions with aggregation performed by summation, followed by prediction via a regression model. In this paper, $f(\cdot)$, is a hand-crafted feature but the same machinery applies to deep learning based features.

As mentioned earlier, we model local features of subject s as samples drawn from its probability distribution p_s. The aggregator maps the probability density to a vector $\boldsymbol{t}_s$ relating the subject to the rest of the population. To do that, we take the following steps. First, we estimate the Kullback-Leibler (KL) divergence between every pair of probability distributions. Second, we convert the distribution distance to a proper similarity kernel. Finally, we use a dimensionality reduction method to estimate $\boldsymbol{t}_s$ from the similarity kernel. The pipeline is shown in Fig.2.

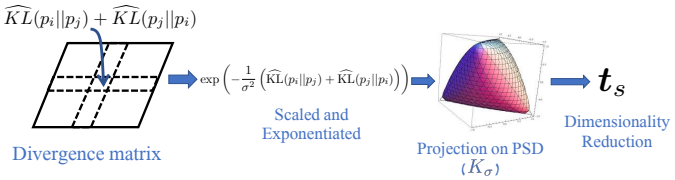


Fig. 2: Construction of the subject-level features ($\boldsymbol{t}_s$) has the following steps: approximating pairwise divergence matrix, exponentiating the matrix, projecting it on the PSD cone, and reducing the dimensionality.

a) Estimating KL divergence: The KL divergence has the following form,

$$\mathrm{KL}(p_i \| p_j) = \int_{\mathbb{R}^d} \log \frac{p_i(x)}{p_j(x)} p_i(x) dx. \tag{2}$$

In this section, we do not assume any explicit parametric form for p_i. Even with a parametric form, estimating the KL divergence is not straightforward. Instead of assuming an explicit parametrization, we use a non-parametric estimator for KL divergence that is consistent and unbiased [30]. The estimator is scalable for high-dimension features and it only requires the nearest neighbor graph that can be approximated using a hashing method [31]. The general idea of the estimator is explained in Appendix A. We use $\widehat{\mathrm{KL}}(p_i \| p_j)$ to denote the estimator for the KL divergence.

b) Computing the Similarity Kernel Matrix: The similarity kernel matrix is a Positive Semi-Definite (PSD) matrix. For example, exponentiating the ℓ_2-distance between features results in a proper similarity kernel matrix known as an RBF kernel. However, the KL divergence is neither symmetric nor a proper metric. First, we compute an $S \times S$ matrix where the entry in row i and column j is

$$[L_\sigma]_{ij} = \exp\left(-\frac{1}{\sigma^2}\left(\widehat{\mathrm{KL}}(p_i \| p_j) + \widehat{\mathrm{KL}}(p_j \| p_i)\right)\right). \tag{3}$$

The variable σ is set to the median of KL divergences (so-called median trick [32]). Then, we project this matrix onto the PSD cone to construct the kernel,

$$K_\sigma = \mathrm{Proj}_{\mathrm{PSD}}(L_\sigma), \tag{4}$$

where $\mathrm{Proj}_{\mathrm{PSD}}$ computes the Singular Value Decomposition of the input matrix and sets the negative singular values to zero.

c) Computing Subject Representation ($\boldsymbol{t}_s$): Since K_σ is a PSD matrix, one can compute $K_\sigma = BB^T$ and view columns of B as an *implicit* characterization of the subjects. However, the columns of B are high dimensional (as many as the number of patients in the dataset). We use Locally Linear Embedding (LLE) to reduce the dimensionality [33]. Other dimensionality reduction methods can be applied as well.

C. Explainer Template

A predictive model can be explained in various ways. In this paper, we would like to allow a practitioner to use knowledge about the disease as a "template" for the explanation. The *Explainer* model approximates the distribution of the image data, and we would like that explanation to be consistent with

the disease severity. To avoid compromising the prediction task, we insert the subject representation ($\boldsymbol{t}_s$) into the posterior estimation of the parameters. In this section, we discuss the population-level and the subject-level modeling assumptions on the distribution. In the next section, we discuss how $\boldsymbol{t}_s$ can inform the posterior distribution.

a) Population-Level Model: Our population model is based on the truncated Hierarchical Dirichlet Process (HDP) [34] . The model assumes that there are K tissue types, "topics", that are shared across subjects in the population. We let a Gaussian distribution with mean vector $\boldsymbol{\mu}_k \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^D \times \mathbb{R}^D$ generate supervoxel descriptors $\boldsymbol{x}_{sn}$. For notational brevity, let $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. As $K \to \infty$, this model converges to a non-parametric HDP [35], [36]. Rather than choosing specific values for K, this model chooses a large enough K and imposes a sparsity term on the allocated topics so that the actual number of topics is discovered from data. To do that the HDP follows the so-called "stick-breaking" construction [34],

$$\boldsymbol{\beta} \sim \text{GEM}(\alpha): \quad \tau_j \sim \text{Beta}(1, \alpha), \quad \beta_k = \tau_k \prod_{j<k}(1-\tau_j) \tag{5}$$

where $\boldsymbol{\beta} \sim \text{GEM}(\alpha)$ denotes sampling from a stick-breaking distribution and α and γ are tunable hyper-parameters of the model. For computational reasons, we also assume a conjugate prior for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$:

$$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim \text{NIW}(\eta)$$

where $\text{NIW}(\eta)$ is the Normal-Inverse-Wishart distribution with hyper-parameters η.

b) Subject-Level Model: For subject s, $\boldsymbol{\pi}_s = [\pi_{s1}, \cdots, \pi_{sK}]$ and $\{z_{sn}\}_{n=1}^{N_s}$ are latent random variables denoting the proportion of topics and the allocation of the supervoxels to the topics (*i.e.,* $z_{sn} \in [K]$) respectively. The $\boldsymbol{\pi}_s$ follows the Dirichlet distribution,

$$\boldsymbol{\pi}_s | \boldsymbol{\beta} \sim \text{Dir}(\alpha\beta_1, \cdots, \alpha\beta_K),$$

where α is a hyper-parameter. The $z_{sn} = k$ indicates supervoxel n of subject s follows the local image descriptor of topic k:

$$z_{sn} | \boldsymbol{\pi}_s \sim \text{Cat}(\boldsymbol{\pi}_s), \qquad I_{sn} | z_{sn}, \{\boldsymbol{\theta}_k\}_{k=1}^K \sim \mathcal{N}(\mu_{z_{sn}}, \Sigma_{z_{sn}});$$

$\text{Cat}(\boldsymbol{\pi}_s)$ represents a categorical distribution.

For notational convenience, we define $\mathcal{D} = \{\mathcal{X}_s\}_{s=1}^S$ to be all image data, $\mathcal{S} = \{z_{sn}, \boldsymbol{\pi}_s\}_{s=1}^S$ to be all subject-specific latent variables, and $\mathcal{P} = \{\boldsymbol{\theta}_k, \boldsymbol{\beta}\}$ to be all population-based latent variables. The joint distribution of all random variables can be written as follows,

$$p(\mathcal{D}, \mathcal{S}, \mathcal{P}) = \prod_{s,n} p(\boldsymbol{x}_{sn} | z_{sn}, \{\boldsymbol{\theta}_k\}) \prod_{s,n} p(z_{sn} | \boldsymbol{\pi}_s) \times \prod_s p(\boldsymbol{\pi}_s | \boldsymbol{\beta}, \boldsymbol{\alpha}) p(\boldsymbol{\beta} | \gamma). \tag{6}$$

The graphical model in Fig.3 summarizes all assumptions of the template model.

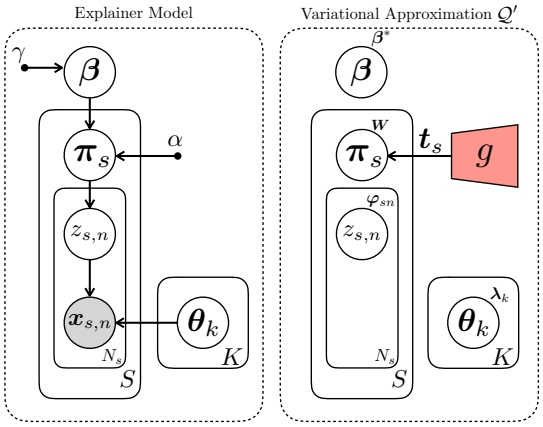


Fig. 3: (*left*) Depicts template for the explainer model, modeled as a topic model. The white and gray circles represent latent and observed random variables and the arrows show conditional dependencies. (*right*) The PGM showing the variational approximation $\mathcal{Q}'$ to the explainer model with subject-level representation $\boldsymbol{t}_s$ injected into the estimate for $q(\boldsymbol{t}_s)$.

D. Inference of Disease Related Posterior Explainer

We propose to account for the disease relevant information in the approximation to the posterior distribution of the latent variables. Assuming that $\boldsymbol{t}_s$ is informative to the prediction of disease severity, it should be incorporated into our approximation. First, we explain the classical approach, and then explain our method to incorporate $\boldsymbol{t}_s$.

Variational Bayes Approximate of the Posterior: We seek the true posterior distribution of the model parameters,

$$p(\mathcal{S}, \mathcal{P} | \mathcal{D}) = \frac{p(\mathcal{D}, \mathcal{S}, \mathcal{P})}{\int p(\mathcal{D}, \mathcal{S}, \mathcal{P}) d\mathcal{S} d\mathcal{P}}. \tag{7}$$

Exact computation of the posterior quantities is computationally intractable since the denominator is hard to compute. Therefore, Variational Bayes [37], [38] approximates the posterior by maximizing the so-called Evidence Lower Bound (ELBO) with respect to q,

$$\max_{q \in \mathcal{Q}} \mathcal{L}(q), \quad \mathcal{L}(q) \triangleq \mathbb{E}_q \left[\ln p(\mathcal{D}, \mathcal{S}, \mathcal{P})\right] - \mathbb{E}_q \left[\ln q(\mathcal{S}, \mathcal{P})\right],$$

where $q \in \mathcal{Q}$ is an approximate distribution from the family of computationally efficient probability densities $\mathcal{Q}$. As it is common in mean-field variational inference [37]–[40], we assume the following form for the approximate posterior, $q(\cdot)$,

$$\mathcal{Q}: q(\mathcal{S}, \mathcal{P}) = q(\boldsymbol{\beta}; \beta^*) \underbrace{\prod_s q(\boldsymbol{\pi}_s; \boldsymbol{\omega}_s)}_{\text{subject-level}} \underbrace{\prod_{s,n} q(z_{sn}; \boldsymbol{\varphi}_{sn})}_{\text{spatial level}} \underbrace{\prod_k q(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k)}_{\text{population-level}}, \tag{8}$$

where β^*, $\boldsymbol{\varphi}_{sn}$, $\boldsymbol{\lambda}_k$, and $\boldsymbol{\omega}_s$ are the variational parameters corresponding to the random variables β, z_{sn}, $\boldsymbol{\theta}_k$, and $\boldsymbol{\pi}_s$ respectively. We use the variational parameters of the approximating distribution $q(\mathcal{S}, \mathcal{P})$ to construct estimates of the relevant model parameters. Specifically, we seek (1) the posterior distribution of $\boldsymbol{\theta}_k$'s as the image descriptors of each subtype (topic) which is on the *population-level*, (2) the posterior distribution of $\boldsymbol{\pi}_s$ as the proportion of subtypes per subject which is on the *subject-level* and (3) the posterior distribution of $z_{s,\cdot}$ that visualizes the spatial distribution of

the subtypes within the lung of patient s which is a *spatial level*. The exact parametric form for each term is given in Appendix B.

Injecting Discriminative Features into Posterior: In the previous sections, we described the standard topic model construction and the corresponding family of variational distributions used to approximate the posterior of the latent variables in the model. As mentioned at the beginning of this section, we also want the explainer model to be informed by features that are known to be highly predictive of disease severity. Thus our goal is to define a new family of approximating posterior distributions, $\mathcal{Q}'$, that is as discriminative as the predictive model. To do that, we use $\boldsymbol{t}_s$, the subject specific representation, to encode the subject-level latent variable. In other words, we use $\boldsymbol{t}_s$ to parameterize $q(\boldsymbol{\pi}_s)$,

$$\mathcal{Q}' : q(\mathcal{S}, \mathcal{P}) = \quad q(\boldsymbol{\beta}; \beta^*) \quad \prod_s \overbrace{q(\boldsymbol{\pi}_s | \boldsymbol{t}_s; \boldsymbol{W})}^{*} \prod_{s,n} q(z_{sn}; \varphi_{sn}) \times \prod_k q(\boldsymbol{\theta}_k; \lambda_k), \tag{9}$$

where $\boldsymbol{W} = \{\boldsymbol{W_\sigma}, \boldsymbol{W_\mu}\}$ is a new parametrization of the latent variables $\boldsymbol{\pi}_s$. Note that whereas before we had different variational parameters $\boldsymbol{\omega}_s$ for each subject, we now have one set of parameters $\boldsymbol{W}$ shared across all subjects. The marginal distribution $q(\boldsymbol{\pi}_s)$ is a natural part of the topic model to introduce $\boldsymbol{t}_s$ because $\boldsymbol{\pi}_s$ is a *subject level* characterization of the topics and $\boldsymbol{t}_s$ characterizes the subject with respect to the rest of the population.

We model $q(\boldsymbol{\pi}_s)$ implicitly by sampling from a Gaussian distribution and passing the samples through a function to normalize them to a simplex (*i.e.,* $\sum_k [\boldsymbol{\pi}_s]_k = 1$). Similar to the idea of a Variational Autoencoder (VAE) [41], we parameterize the mean and the variance of the Gaussian by a neural network. However, instead of inputting the original image, we use the subject-level representation, $\boldsymbol{t}_s$, as input:

$$\begin{aligned} \boldsymbol{\epsilon} &\sim \mathcal{N}(0, I_{K\times K}) \\ \boldsymbol{\psi}_s &= \boldsymbol{\mu}(\boldsymbol{t}_s; \boldsymbol{W_\mu}) + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}(\boldsymbol{t}_s; \boldsymbol{W_\sigma}) \\ \boldsymbol{\pi}_s &= h_{SB}(\boldsymbol{\psi}_s), \end{aligned} \tag{10}$$

where $\boldsymbol{\mu}(\boldsymbol{t}_s; \boldsymbol{W_\mu})$ and $\boldsymbol{\sigma}(\boldsymbol{t}_s; \boldsymbol{W_\sigma})$ are neural networks computing the mean and variance vector of $\boldsymbol{\psi}_s$, respectively. The $h_{SB}(\cdot)$ is a function transforming the unbounded values of $\boldsymbol{\psi}_s$ drawn from a Gaussian distribution to a random variable on a simplex, *i.e.,* $h_{SB} : \mathbb{R}^K \to \Delta^K$. Many choices are possible for $h_{SB}(\cdot)$, such as the *softmax* function. However, computing the probability density of the transformed random variable is not always straightforward. Here, we choose the following form that enables us to have a closed-form probability density for $\boldsymbol{\pi}_s$ [42],

$$h_{SB}(\boldsymbol{\psi}_s) : \quad \boldsymbol{\pi}_{sk} = \sigma(\boldsymbol{\psi}_{sk})(1 - \sum_{j<k} \boldsymbol{\pi}_{sj}). \tag{11}$$

The $\boldsymbol{\pi}_s$, which is the result of a change of variable, has the following probability density,

$$q(\boldsymbol{\pi}_s | \boldsymbol{t}_s; \boldsymbol{W}) = \mathcal{N}(\boldsymbol{\epsilon}; 0, \mathrm{diag}(\boldsymbol{\sigma}^2)) \left| \left\{ \frac{\partial [\boldsymbol{\pi}_s]_i}{\partial [\boldsymbol{\psi}_s]_j} \right\} \right|^{-1}, \tag{12}$$

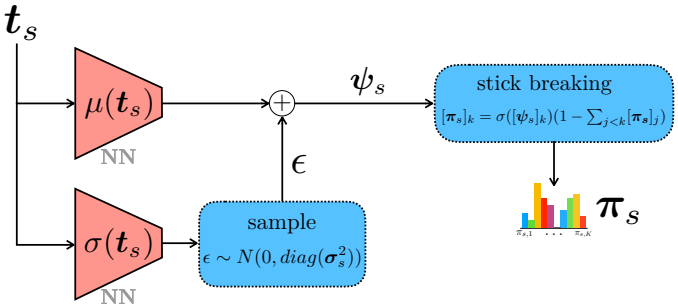


Fig. 4: Schematic showing how the topic proportions ($\boldsymbol{\pi}_s$) are constructed as a function of the subject-level features $\boldsymbol{t}_s$. Learned functions $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ are neural network parameterized by $\boldsymbol{W_\mu}$ and $\boldsymbol{W_\sigma}$ respectively. The stick-breaking constructions of $\boldsymbol{\pi}_s$ normalizes $\boldsymbol{\psi}_s$.

where $\left| \left\{ \frac{\partial [\boldsymbol{\pi}_s]_i}{\partial [\boldsymbol{\psi}_s]_j} \right\} \right|$ is the determinant of the Jacobian which is easily computable (see Appendix B). This is a computationally appealing property for our optimization-based inference because we can easily plug it into the factorization in Eq. 9. The schematic is shown in Fig.4.

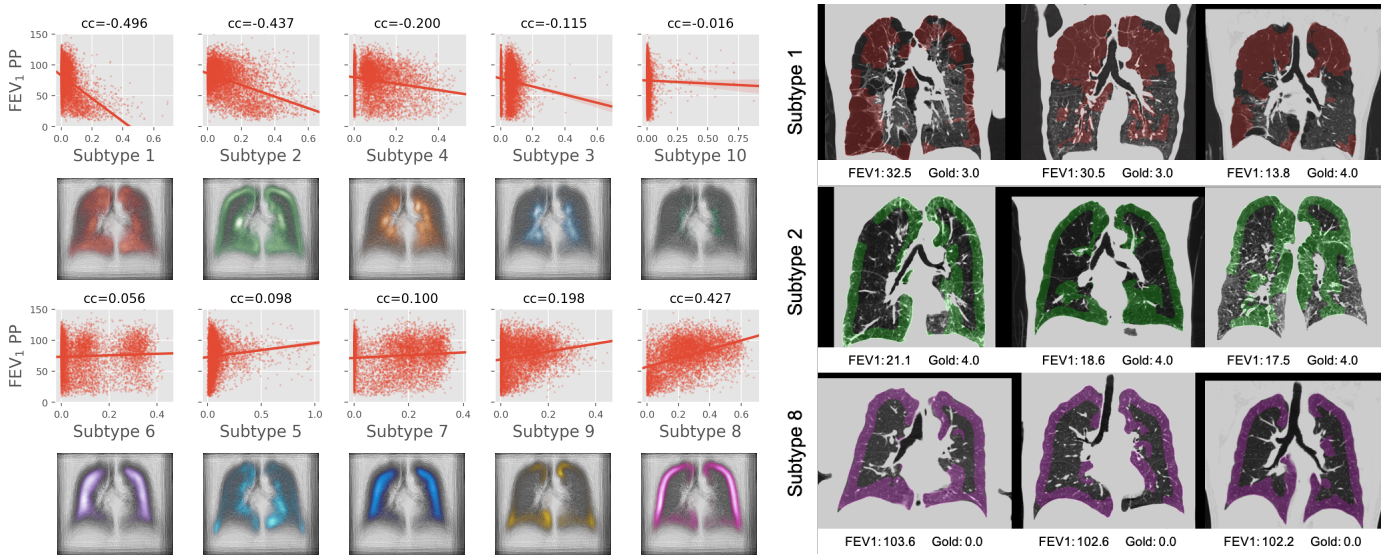
Similar to the classical model in Eq. 8, the parameters of this model are learned by maximizing the ELBO. All updates have a similar form as before except $\boldsymbol{W_\mu}$ and $\boldsymbol{W_\sigma}$, for which we use stochastic gradient descent (see Appendix B for more details).

III. Experiments

In this section, we evaluate the proposed method for lung tissue subtyping on a large-scale dataset from the COPDGene study [43]. First, we evaluate our discriminative model by predicting a few clinical measurements that are indicative of disease severity. We will explain and justify our choices of the feature extractor, $f(\cdot)$, and aggregator $g(\cdot)$. Then, we compare the predictive power of our explainer model, that can exploit the discriminative information used by the predictive model, with that of an unsupervised model. We use topic modeling as the unsupervised model since that is the template for the explainer. Finally, we visualize the subtypes on the subject and population-level and explain the clinical interpretation of each subtype. We further justify the discovered subtypes by studying the genetic heritability of each subtype.

A. Setup

a) ***Feature Extractor****:* We apply our method to lung CT inspiratory images of 7,292 subjects from the COPDGene study [43]. We first oversegment the lung volume into spatially homogeneous regions that align with image boundaries using the SLIC superpixel segmentation algorithm [44]. For each 3D superpixel, intensity histogram and texture based local image features are extracted, as has been shown to be important in characterizing emphysema [45], [46]. We compute Haralick features (`Hara`) from the Gray-Level Cooccurance Matrix (GLCM) that encode image texture but also incorporate intensity [47]. We also separately compute 32-bin intensity histogram features (`Hist`) for each superpixel, following Sorensen *et al.* [46]. For an alternate texture feature, we use



(a) (b)

Fig. 5: (a) *Odd Rows:* Pearson correlation between load of subtype and FEV_1. The x- and y-axis are the load of the subtype and FEV_1 respectively. *Even Rows:* Visualization of spatial average of the learned subtypes across the population shown on a coronal slice of a lung atlas. (b) *Subtypes 1*, *2*, and *8* depicted on a set of nine patients. *Subtypes 1* and *2* are correlated with increase in severity of COPD (negatively correlated with FEV_1), whereas *subtype 8* appears to be healthy tissue (positively correlated with FEV_1).

a rotationally invariant descriptor proposed by Liu *et al.* [48] which computes the histogram of gradients of pixels belonging to a superpixel on a unit sphere using spherical harmonics; we refer to it as `sHOG`.

The aggregation function, $g(\cdot)$, constructs subject-level features from the local image features which has the following steps (described in Sec. II-B): (1) we use a non-parametric estimate of the Kullback Leibler divergence (KL), (2) we symmetrize the KL matrix and projection it on the PSD cone, then (3) apply the Cholesky factorization and use the Locally Linear Embedding (LLE) method [33] to reduce the dimensionality of the factors to a 100-dimensional subject-level feature (*i.e.,* $\boldsymbol{t}_s$).

We compare the predictive performance of $\boldsymbol{t}_s$ with two baselines. First, Low Attenuation Area below Hounsfield Unit of -950 on Inspiration CT image (%LAA-950Insp) which is commonly used as a clinical measure of emphysema. Second, a subject-level representation learned by a traditional bag-of-words (BOW) model which is the $K-$means algorithm, setting $K = 100$ to make it comparable with our representation.

b) ***Initialization of Template and Explainer Models****:* To initialize the parameters of the NIW distribution, $\{\boldsymbol{\theta}_k\}_{k=1}^K$, in the template and explainer models, we ran unsupervised hierarchical clustering [49] on local image features extracted from supervoxels of the training set. The hierarchical clustering cut-off threshold was set to match the number of tissue subtypes K. Each subtype distribution was subsequently initialized with the sufficient statistics computed from the corresponding cluster.

c) ***Evaluation Metric****:* To evaluate our subject-level representation objectively, we use the representation to predict a few clinical variables that are indicative of disease severity. We compare the performance with that of BOW and %LAA-950Insp. More specifically, we use the following measurements:

- Percent Predicted Forced Expiratory Volume in one second (FEV_1 PP): A measure of lung function which is the percentage of normal predicted values of FEV_1 for individuals in the population with similar age, height, weight, gender and ethnicity. Lower values indicate more severe disease.
- Ratio of FEV_1 to Forced Vital Capacity (FEV_1/FVC): Forced Vital Capacity (FVC) is the total amount of air an individual can exhale forcefully after taking the deepest breath possible. This ratio represents the proportion of an individual's vital capacity that they can breathe out in one second.
- Global Initiative for Obstructive Lung Disease (GOLD): GOLD is a discrete value derived from two Spirometry measurements and is between zero and four where zero is used for people at risk (Normal Spirometry but Chronic Symptoms), 1-4 denote Mild to Very Severe COPD. The -1 is used for subjects who have Preserved Ratio Impaired Spirometry (PRISm), which indicates that they have reduced FEV_1 while having preserved FEV_1/FVC.

B. Qualitative Evaluation of Subtypes

a) ***Population-Level Interpretation****:* To summarize the results of the *Explainer* model, we compute the posterior distribution of z_{sn}. The $P(z_{sn} = k|\mathcal{D})$ represents the posterior probability of supervoxel n of subject s being assigned to subtype k which can be visualized as a label mask. Examples of such masks are shown in Figure 5b for a few subjects and subtypes. We register the label masks of all the subtypes to a common space to compute the average distribution of each

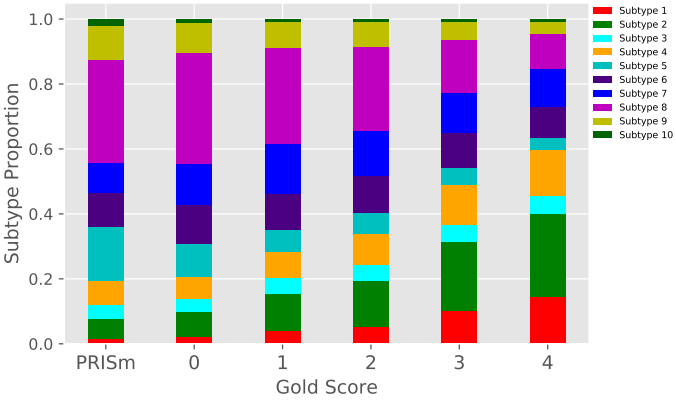


Fig. 6: Subtype proportions averaged over subsets of the population with GOLD score values PRISm, 0, 1, 2, 3, and 4.

subtype across the population. Figure 5a shows these average distributions for each subtype along with corresponding scatter plots denoting the correlation between the load of the subtype and FEV_1 PP. Each dot in the scatter plot denotes one subject where $y-$axis corresponds to FEV_1 PP and $x-$axis is the average of the probabilities of that subtype over all supervoxels of the subject. A positive correlation suggests that tissue type is healthy and negative correlation suggests a disease-related subtype.

We also study the average distributions of the subtypes and their variations among patients with different GOLD scores. The result is shown in Figure 6. Each bar represents a sub-population of patients with a particular GOLD score and colors within the bar represent the average proportion of a subtype within that sub-population. All bars have equal sizes but the proportion of subtypes varies. The proportion of *subtype 1* and *2* increase as we move from PRISm to GOLD score 4 (indicating severely diseased). *Subtype 8*, in contrast, decreases with increased severity. *Subtype 5* is notable because even though it is not significantly correlated with disease, it is prevalent in PRISm sub-population relative to other GOLD scores.

b) ***Patient-Level Interpretation****:* To have a better understanding of subtypes, we visualize $P(z_{sn} = k|\mathcal{D})$ on lung CT's of nine subjects for $k = 1, 2, 8$ which have the strongest correlation with FEV_1. Figure 5b shows that *subtype 1* is found primarily on pulmonary bullae and *subtype 2* captures patients with peripheral bronchiolitis in patients with severe pulmonary disease (*i.e.,* Gold score $\geq$ 3). On the other hand *subtype 8* is very pronounced on the rind of three subjects with healthy lungs.

To get a clinical understanding of these subtypes we asked a clinical expert to inspect all subtypes showing average and subject-level representation. Tissue subtypes 1, 2, 3, 4, and 10 are negatively correlated with FEV_1PP. Thus these subtypes are correlated with increased disease severity. Tissue *subtype 1* tends to characterize paraseptal emphysema and is often found in regions containing pulmonary bullae. *Subtype 1* tends to pick up low attenuation areas on the surface. *Subtype 2* is often indicative of peripheral bronchiolitis, picking up peripheral rind linear opacities in the lung, in some case blood vessels or lymphatics, as well as tree-in-bud opacities. *Subtype 3* predominantly captures different pathological features. It is associated mostly with large high attenuation areas like scarring and vessels as well as airways. *Subtype 4* picks up on more preserved (*i.e.,* less destruction) areas in patients with emphysema. *Subtype 10* is mostly related to the unexplained image statistics associated with large high attenuation areas.

In contrast subtypes 5, 6, 7, 8, and 9 are negatively correlated with increased disease severity. *Subtype 5* captures regions that are more relatively hyperattenuated than surrounding regions. *Subtype 6* picks up on some dimensional feature of the thorax, maintaining a distance on structure – though it is not clear what it is picking up. This is also true for *subtype 7*, which was difficult for the clinical expert to characterize. Subtypes 5, 6, and 7 tend to be attenuation agnostic. *Subtype 8* is associated with more normal and blotchy regions on the rind of the lung. *Subtype 9* is characteristic of thicker peripheral opacities and lines on the apex of the lung which might be indicative of higher diffusing capacity.

C. Quantitative Evaluation of the Subtypes

In this section, we evaluate the discriminative power of the subject level representation ($\boldsymbol{t}_s$) introduced in Section II-B. Then, we examine the inferred explainer model (Section II-D). To do that, we compare our method with the unsupervised method of topic modeling. For both models, we compute the posterior means of the subtype proportions (*i.e.,* $\mathbb{E}_q[\boldsymbol{\pi}_s|\mathcal{D}]$) and use it as feature vectors for regression tasks predicting a few measures of disease severity. For FEV_1 PP, FEV_1/FVC, and distance walk, we report the coefficient of determination R^2, and for GOLD score, we predict the accuracy. We also compute the genetic heritability for each subtype which is the proportion of the variance explained by the genetic similarity between subjects.

a) ***Discriminative Power of the Representation****:* We compare the discriminative performances of the three local image descriptors along with two methods of building the subject-level representation. We separately train linear regression models (via Ridge Regression) to predict FEV_1 PP and FEV_1/FVC from the subject-level features ($\boldsymbol{t}_s$). We use the predicted values to compute the GOLD score. We report the average accuracy rate performed on 5-fold cross validation. Since the GOLD score is a discrete but ordered value, we report the percentage of cases whose classification lays within one class of the true value (one-off) as well as exact value. Table I compares the performance of our subject-level descriptor (Distribution Distance (KL)) with classical BOW model (K-means) for various choices of local image features. Our approach outperforms the threshold-based approach (%LAA-950Insp) as well as BOW across all choices of local image descriptors. While all three choices of local image descriptors perform equally well when used by our method, there is significant variation in performances when BOW is used. In the rest of the experiments, we opt to use Hist+sHOG as the local image features for computing the subject-level representation due to the slight advantage in performance.

To evaluate which subject-level representation is best suited for characterizing the severity of the disease, we trained

Local Image Feature	Subject-level Descriptor	Exact Acc (Std dev)	One-off Acc (Std dev)
Baseline	%Low Attenuation Level (-950)	0.56 (0.03)	0.76 (0.02)
`Hara`	BOW (K-means)	0.47 (0.02)	0.71 (0.02)
	Distribution Distance (KL)	0.58 (0.03)	0.83 (0.02)
`Hist`	BOW (K-means)	0.54 (0.04)	0.79 (0.01)
	Distribution Distance (KL)	0.57 (0.03)	0.82 (0.01)
`Hist+sHOG`	BOW (K-means)	0.57 (0.03)	0.82 (0.01)
	Distribution Distance (KL)	**0.59** (0.03)	**0.84** (0.01)

TABLE I: Average classification accuracy of predicting GOLD 5 classes from subject-level descriptors. Subject-level descriptors are computed from corresponding local image features in each row. `Hara`, `Hist`, `Hist+sHOG` denote Haralick, Histogram, Histogram combined with Spherical Histogram of Gradient descriptors respectively. Results are averaged across 5 cross-validation folds. *One-off Acc* is the percentage of times the predictor was at most one-off in predicting GOLD score.

	R^2			
Subject-Level Descriptor	FEV_1 PP	FEV_1/FVC	FVC	Distance Walked
%Low Attenuation Level (-950)	0.44	0.61	0.03	0.07
BOW (K-means)	0.55	0.66	**0.48**	0.19
Fully Unsupervised (Template)	-1.16	-9.40	-518.92	-20.47
Proposed Method ($\boldsymbol{t}_s$)	**0.58**	**0.69**	0.38	**0.20**

TABLE II: Performance of predicting FEV_1 PP, FEV_1/FVC, FVC, and distance walked compared across *Bag-Of-Words (BOW)*, unsupervised topic model, our method ($\boldsymbol{t}_s$), and *% Low Attenuation Level (-950) (classic)* subject-level descriptors using ridge regression. Our method outperforms the rest in almost all metrics. *Fully Unsupervised (Template)* reports prediction accuracy when using topic proportions inferred by the template explainer model, learned in a fully unsupervised fashion.

separate linear models to predict FEV_1 PP, FEV_1/FVC, and distance walk from different subject-level features. Table II reports regression accuracy (R^2) on predicting these different metrics; it shows that our subject-level representation, $\boldsymbol{t}_s$, outperforms the standard bag-of-words representation and is significantly better than using %LAA-950Insp.

b) ***Explainer Model****:* We trained both the fully unsupervised template model (*i.e.,* topic model) and our proposed supervised model, described in Sec. II-D. For both models, we set $K = 10$. After training the models, we compute the posterior mean of the subtype proportion (*i.e.,* $\mathbb{E}_q[\boldsymbol{\pi}_s|\mathcal{D}]$) on the test data for evaluation. These values are used to train linear regression models predicting the disease severity measures. The results are shown in Table II. The template topic model, *without* injected subject-level features $\boldsymbol{t}_s$, learns subtypes that are not predictive of disease severity. Table II shows that subject-level features, $\boldsymbol{t}_s$, are most predictive of disease severity. This motivates our approach, in Section II-D, for building a variational approximation to the template model using features $\boldsymbol{t}_s$. Since our model uses $\boldsymbol{t}_s$ for inference, our prediction performance is the same. Our inference algorithm (Section II-D) transforms $\boldsymbol{t}_s$ to compute $\mathbb{E}_q[\boldsymbol{\pi}_s|\mathcal{D}]$. If this transformed value is used for the prediction, R^2 of predicting FEV_1 PP and FEV_1/FVC are 0.42 and 0.58 respectively. The gap between these values and the performance of $\boldsymbol{t}_s$ is the cost we pay to gain interpretation, which is much better than the fully unsupervised method. This confirms that the *Explainer* model learns tissue subtypes that are relevant to disease prediction, and not simply capturing irrelevant image statistics in the subject CT's.

Subtype	h^2 (%)	SE (%)	p-value
1	**23.69**	**8.42**	**2.3e-03**
2	**23.37**	**8.29**	**1.8e-03**
3	5.83	7.92	2.2e-01
4	9.96	8.26	1.1e-01
5	≈ 0	8.17	5e-01
6	≈ 0	8.38	5e-01
7	8.37	8.48	1.7e-01
8	**18.74**	**8.34**	**1.1e-02**
9	1.46	8.00	4.3e-01
10	2.16	8.00	3.9e-01

TABLE III: Heritability of tissue subtypes. h^2 measures the fraction of phenotypic variance (*i.e.,* variance in subject subtype proportion) explained by the total genetic variance.

c) ***Genetic Heritability****:* To understand the genetic etiology of each subtype, we perform the so-called genetic heritability analysis. In brief, the genetic heritability analysis studies the correlation between a quantitive trait and genetic data by estimating the proportion of the variance explained by genetic random effects. The variance ratio (h^2) is estimated under a linear mixed effect model where the fixed effects are nuisance variables, and the random effect is the linear effect of the genotyped variants. The higher the h^2, the stronger the genetic contribution to the trait. For each subtype, we view the proportion as a quantitive trait and estimate h^2 using the Restricted Maximum Likelihood (REML) method using GCTA software [50]. We use age, gender, number of smoking packs per year, and the first six principal components of the genetic kinship matrix as nuisance parameters (fixed effect). The results are shown in Table III. *Subtype 1*, *2*, and *8* show significant heritability of approximately $18 - 24\%$, providing strong evidence that these subtypes are biologically driven. While *subtypes 1, 2* have the strongest negative correlation with FEV_1, *subtype 8* has the strongest positive correlation with the FEV_1.

d) ***Sensitivity to*** K*:* We investigate the sensitivity of the explainer model to the choice of the number of subtypes, K, which is the most important one amongst the hyperparameters. Figure 7 shows the results of training the explainer model for varying values of parameter K. We measure the explainer model's ability to explain the observed data (*i.e.,* image features of the lung) on the test set by computing the log-likelihood of the data under the model. Each point is an average over two separate training runs of the explainer model with random initialization. When the assumed number of subtypes is less than 10 the explainer model's performance suffers but for values ≥ 10 we see relatively stable performance. This suggests that our choice of 10 subtypes is a reasonable approximation of the number of image feature clusters.

IV. Discussion and Conclusion

In the context of machine learning for clinical application, accurate prediction and meaningful clinical interpretation are equally important. While sophisticated models achieve state-of-the-art prediction performance, they are challenging to interpret. On the other hand, explainable models tend to be too simple to have excellent predictive performance. In this

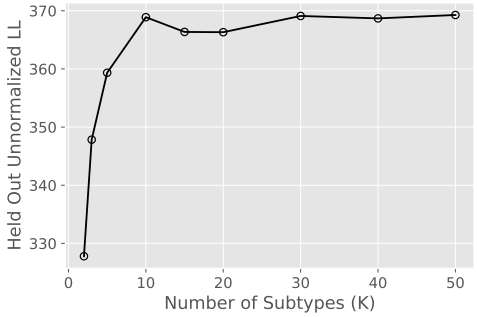


Fig. 7: Log-likelihood (LL) of explainer model on the held out set for different values of K. Each point is an average over two separate training runs of the explainer model with random initialization.

paper, we propose to use a probabilistic graphical model (PGM) to bridge this gap. Their expressive structure allows incorporating underlying domain knowledge (via a template). The approach lets the practitioner build a template for the explanation using knowledge about the disease. In this paper, we showed an application of our method for COPD, which is a highly heterogeneous disease. We viewed every patient as a mixture of different subprocesses; hence, a topic model is a proper template for the explanation. Our framework is general, and other choices of PGM are applicable depending on the application.

We showed that one could incorporate the discriminative information into the space of the posterior distributions to avoid loss of predictive performance. The idea is that the predictive model shares covariates relevant to prediction ($\boldsymbol{t}_s$) with the generative model. Therefore, they have the same prediction performance. We inject $\boldsymbol{t}_s$ into the approximation of the template's posterior distribution. To make the inference computationally efficient, we presented a specific transformation of $\boldsymbol{t}_s$ that results in a closed-form parameterization of the posterior distribution of the subtype proportion.

We apply our model on CT images of the COPDGene dataset. Our predictive model estimates subject dissimilarity using a non-parametric estimation of the KL divergence (see Section II-B) and converts it to a subject-level representation ($\boldsymbol{t}_s$). We compare our method with K-means as a standard Bag-of-Words method. One can view K-means as a simple graphical model which implicitly assumes a spherical Gaussian distribution for input features. Table I shows the effectiveness of our non-parametric method for predicting the severity of the disease. The table shows that our approach achieves the best prediction performance regardless of the input local image descriptor while there is significant variation in the performance of K-means. This result suggests that the implicit distribution assumption of K-means is not compatible with all choices of input features. However, our non-parametric model does not rely on any distributional assumption. The results in Table II validate the main idea of the paper, namely using $\boldsymbol{t}_s$ to build the posterior distribution. It shows that the vanilla topic modeling, which is fully unsupervised, completely loses discriminative power.

The main advantage of a PGM is that it is highly interpretable. The posterior probability of the different latent random variables in our model provides insight into the disease. Figures III-A visualizes the population-level and subject-level distributions of the subtypes. However, not all inferred subtypes aligned with the current clinical understanding of the disease; (e.g., subtypes six, seven, and ten). The fact that subtype ten is positively correlated with FEV_1 suggests that it represents healthy tissue. We observed that the proportion of subtype five is higher in the PRISm sub-population than the rest of the population (Figure 6). This is an exciting area for further investigation since the PRISm patients are difficult to characterize. However, this subtype does not show a significant correlation with the genetic data. Interestingly, the most significant subtypes in term of genetic heritability are the ones with the strongest correlation with FEV_1. Understanding the biological etiology of those subtypes requires further causal analysis, which is another avenue for future research.

In this paper, we applied our model on pre-engineered local image descriptors. However, the proposed framework is general and can be applied to interpret deep learning models. For example, $\boldsymbol{t}_s$ can be the output of the last fully connected layer in a Convolutional Neural Network (CNN). Currently, we use a generative model to interpret a pre-trained discriminative model. An exciting direction for future research is to train both models in an end-to-end fashion. One can use the genetic relatedness matrix as an extra covariate to the explainer model. Lastly, we would like to use the subtypes that are heritable and related to the disease as phenotypic traits for genetic association studies.

REFERENCES

[1] H. Ay, "Advances in the diagnosis of etiologic subtypes of ischemic stroke," *Current neurology and neuroscience reports*, vol. 10, pp. 14–20, 01 2010.

[2] F. Blows and et al., "Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: A collaborative analysis of data for 10,159 cases from 12 studies," *PLoS Medicine*, vol. 7, no. 5, 2010.

[3] S. Saria and A. Goldenberg, "Subtyping: What it is and its role in precision medicine," *IEEE Intelligent Systems*, vol. 30, no. 4, pp. 70–75, July 2015.

[4] K. Johnson, S. Khader, B. Glicksberg, B. Readhead, P. Sengupta, J. L.M. Bjrkegren, J. Kovacic, and J. T. Dudley, "Enabling precision cardiology through multiscale biology and systems medicine," *JACC: Basic to Translational Science*, vol. 2, pp. 311–327, 06 2017.

[5] P. J. Castaldi, M. Benet, H. Petersen, N. Rafaels, J. Finigan, M. Paoletti, H. Marike Boezen, J. M. Vonk, R. Bowler, M. Pistolesi, M. A. Puhan, J. Anto, E. Wauters, D. Lambrechts, W. Janssens, F. Bigazzi, G. Camiciottoli, M. H. Cho, C. P. Hersh, K. Barnes, S. Rennard, M. P. Boorgula, J. Dy, N. N. Hansel, J. D. Crapo, Y. Tesfaigzi, A. Agusti, E. K. Silverman, and J. Garcia-Aymerich, "Do copd subtypes really exist? copd heterogeneity and clustering in 10 independent cohorts," *Thorax*, vol. 72, no. 11, pp. 998–1006, 2017.

[6] X. Chen, X. Xu, and F. Xiao, "Heterogeneity of chronic obstructive pulmonary disease: from phenotype to genotype." *Frontiers of medicine*, vol. 7, no. 4, pp. 425–32, 12 2013.

[7] G. Viegi, F. Pistelli, D. L. Sherrill, S. Maio, S. Baldacci, and L. Carrozzi, "Definition, epidemiology and natural history of copd," *European Respiratory Journal*, vol. 30, no. 5, pp. 993–1013, 2007.

[8] M. Decramer, W. Janssens, and M. Miravitlles, "Chronic obstructive pulmonary disease," *The Lancet*, vol. 379, no. 9823, pp. 1341 – 1351, 2012.

[9] World Health Organization, "The top 10 causes of death," , 5 2018, [Online; accessed 12-June-2018].

[10] S. D. Shapiro, "Evolving concepts in the pathogenesis of chronic obstructive pulmonary disease," *Clin Chest Med*, vol. 21, no. 4, pp. 621–632, 2000.

[11] Y. S. Park, J. B. Seo, N. Kim, E. J. Chae, Y. M. Oh, S. D. Lee, Y. Lee, and S.-H. Kang, "Texture-based quantification of pulmonary emphysema on high-resolution computed tomography: comparison with density-based quantification and correlation with pulmonary function test." *Investigative radiology*, vol. 43 6, pp. 395–402, 2008.

[12] J. C. Ross, P. J. Castaldi, M. H. Cho, J. Chen, Y. Chang, J. G. Dy, E. K. Silverman, G. R. Washko, and R. S. J. Estépar, "A bayesian nonparametric model for disease subtyping: Application to emphysema phenotypes," *IEEE Transactions on Medical Imaging*, vol. 36, pp. 343–354, 2016.

[13] J. Song, J. Yang, B. M. Smith, P. P. Balte, E. A. Hoffman, R. G. Barr, A. F. Laine, and E. D. Angelini, "Generative method to discover emphysema subtypes with unsupervised learning using lung macroscopic patterns (lmps): The mesa copd study," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 375–378, 2017.

[14] J. Yang, E. D. Angelini, P. P. Balte, E. A. Hoffman, J. H. M. Austin, B. M. Smith, J. Song, R. G. Barr, and A. F. Laine, "Unsupervised discovery of spatially-informed lung texture patterns for pulmonary emphysema: The mesa copd study," *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 10433, pp. 116–124, 2017.

[15] Y. Hme, E. D. Angelini, M. A. Parikh, B. M. Smith, E. A. Hoffman, R. G. Barr, and A. F. Laine, "Sparse sampling and unsupervised learning of lung texture patterns in pulmonary emphysema: Mesa copd study," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, April 2015, pp. 109–113.

[16] R. Uppaluri, T. Mitsa, M. Sonka, E. Hoffman, and G. McLennan, "Quantification of pulmonary emphysema from lung computed tomography images," *American Journal of Respiratory and Critical Care Medicine*, vol. 156, no. 1, pp. 248–254, 1997, pMID: 9230756.

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

[18] A. Depeursinge, D. Sage, A. Hidki, A. Platon, P. Poletti, M. Unser, and H. Muller, "Lung tissue classification using wavelet frames," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2007, pp. 6259–6262.

[19] M. Prasad, A. Sowmya, and P. Wilson, "Multi-level classification of emphysema in hrct lung images," *Pattern Analysis and Applications*, vol. 12, no. 1, pp. 9–20, Feb 2009.

[20] P. Binder, N. K. Batmanghelich, R. S. J. Estepar, and P. Golland, "Unsupervised Discovery of Emphysema Subtypes in a Large Clinical Cohort," in *Machine Learning in Medical Imaging: 7th International Workshop, MLMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings*, L. Wang, E. Adeli, Q. Wang, Y. Shi, and H.-I. Suk, Eds. Cham: Springer International Publishing, 2016, pp. 180–187.

[21] Z. Aziz, A. U. Wells, D. M. Hansell, G. A. Bain, S. J. Copley, S. R. Desai, S. M. Ellis, F. V. Gleeson, S. Grubnic, A. G. Nicholson, S. P. G. Padley, K. S. Pointon, J. H. Reynolds, R. Robertson, and M. Rubens, "Hrct diagnosis of diffuse parenchymal lung disease: inter-observer variation." *Thorax*, vol. 59 6, pp. 506–11, 2004.

[22] P. J. Castaldi, M. Benet, H. Petersen, N. Rafaels, J. Finigan, M. Paoletti, H. M. Boezen, J. M. Vonk, R. Bowler, M. Pistolesi *et al.*, "Do copd subtypes really exist? copd heterogeneity and clustering in 10 independent cohorts," *Thorax*, vol. 72, no. 11, pp. 998–1006, 2017.

[23] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[26] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014.

[28] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II–1278–II–1286.

[29] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, June 2005, pp. 524–531 vol. 2.

[30] B. Poczos and J. Schneider, "On the estimation of alpha-divergences," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. Dudk, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 11–13 Apr 2011, pp. 609–617.

[31] J. Schabdach, W. M. Wells, M. Cho, and K. N. Batmanghelich, "A likelihood-free approach for characterizing heterogeneous diseases in large-scale studies," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 170–183.

[32] L. Song, S. M. Siddiqi, G. Gordon, and A. Smola, "Hilbert Space Embeddings of Hidden Markov Models," in *The 27th International Conference on Machine Learning (ICML2010)*, 2010, pp. 991–998.

[33] Z. Zhang and J. Wang, "MLLE: Modified Locally Linear Embedding Using Multiple Weights," *Advances in Neural Information Processing Systems*, pp. 1593–1600, 2006.

[34] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[35] M. Bryant and E. B. Sudderth, "Truly nonparametric online variational inference for hierarchical dirichlet processes," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 2699–2707.

[36] M. Johnson and A. Willsky, "Stochastic Variational Inference for Bayesian Time Series Models," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1854–1862.

[37] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, 01 2016.

[38] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, Nov 1999.

[39] C. Peterson and J. R. Anderson, "A mean field theory learning algorithm for neural networks," *Complex Systems*, vol. 1, pp. 995–1019, 1987.

[40] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[41] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint*, no. Ml, pp. 1–14, 2013.

[42] S. W. Linderman, M. J. Johnson, and R. P. Adams, "Dependent multinomial models made easy: Stick breaking with the pólya-gamma augmentation," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 3456–3464.

[43] E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo, "Genetic epidemiology of COPD (COPDGene) study design," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 7, no. 1, pp. 32–43, 2011.

[44] M. Holzer and R. Donner, "Over-Segmentation of 3D Medical Image Volumes based on Monogenic Cues," *Cvww*, no. JANUARY 2014, pp. 35–42, 2014.

[45] S. B. Shaker, M. D. Bruijne, L. Sorensen, S. B. Shaker, and M. De Bruijne, "Quantitative analysis of pulmonary emphysema using local binary patterns," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 2, pp. 559–569, 2010.

[46] L. Sorensen, M. Nielsen, P. Lo, H. Ashraf, J. H. Pedersen, and M. De Bruijne, "Texture-based analysis of COPD: A data-driven approach," *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 70–78, 2012.

[47] W. D. Vogl, H. Prosch, C. Muller-Mang, U. Schmidt-Erfurth, and G. Langs, "Longitudinal alignment of disease progression in fibrosing interstitial lung disease," in *Lecture Notes in Computer Science*, vol. 8674 LNCS, no. PART 2, 2014, pp. 97–104.

[48] K. Liu, H. Skibbe, T. Schmidt, T. Blein, K. Palme, T. Brox, and O. Ronneberger, "Rotation-Invariant HOG Descriptors Using Fourier Analysis in Polar and Spherical Coordinates," *International Journal of Computer Vision*, vol. 106, no. 3, pp. 342–364, 2014.

[49] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172.

[50] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, and Others, "Common {SNPs} explain a large proportion of the heritability for human height," *Nat Gen*, vol. 42, no. 7, pp. 565–569, 2010.