Transfer learning for multi-center classification of chronic obstructive pulmonary disease

Veronika Cheplygina, Isabel Pino Peña, Jesper Holst Pedersen, David A. Lynch, Lauge Sørensen, and Marleen de Bruijne

***Abstract*—Chronic obstructive pulmonary disease (COPD) is a lung disease which can be quantified using chest computed tomography (CT) scans. Recent studies have shown that COPD can be automatically diagnosed using weakly supervised learning of intensity and texture distributions. However, up till now such classifiers have only been evaluated on scans from a single domain, and it is unclear whether they would generalize across domains, such as different scanners or scanning protocols. To address this problem, we investigate classification of COPD in a multi-center dataset with a total of 803 scans from three different centers, four different scanners, with heterogenous subject distributions. Our method is based on Gaussian texture features, and a weighted logistic classifier, which increases the weights of samples similar to the test data. We show that Gaussian texture features outperform intensity features previously used in multi-center classification tasks. We also show that a weighting strategy based on a classifier that is trained to discriminate between scans from different domains, can further improve the results. To encourage further research into transfer learning methods for classification of COPD, upon acceptance of the paper we will release two feature datasets used in this study on http://TBA.**

***Index Terms*—Transfer learning, multiple instance learning, domain adaptation, importance weighting, computed tomography (CT), chronic obstructive pulmonary disease (COPD), lung**

I. Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by chronic inflammation of the lung airways and emphysema, i.e., degradation of lung tissue [1]. Emphysema can be visually assessed in vivo using chest computed tomography (CT) scans, however, to overcome limitations of visual assessment, automatic quantification of emphysema has been explored [2], [3], [4], [5], [6], [7]. Several of these methods rely on supervised learning and require manually annotated regions of interest (ROIs) [2], [3], [4], while other approaches using multiple instance learning (MIL) only require patient-level labels indicating overall disease status [5], [6], [7]. In this work we address this weakly-supervised setting, i.e., the scans are only labeled as belonging to a healthy or COPD subject, and no information on ROI level is available.

A challenge for classification of COPD in practice is that the training data may not be representative of the test data, i.e. the distributions of the training and the test data are different. This can happen if the data originates from different *domains*, such as different subject groups, scanners, or scanning protocols. One approach to overcome this problem is to search for features that are robust to such variability. For example, in a multi-cohort study with different CT scanners [4], the authors compare intensity distribution features to local binary pattern (LBP) texture features, and suggest that intensity might be more effective in multi-scanner situations.

Another way to explicitly address the differences in the distributions of the training and test data is called *transfer learning* [8] or domain adaptation. These differences can be caused by different marginal distributions $P(\mathbf{x})$, different labeling functions $P(y|\mathbf{x})$ or even different feature and label spaces. Based on these differences, different transfer learning scenarios can be distinguished. One of these scenarios is transductive transfer learning, where labeled training data (or source data), as well as unlabeled test data (or target data), are assumed to be available. This is the scenario we investigate.

Transfer learning methods can be divided into instance-transfer approaches and feature-transfer approaches. This paper presents an instance-transfer approach, but we briefly discuss both approaches to contrast our work from the literature. In short, feature-transfer approaches aim to find features which are good for classification, possibly in a different classification problem. In contrast, instance-transfer methods aim to select source samples which help the classifier to generalize well. An intuitive instance-transfer approach is importance weighting [9], [10], [11], i.e., assigning weights to the source samples, based on their similarity to the unlabeled target samples, and subsequently training a weighted classifier. This strategy assumes that only the marginal distributions are different, and that the labeling functions are the same. However, in practice, importance weighting can also be beneficial in cases where the labeling functions are different [12].

Transfer learning techniques are relatively new in the medical imaging domain, and have shown to be successful in several applications, such as classification of Alzheimer's disease [13], [14] and segmentation of magnetic resonance (MR) images [12], [15], [16] and microscopy images [17],

V. Cheplygina was with the Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology of the Erasmus MC - University Medical Center Rotterdam, Rotterdam, The Netherlands when his work was performed. She is now with the Medical Image Analysis group, Eindhoven University of Technology. E-mail: v.cheplygina@tue.nl

I. Pino Peña is with the Department of Health Science and Technology, Aalborg University, Aalborg, Denmark.

J. H. Pedersen is with the Department of Thoracic Surgery, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark.

D. A. Lynch is with the Department of Radiology, National Jewish Health, Denver, Colorado, United States of America.

L. Sørensen and M. de Bruijne are with the Image Section, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark.

M. de Bruijne is also with the Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology of the Erasmus MC - University Medical Center Rotterdam, Rotterdam, Netherlands.

[18]. In chest CT scans, transfer learning has been used for classification of different abnormalities in lung tissue [19], [20]. However, these approaches focus on feature-transfer between datasets, possibly even from non-medical datasets, while we investigate an instance transfer approach which focuses on differences between data acquired at different sites. To the best of our knowledge, our work is the first to investigate the use of transfer learning for classification of COPD.

The contributions of this paper are twofold. Our first contribution is a comparison of different types of intensity- and texture-based features for the task of classifying COPD in chest CT scans, to assess the features' robustness across scanners. The second contribution is a proposed approach which combines transfer learning with a weakly-supervised classifier. To this end, we investigate three different weighting strategies. We use four datasets, which differ with respect to the subject group, site of collection, scanners and scanning protocols used. Furthermore, we publicly release two feature datasets used in this study to further the progress in transfer learning in classification of COPD and in medical image analysis in general.

II. Methods

Following Sørensen et al. [5], we represent each chest CT image by a set of 3D ROIs. Each ROI is represented by a feature vector describing the intensity and/or texture distribution in that ROI. In order to classify each individual test scans, we assign weights to the training scans based on their similarity to the single test scan, and subsequently train a weighted multiple instance classifier. The procedure is illustrated in Fig. 1.

A. Notation and Feature Representation

Each scan is represented by a bag $X_i = \{\mathbf{x}_{ij} | j = 1, ..., n_i\} \subset \mathbb{R}^d$ of n_i instances (ROIs), where the j-th instance or ROI is described by a m-dimensional feature vector $\mathbf{x}_{ij}$. The bags have labels $y_i \in \{+1, -1\}$, in our case COPD and non-COPD, but the instances are unlabeled: the problem is thus called weakly supervised. The bags originate from two different datasets: training (source) $\mathcal{X}$ and test (target) data $\mathcal{Z}$. We will denote bags and instances from the source data by X and $\mathbf{x}$, bags and instances from the target data are denoted by Z and $\mathbf{z}$.

We represent each CT scan by a bag of 50 possibly overlapping, volumetric ROIs of size $41 \times 41 \times 41$ voxels, extracted at random locations inside the lung mask. The lung masks were obtained prior to this study. For three datasets (DLCST and both COPDGene datasets), the lung masks were obtained with a region-growing algorithm and postprocessing step used in [21], and for one dataset (Frederikshavn) with a method based on multi-atlas registration and graph cuts, similar to [22].

We use Gaussian scale space (GSS) features, which capture the image texture, to represent the ROIs. Each image is first convolved with a Gaussian function at scale σ using normalized convolution within the lung mask. We use four different scales, $\{0.6, 1.2, 2.4, 4.8\}$ mm, and compute eight filters: smoothed image, gradient magnitude, Laplacian of Gaussian, three eigenvalues of the Hessian, Gaussian curvature and eigen magnitude. The filtered outputs are summarized with histograms, where adaptive binning [23] is used to best describe the data while reducing dimensionality. We quantize the output of each filter into ten bins, where the bin edges used for adaptive binning (i.e. volume in each bin must be equal) of all datasets have been determined on an independent sample from one of the datasets (DLCST). This leads to in $8 \times 4 \times 10 = 320$ features in total.

B. Classifier

To learn with weakly labeled scans, we use a MIL classifier. A straightforward approach called SimpleMIL propagates the training bag labels to the training instances, and trains an instance classifier. For a test bag, a label is obtained by classifying that bag's instances, and combining the instance posteriors. Here we apply the average rule,

$$\frac{p(y_i = +1|X_i)}{p(y_i = -1|X_i)} = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{p(y_{ij} = +1|\mathbf{x}_{ij})}{p(y_{ij} = -1|\mathbf{x}_{ij})}, \quad (1)$$

which assumes that all instances contribute to the bag label. Despite its simplicity, this strategy has achieved good results in previous experiments on weakly-labeled single-domain chest CT data [6], [5]. In [6], this method was used with a logistic and a nearest neighbor classifier, and the logistic classifier achieved the best performance. Here we therefore use SimpleMIL with a weighted logistic classifier. The logistic classifier is a linear classifier $\mathbf{w}^*$, defined as follows:

$$\mathbf{w}^* = \text{argmin}_{\mathbf{w}} \Big(\sum_{(\mathbf{x}_{ij}, y_{ij})} s_{ij} L(\mathbf{w}, \mathbf{x}_{ij}, y_{ij}) + \lambda ||\mathbf{w}||_2^2 \Big), \quad (2)$$

where $\mathbf{w}$ is a vector of m feature coefficients (we drop the intercept for ease of notation), the loss is defined as $L(\mathbf{w}, \mathbf{x}, y) = \frac{1}{\ln 2} \ln (1 + \exp(-y\mathbf{w}^{\intercal}\mathbf{x}))$, λ is a regularization term controlling the complexity of the weight vector, and s_{ij} is the importance weight associated with the j-th instance from the i-th bag.

To make sure that the total volume of weights is the same across different weighting strategies, before training the classifier we multiply the weights by $N / \sum_{i,j} s_{ij}$, such that the sum of the weights is equal to the number of training instances N.

C. Instance Weighting

We estimate the weights of the source bags with three different weight measures:

- using the distance from the source bags to the target bag
- using the distance from the target bag to the source bags
- using the estimated probability of the source bag belonging to the target class

In the traditional instance weighting approach, the weights are assigned to instances, which are considered independent. However, for MIL, this is not the most intuitive approach. Rather than finding similar instances in the training data, we

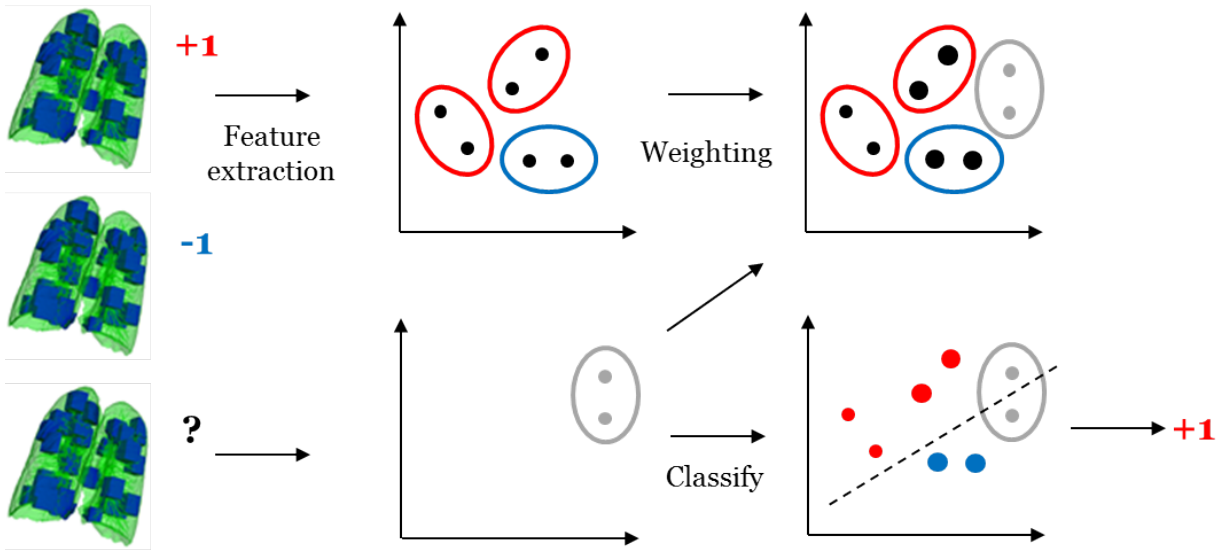

Fig. 1. Overview of the procedure. Each scan is represented as a bag of instances (feature vectors). The bag is labeled as COPD (+1) or healthy (-1). At test time, the bags are weighted by their similarity to the unlabeled (?) test scan. The weighted bags are used to train a classifier, which in turn is applied to classify the test instances. The outputs are aggregated into an overall scan classification.

are more interested in similar bags. Therefore, we want to assign the weights on bag-level, therefore in what follows we describe how to obtain a weight s_i for each bag X_i. In training the SimpleMIL classifier, however, each instance is associated with a weight $s_{ij} = s_i$.

1) Source to target weights: The first approach is based on a bag distance between the source bag, and the target bag. We use weights that are inversely proportional to the source-to-target (*s2t*) distance of source bag X_i to a target bag Z. In converting the distances to weights, we scale the weights to the interval [0,1], which assumes that there are always relevant and irrelevant source samples. The *s2t* weights are then defined as follows:

$$s_i^{s2t} = \frac{d_{max}^{s2t} - d_i^{s2t}}{d_{max}^{s2t} - d_{min}^{s2t}} \tag{3}$$

where

$$d_i^{s2t} = \frac{1}{|X_i|} \sum_{\mathbf{x}_{ij} \in X_i} \min_{\mathbf{z}_k \in Z} ||\mathbf{x}_{ij} - \mathbf{z}_k||^2. \tag{4}$$

and $d_{max} = \max_i d_i$ and $d_{min} = \min_i d_i$ are the maximum and minimum bag distances found in the training set.

In other words, for each instance in the source bag, we find its nearest neighbor in the target bag Z, and average the nearest neighbor distances. A divergence measure that is analogous to this distance has been successfully used in previous works on transfer learning in medical image analysis [12], [15]. The distance we propose is more efficient to compute, and has shown to be robust in high-dimensional situations [24], [25] than related divergences.

2) Target to source weights: The matching of instances with their nearest neighbors makes the bag distance asymmetric. In previous work on medical imaging such asymmetry was important for classification performance [25]. The rationale is that for a test scan with unusual ROIs (i.e., outliers in feature space), we want to ensure that these outliers influence the training weights as much as possible. However, with the *s2t* distance, it is possible that the test outliers do not participate in the weighting process at all. Therefore we also examine weights based on the counterpart of the source-to-target distance, i.e. the target-to-source (*t2s*) distance:

$$s_i^{t2s} = \frac{d_{max}^{t2s} - d_i^{t2s}}{d_{max}^{t2s} - d_{min}^{t2s}} \tag{5}$$

where

$$d_i^{t2s} = \frac{1}{|Z|} \sum_{\mathbf{z}_k \in Z} \min_{\mathbf{x}_{ij} \in X_i} ||\mathbf{x}_{ij} - \mathbf{z}_k||^2 \tag{6}$$

and d_{max}^{t2s} and d_{min}^{t2s} are defined analogously to d_{max}^{s2t} and d_{min}^{s2t}.

Note that the fact that we can use the *t2s* distance for weighting relies on the fact that we are computing bag distances. If we would weight the training instances independently, some of the training instances might not get matched with target instances, and therefore might not receive a weight.

3) Logistic weights: The last weighting approach is based on how well a logistic classifier $\mathbf{w}^s$, which models posterior probabilities, can separate the source and target data. That is, all the instances in the source data are labeled as class -1, and samples in the target data are labeled as class 1, and the classifier $\mathbf{w}^s$ is trained on these two classes. The source samples are then evaluated by the classifier to obtain their probabilities of belonging to the target class $p(y = 1|\mathbf{x}_{ij}) = \exp{(-\mathbf{w}^{s\intercal}\mathbf{x}_{ij})} / \sum_{y_{ij} in \{-1,+1\}} \exp{(-y_{ij}\mathbf{w}^{s\intercal}\mathbf{x}_{ij})}$. For a training bag, we therefore have the following:

$$s_i^{log} = \frac{1}{|X_i|} \sum_{\mathbf{x}_{ij} \in X_i} \frac{\exp{(-\mathbf{w}^{s\intercal}\mathbf{x}_{ij})}}{\sum_{y_i \in \{-1,+1\}} \exp{(-y_i\mathbf{w}^{s\intercal}\mathbf{x}_{ij})}} \tag{7}$$

This approach is common in transfer learning literature in the field of machine learning [26], and, in the infinite-sample case and no change in labeling function, has shown to be equivalent to a classifier trained on the source samples [9]. In medical image analysis, this approach has been used for segmentation of tumors in brain MR images [27] for a domain adaptation setting where only the sampling of the training and test data is different.

III. Experiments

A. Data

We use four datasets from different scanners in the experiments (Table I). The first dataset consists of 600 baseline inspiratory chest CT scans from the Danish Lung Cancer Screening Trial (DLCST) [28]. The second (120 inspiratory scans) and third (67 inspiratory scans) datasets consist of subjects from the COPDGene study [29], both acquired at the National Jewish Center in Denver, Colorado. The fourth dataset (16 scans) consists of subjects with respiratory problems referred to the out-patient clinic of the Frederikshavn hospital in Denmark.

All scans are acquired at full inspiration, and the COPD diagnosis is determined according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria [30], i.e., $FEV_1/FVC < 0.7$.

We consider DLCST and the two COPDGene datasets both as source data and as target data, and the Frederikshavn only as the target data, due to its small size.

B. Feature Datasets

In the proposed approach, each ROI is represented by GSS as described in Section II-A, resulting in a feature vector with 320 dimensions. We compare our method with intensity features based on kernel density estimation (KDE) used in [4]. We use 256 bins in order for the dimensionality to be comparable to the Gaussian features. To focus on the more informative part of the intensities, we apply the KDE to the range [-1100 -600]. We originally used a larger range and 4096 bins, following correspondence with the authors of [4]. However, this gave poor results in preliminary experiments on DLCST and Fredrikshavn data. We concluded that the classifier suffered from overfitting, and adapted the range and dimensionality to produce reasonable results for those two datasets.

Furthermore, we compare our feature set to two of its subsets: a subset with 40 features describing the intensity of the scan at different scales (GSS-i), and its complement with 280 features describing with derivatives only, thus more describing the texture (GSS-t). These comparisons will allow us to evaluate whether it is the intensity information that is responsible for differences between GSS and KDE, or the particular choice of implementation used in KDE.

C. Classifiers without Transfer Learning

We first use SimpleMIL with a logistic classifier without any weighting. We train classifiers on each of the three source datasets (D, C1 and C2). We then apply the trained classifiers to the four target datasets (D, C1, C2 and F). When the source and target datasets are the same, this experiment is performed in a leave-one-scan-out procedure.

The logistic classifier has only one free parameter, the regularization parameter λ. For both $\mathbf{w}^*$ (the SimpleMIL classifier) and $\mathbf{w}^s$ (the classifier used to determine the logistic weights) we fix $\lambda = 1$, because in preliminary experiments choosing other values did not have a large effect on the results.

D. Classifiers with Transfer Learning

We then use SimpleMIL with a weighted logistic classifier. For each of the nine combinations of source and different-domain target datasets, we perform a leave-one-image-out procedure. For each target image, we determine the weights using three different methods: *s2t*, *t2s* and logistic. We then train the weighted classifiers and evaluate them on the target image. Again, the regularization parameter λ is fixed to 1.

E. Evaluation

The evaluation metric is the area under the receiver-operating characteristic curve (ROC), or AUC. We test for significant differences using the DeLong test for ROC curves [31].

To summarize results over nine pairs of source and target data, we also rank the different weighting methods and different feature methods, and report the average ranks. To assess significance, we perform a Friedman/Nemenyi test [32] at the $p = 0.05$ level. This test first checks whether there are any significant differences between the ranks, and if so, determines the minimum difference in ranks (or critical difference) required for any two individual differences to be significant. For nine pairs of datasets and four methods, the critical difference is 1.56.

IV. Results

A. Performance without Transfer Learning

Fig. 2 shows the results of different features for the SimpleMIL logistic classifier, without using any transfer learning. In this section we summarize the results per test dataset.

For DLCST, the best results are obtained when training within the same dataset using GSS or GSS-t features. The AUCs are not very high compared to those of other datasets, but they are consistent with previous results on DLCST [6], [5].

On the COPDGene datasets we obtain much higher AUCs. When training on one dataset and testing on the other, the performances are similar to when training within a single dataset, suggesting the protocol was well-standardized and that using a slightly different scanner did not have a large effect on the scans. In this scenario, all features give good results, with GSS-i being slightly better than the others. However, when training on a very different dataset, DLCST, the situation changes: the best results are still provided by GSS-i, but the gap between GSS-i and the other features now increases. In particular, the performance of KDE-i drops dramatically.

The Frederikshavn dataset also can be classified well, but the success is more dependent on the dataset and the features

Dataset	Subjects	Age	GOLD (1/2/3/4)	Smoking (c/f/n)	Scanner	Resolution (mm)	Exposure	Reconstruction
DLCST (D)	300 +	59 [50, 71]	69/28/2/0	77/23/0	Philips	0.72×0.72×1 to	40 mAs	Philips D
	300 -	57 [49, 69]		74/26/0	16 rows Mx 8000	0.78×0.78×1		hard
COPDGene 1 (C1)	74 +	64 [45, 80]	21/18/19/16	17/57/0	Siemens	0.65×0.65×0.75	200 mAs	B45f sharp
	46 -	59 [45, 78]		23/20/3	Definition			
COPDGene 2 (C2)	42 +	65 [45, 78]	9/13/7/13	12/30/0	Siemens	0.65×0.65×0.75	200 mAs	B45f sharp
	25 -	60 [47, 78]		9/11/5	Definition AS+			
Frederikshavn (F)	8 +	66 [48, 77]	1/3/3/1	1/7/0	Siemens	0.58×0.58×0.6	95 mAs	I70f very sharp
	8 -	56 [25, 73]		1/2/5	Definition Flash			

TABLE I
DETAILS OF DATASETS. FOR SUBJECTS, + = COPD, - = NON-COPD. AGES REPORTED AS MEAN [MIN, MAX], ROUNDED TO NEAREST INTEGER. GOLD REFERS TO THE COPD STAGE AS DEFINED BY THE GLOBAL INITIATIVE FOR CHRONIC OBSTRUCTIVE LUNG DISEASE. FOR SMOKING STATUS, C=CURRENT, F=FORMER, N=NEVER.

used, than is the case for COPDGene. The best performances on Frederikshavn are obtained with GSS-t features, followed by GSS. The two types of intensity features perform the worst, with GSS-i doing slightly better.

B. *Performance with Transfer Learning*

We now examine the performances of the importance-weighted classifiers. The performances are shown in Table II for completeness, but for better interpretation of the numbers, a summary is provided in Table III. In total we considered nine across-domain experiments. Averaged over these nine experiments, we report the AUC, the rank of each weighting method (per feature), and the rank of each feature (per weighting method).

The average AUCs do not give a conclusive answer about whether weighting is beneficial. For both types of intensity features, weighting always improves performance, but for GSS and GSS-t weights can also deteriorate the performance slightly. The best results are obtained with the logistic weights, which improves average performance for all feature types.

The average ranks for the weights, per feature, tell a slightly different story, although here almost none of the differences are significant. For GSS, none of the weighting methods rank higher than the unweighted case. For the other features, it is always beneficial to do some form of weighting, but the best method varies per feature. In general, the differences between the ranks are quite small and not significant. The only exception is GSS-i, where *s2t* has a much better rank than the other methods, and it is also the only feature for which any significant differences in ranks are found.

When comparing the ranks of the features, the differences are much larger. Now significant differences are found for each weight type. GSS features are clearly the best overall, with ranks close to 1, followed by GSS-i and GSS-t (although these differences are not significant), and KDE-i are the worst, with ranks close to 4. The last difference is significant for all weighting strategies.

V. DISCUSSION

The main findings from the previous section are: (i) there are large differences between datasets, (ii) there are large differences between features, and (iii) weighting, in particular with logistic classifier-based weights, can improve performance. In this section we discuss these results in more detail. We then discuss limitations of our method, and provide some recommendations for classification of COPD in multi-center datasets.

	gss	gss-t	gss-i	kde-i
none	82.6	83.6	79.7	71.8
s2t	82.1	83.4	**81.0**	73.3
t2s	82.3	83.8	80.2	73.8
log	**83.1**	**84.4**	80.2	**73.9**
none	**2.11**	**2.78**	3.17	**3.06**
s2t	**2.78**	**2.72**	**1.50**	**2.78**
t2s	**2.78**	**2.61**	**2.67**	**2.06**
log	**2.33**	**1.89**	**2.67**	**2.11**
none	1.67	2.22	2.33	3.78
s2t	1.78	2.44	2.00	3.78
t2s	1.72	2.22	2.33	3.72
log	1.67	2.22	2.44	3.67

TABLE III
TOP: AVERAGE AUC, IN PERCENTAGE, OVER NINE TRANSFER EXPERIMENTS. BEST WEIGHTING METHOD IS IN BOLD, BEST FEATURE IS UNDERLINED. MIDDLE: RANKS OF EACH WEIGHT TYPE (1=BEST, 4=WORST), COMPARE PER COLUMN. BEST WEIGHT, OR WEIGHTS THAT ARE NOT SIGNIFICANTLY WORSE (FRIEDMAN/NEMENYI TEST, CRITICAL DIFFERENCE = 1.56) ARE IN BOLD. BOTTOM: RANKS OF EACH FEATURE TYPE (1=BEST, 4=WORST), COMPARE PER ROW. BEST FEATURE, OR FEATURES THAT ARE NOT SIGNIFICANTLY WORSE (FRIEDMAN TEST, CRITICAL DIFFERENCE = 1.56) ARE UNDERLINED.

A. *Datasets*

The results of COPD classification trained on data from within the same domain tell us something about the differences between datasets. In particular, DLCST is more difficult to classify than the other datasets. The highest AUC for DLCST is 0.79 (when training on the same domain), whereas the AUCs for the other datasets are often higher than 0.9. This difference can be explained by the differences in COPD severity between datasets. DLCST contains many cases of mild COPD, which can be easily misclassified as healthy subjects. The other datasets contain more severe cases of COPD. This is supported by the fact that, if we remove the GOLD 2-4 subjects from the COPDGene datasets, the AUC decreases to around 0.8 for the best features.

The datasets have different sizes, which can also affect the results of the classifiers. When the training and test data are from the same or similar domain, the training dataset should

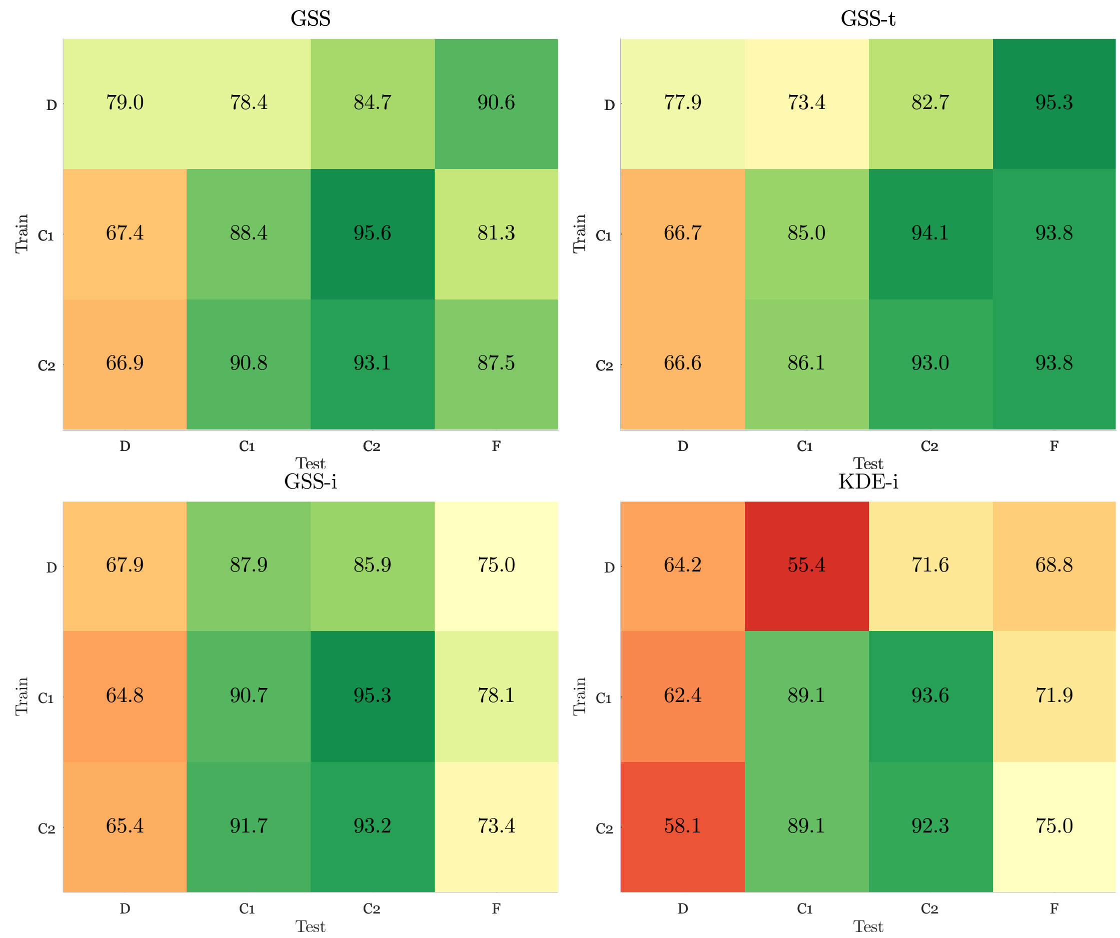

Fig. 2. AUC ×100 of SimpleMIL across datasets, without transfer, for four different feature types. Three datasets (rows) are used for training and four datasets (columns) are used for testing. Diagonal elements (for D, C1 and C2) show leave-one-out performance within a single dataset.

be sufficiently large to describe all possible variations. As a result, when testing on COPDGene 2, it is actually better to train on COPDGene 1 (which has similar scans, but is larger than COPDGene2), than to do same-domain training on COPDGene 2. Another example is Frederikshavn: since both DLCST and the COPDGene datasets are rather dissimilar, the larger DLCST training data tends to give better results. As such, it would be interesting to compare results of different methods, when sampling the same number of training scans from each dataset.

B. Features

Our results show that intensity is not always a robust choice of features when classifying across domains. Gaussian scale space features, which combine intensity and texture components, had higher performances overall, and in some cases, the intensity components could even deteriorate the performance.

These findings are interesting with respect to previous results from the literature. On a task of classifying ROIs within a single domain, [3] showed that local binary pattern (LBP) texture features combined with intensity features can give good classification performance. However, [4] showed that intensity features alone performed better than a different implementation of LBP in across domain classification.

We note that there are several differences between [4] and the current study. We focus on weakly-supervised classification of entire chest CT scans, whereas [4] deals with a multi-class ROI classification problem. Furthermore, in our transfer learning experiments the training and test domains are disjoint, i.e., the classifier does not have access to any labeled data from the test domain. On the other hand, in [4] scans from the same domain are present in the training set. Combined with their use of the nearest neighbor classifier, this could enable intensity features to perform well even if intensities are different across domains. A further difference is that to avoid overfitting, we reduced the dimensionality of the intensity representation.

C. Weights

Weighting can improve performance across domains, but does not guarantee improved performance. In our experiments, no weighting method was always (for each dataset combination and feature type) better than the unweighted baseline. However, on average the logistic classifier-based weights performed quite well. The logistic weights had the highest average performance on each of the four feature types, and the highest rank on three out of four features, although the differences were not significant.

	gss	gss-t	gss-i	kde-i	gss	gss-t	gss-i	kde-i	gss	gss-t	gss-i	kde-i
Train D	Test C1				Test C2				Test F			
none	**78.4**	**73.4**	**87.9**	**55.4**	**84.7**	**82.7**	85.9	**71.6**	**90.6**	**95.3**	**75.0**	**68.8**
s2t	**78.6**	**75.2**	**89.1**	55.6	**85.8**	**83.8**	**88.5**	**72.8**	**89.1**	**93.8**	**76.6**	**76.6**
t2s	77.0	**73.2**	86.3	**57.8**	**84.0**	**83.0**	86.0	**73.5**	**90.6**	**95.3**	76.6	**76.6**
log	**77.9**	**73.1**	**87.4**	**57.1**	**84.0**	**82.3**	**86.8**	**73.3**	**93.8**	**96.9**	**75.0**	**78.1**
Train C1	Test D				Test C2				Test F			
none	**67.4**	**66.7**	64.8	**62.4**	**95.6**	**94.1**	**95.3**	**93.6**	**81.3**	**93.8**	**78.1**	**71.9**
s2t	**67.0**	65.6	**65.8**	**62.1**	**95.7**	**95.0**	**95.0**	**93.4**	81.3	**95.3**	**79.7**	71.9
t2s	**67.2**	**66.3**	**65.8**	**62.0**	**96.2**	**94.5**	**95.8**	**93.5**	**79.7**	**92.2**	79.7	71.9
log	67.0	**66.7**	65.3	**62.1**	**95.5**	**94.6**	**95.2**	**93.4**	81.3	**96.9**	79.7	71.9
Train C2	Test D				Test C1				Test F			
none	66.9	66.6	**65.4**	58.1	**90.8**	**86.1**	**91.7**	**89.1**	**87.5**	**93.8**	**73.4**	**75.0**
s2t	65.4	62.7	**65.6**	**61.7**	**89.9**	**85.6**	**91.9**	**88.9**	**85.9**	**93.8**	**76.6**	**76.6**
t2s	**67.9**	66.3	65.2	**61.6**	**90.7**	**86.2**	**91.5**	**89.4**	**87.5**	**96.9**	**75.0**	**78.1**
log	**68.4**	**67.7**	**65.5**	**61.6**	**90.7**	**86.5**	**91.6**	**89.4**	**89.1**	**95.3**	**75.0**	78.1

TABLE II

AUC of SimpleMIL, in percentage. In each of the nine experiments, the AUCs are compared with a DeLong test for AUCs. Per column of 4 methods, **bold**: best or not significantly worse than best different-domain method. Per row of 4 features, underline: best or not significantly worse than best feature.

The small difference between *s2t* and *t2s*, the different ways in which source and target bags can be compared, is interesting. In a study of brain tissue segmentation across scanners [25], weighting trained *classifiers* based on the *t2s* distance was more effective than weighting them based on the *s2t* distance. We thus hypothesized that *t2s* might also be a better strategy for weighting training samples, but our results show that this is not the case.

To further understand the differences between the weights, we looked at the weights assigned to each training bag. In Fig. 3 we show the weights when training on D and testing on C2 for two of the feature types: GSS with 320 features and GSS-i with 40 features. In each case, we first find the mean and the standard deviation of the weights, assigned to each training bag. We then sort the training bags by their mean weight, and plot the mean and the standard deviations with error bars.

Per training bag, the distance-based weights have a higher variance than the logistic weights. Furthermore, with distance-based weights, the distributions are more steep, i.e. more training bags have a very low, or a very high average weight. Setting many weights (almost) to zero, as is the case for the distance-based weights, effectively decreases the sample size, possibly resulting in lower performance.

One of the reasons for this behavior is the way that the weights are scaled. With the logistic weights, the exponential function provides a more natural scaling of the weights. For example, if all the source bags are similar to the target bag, they will all receive similar weights. The scaling we apply for the distance-based weights is more "artificial", because the most similar bag is assumed to have weight 1, and the least similar bag is assumed to have weight 0. Furthermore, logistic weights are based on all the source bags, i.e., they are assigned by a classifier trained to distinguish the target bag from all the source bags. On the other hand, the distance-based weights are based only on the distance between the target bag and each individual source bag, which leads to noise.

In Fig. 3 we also see that the differences between the weight types are much larger for GSS-i. This is consistent with the fact that we observe smaller differences in AUC performances for GSS. This might be caused by the differences in dimensionality: in higher dimensions, distances become more and more similar, reducing the differences in the weights. The logistic weights are the most robust to the difference in dimensionality.

We now focus on the logistic weights, as these weights perform better on average. Examining their effect on different combinations of source and target datasets, we see that they have the most benefit when the datasets have different scan protocols. Using logistic weights when training on C1 and testing on C2 and vice versa has only small improvements, or even deteriorates the performance. This suggests that both the marginal distributions $P(\mathbf{x})$ and the labeling functions $P(y|\mathbf{x})$ and of these datasets are very similar, due to a similar distribution of subjects and the same scanning protocol.

D. Limitations

In this paper we have not considered intensity normalization as a way of dealing with differences between scanners. Normalization by trachea air intensity has been shown to reduce differences between quantitative emphysema measures from different scanners [33] and to improve correlations between emphysema and spirometry, in scans reconstructed with different kernels [34]. In a study where transfer learning is applied to segmentation in brain MR images [35], intensity normalization was performed, but transfer learning still improved the results. Therefore, we hypothesize that including a normalization step could add a further improvement to our results as well. However, as there is no widely accepted way to perform normalization in chest CT and in theory, Housefield units should be comparable between scans, we did not perform normalization.

E. Recommendations

Based on our observations, we offer some advice to researchers who might be faced with classification problems involving scans from different scanners.

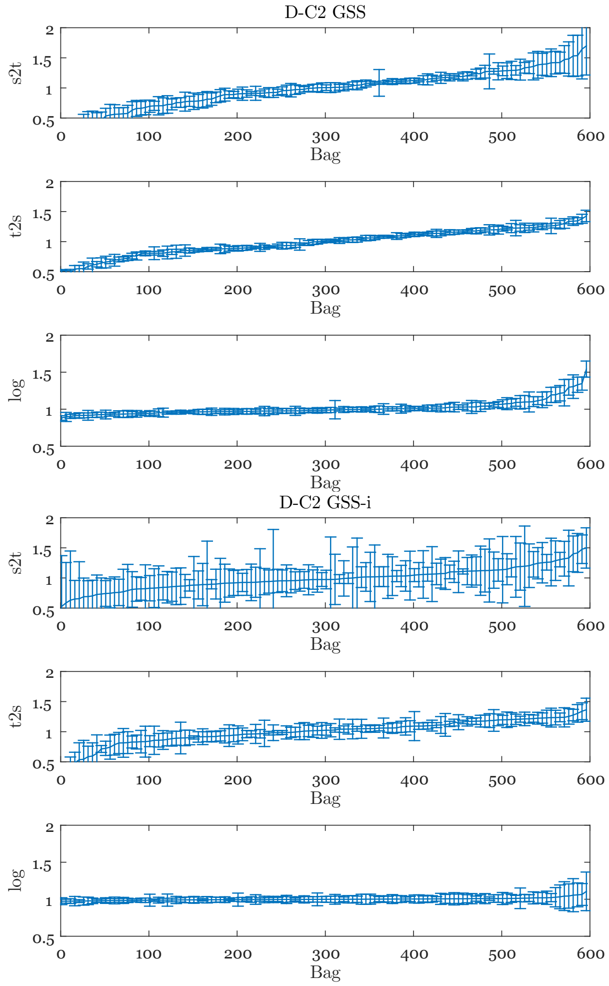

Fig. 3. Distribution of weights when training on DLCST and testing on COPDGene2, for the GSS (top three plots) and GSS-i (bottom three plots) features. The mean and standard deviation of the weight per training bag is shown, for every 5th (due to the large number of bags) training bag in DLCST. The training bags are sorted by average weight for better visualization, there is therefore no correspondence between different x-axes.

- Adaptive histograms of multi-scale Gaussian derivaties are a robust choice of features. Although originally this specific filterbank was used for classifying ROIs [3] and later classifying DLCST scans [5], we did not need any modifications to successfully apply them to independent datasets.
- If using intensity histogram features, adaptive binning is a good way to focus on the more informative intensity ranges, while keeping the dimensionality low. Reducing the dimensionality in KDE only reduces the number of bins, but does not consider their information content. As such, bins in informative intensity ranges become too wide, reducing the classification performance.
- Randomly sampled ROIs together with a SimpleMIL logistic classifier that uses the averaging rule is a good starting point for distinguishing COPD from non-COPD scans, achieving at most 79.0 (DLCST), 91.7 (COPDGene1), 95.6 (COPDGene2) and 95.3 (Frederikshavn) AUC, in %.
- Importance weighting appears not to be needed when the same cohort and only a slightly different scan protocol are used, such as with the COPDGene datasets.
- Importance weights based on a logistic classifier trained to discriminate between source data and target data, are a good starting point. These weights gave the best results overall, eliminate the scaling problem, and were much faster to compute (2 seconds per test image) than the distance-based weights (2 minutes per test image) in this study.

VI. Conclusions

We presented a method for COPD classification using a chest CT scan which generalizes well to datasets acquired at different sites and scanners. Our method is based on Gaussian scale-space features and multiple instance learning with a weighted logistic classifier. Weighting the training samples according to their similarity to the target data could further improve the performance, demonstrating the potential benefit of transfer learning techniques in this problem. Transfer learning methods beyond instance-transfer approaches could be interesting in the future. To this end, upon acceptance of the paper we will publicly release the DLCST and Frederikshavn datasets to encourage more investigation into transfer learning methods in medical imaging. We believe that developing methods that are robust across domains is an important step for adoption of automatic classification techniques in clinical studies and clinical practice.

Acknowledgements

This research was performed as part of the research project "Transfer learning in biomedical image analysis" which is financed by the Netherlands Organization for Scientific Research (NWO) grant no. 639.022.010. We thank Morten Vuust MD (Department of Diagnostic Imaging, Vendsyssel Hospital, Frederikshavn) and Ulla Møller Weinreich (Department of Pulmonology Medicine and Clinical Institute, Aalborg University Hospital) for their assistance in the acquisition of the data used in this study. We thank Wouter Kouw (Pattern Recognition Laboratory, Delft University of Technology) for discussions on logistic weights.

References

[1] R. A. Pauwels, A. S. Buist *et al.*, "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: National heart, lung, and blood institute and world health organization global initiative for chronic obstructive lung disease (gold): executive summary." *Respiratory care*, vol. 46, no. 8, p. 798, 2001.

[2] Y. S. Park, J. B. Seo, N. Kim, E. J. Chae, Y. M. Oh, S. Do 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, no. 6, pp. 395–402, 2008.

[3] L. Sørensen, 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, 2010.

[4] C. S. Mendoza, G. R. Washko, J. C. Ross, 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," in *International Symposium on Biomedical Imaging*. IEEE, 2012, pp. 474–477.

[5] L. Sørensen, 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.

[6] V. Cheplygina, L. Sørensen, D. M. J. Tax, J. H. Pedersen, M. Loog, and M. de Bruijne, “Classification of COPD with multiple instance learning,” in *ICPR*, 2014, pp. 1508–1513.

[7] V. Cheplygina, L. Sørensen, D. M. J. Tax, M. de Bruijne, and M. Loog, “Label stability in multiple instance learning,” in *Medical Image Computing and Computer-Assisted Interventions*. Springer, 2015.

[8] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[9] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[10] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, “Correcting sample selection bias by unlabeled data,” in *Advances in neural information processing systems*, 2006, pp. 601–608.

[11] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, “Covariate shift by kernel mean matching,” *Dataset shift in machine learning*, vol. 3, no. 4, p. 5, 2009.

[12] A. van Opbroek, M. W. Vernooij, M. A. Ikram, and M. de Bruijne, “Weighting training images by maximizing distribution similarity for supervised segmentation across scanners,” *Medical Image Analysis*, 2015.

[13] B. Cheng, M. Liu, H.-I. Suk, D. Shen *et al.*, “Multimodal manifold-regularized transfer learning for MCI conversion prediction,” *Brain imaging and behavior*, pp. 1–14, 2015.

[14] R. Guerrero, C. Ledig, and D. Rueckert, “Manifold alignment and transfer learning for classification of alzheimers disease,” in *Machine Learning in Medical Imaging*. Springer, 2014, pp. 77–84.

[15] A. van Opbroek, M. A. Ikram, M. W. Vernooij, and M. De Bruijne, “Transfer learning improves supervised image segmentation across imaging protocols,” *IEEE Transactions on Medical Imaging*, vol. 34, no. 5, 2014.

[16] A. van Engelen, A. C. van Dijk, M. T. Truijman, R. van’t Klooster, A. van Opbroek, A. van der Lugt, W. J. Niessen, M. E. Kooi, and M. de Bruijne, “Multi-center mri carotid plaque component segmentation using feature normalization and transfer learning,” *Medical Imaging, IEEE Transactions on*, vol. 34, no. 6, pp. 1294–1305, 2015.

[17] C. Becker, C. Christoudias, and P. Fua, “Domain adaptation for microscopy imaging,” *IEEE Transcations on Medical Imaging*, vol. 34, no. 5, pp. 1125–1139, 2014.

[18] V. H. Ablavsky, C. J. Becker, and P. Fua, “Transfer learning by sharing support vectors,” Tech. Rep., 2012.

[19] T. Schlegl, J. Ofner, and G. Langs, “Unsupervised pre-training across image domains improves lung tissue classification,” in *Medical Computer Vision: Algorithms for Big Data*. Springer, 2014, pp. 82–93.

[20] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *IEEE Transactions on Medical Imaging, in press*, 2016.

[21] P. Lo, J. Sporring, H. Ashraf, J. J. Pedersen, and M. de Bruijne, “Vessel-guided airway tree segmentation: A voxel classification approach,” *Medical image analysis*, vol. 14, no. 4, pp. 527–538, 2010.

[22] A. S. Korsager, V. Fortunati, F. van der Lijn, J. Carl, W. Niessen, L. R. Østergaard, and T. van Walsum, “The use of atlas registration and graph cuts for prostate segmentation in magnetic resonance images,” *Medical physics*, vol. 42, no. 4, pp. 1614–1624, 2015.

[23] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[24] V. Cheplygina, D. M. J. Tax, and M. Loog, “Multiple instance learning with bag dissimilarities,” *Pattern Recognition*, vol. 48, no. 1, pp. 264–275, 2015.

[25] V. Cheplygina, A. van Opbroek, M. A. Ikram, M. W. Vernooij, and M. de Bruijne, “Asymmetric similarity-weighted ensembles for image segmentation,” in *ISBI*, 2016.

[26] W. M. Kouw, L. J. P. van der Maaten, J. H. Krijthe, and M. Loog, “Feature-level domain adaptation,” *Journal of Machine Learning Research*, vol. 17, no. 171, pp. 1–32, 2016.

[27] M. Goetz, C. Weber, F. Binczyk, J. Polanska, R. Tarnawski, B. Bobek-Billewicz, U. Koethe, J. Kleesiek, B. Stieltjes, and K. H. Maier-Hein, “DALSA: domain adaptation for supervised learning from sparsely annotated MR images,” *IEEE transactions on medical imaging*, vol. 35, no. 1, pp. 184–196, 2016.

[28] J. H. Pedersen, H. Ashraf, A. Dirksen, K. Bach, H. Hansen, P. Toennesen, H. Thorsen, J. Brodersen, B. G. Skov, M. Døssing *et al.*, “The Danish randomized lung cancer CT screening trial-overall design and results of the prevalence round,” *Journal of Thoracic Oncology*, vol. 4, no. 5, pp. 608–614, 2009.

[29] 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.

[30] J. Vestbo, S. S. Hurd, A. G. Agustí, P. W. Jones, C. Vogelmeier, A. Anzueto, P. J. Barnes, L. M. Fabbri, F. J. Martinez, M. Nishimura *et al.*, “Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary,” *American journal of respiratory and critical care medicine*, vol. 187, no. 4, pp. 347–365, 2013.

[31] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach,” *Biometrics*, pp. 837–845, 1988.

[32] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[33] C. Mol, B. van Ginneken, M. de Bruijne, P. de Jong, M. Oudkerk, A. Dirksen, and P. Zanen, “Correction of Quantitative Emphysema Measures with Density Calibration Based on Measurements in the Trachea,” in *Annual Meeting of the Radiological Society of North America*, 2010.

[34] L. Gallardo-Estrella, D. A. Lynch, M. Prokop, D. Stinson, J. Zach, P. F. Judy, B. van Ginneken, and E. M. van Rikxoort, “Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification,” *European radiology*, vol. 26, no. 2, pp. 478–486, 2016.

[35] A. van Opbroek, H. C. Achterberg, and M. de Bruijne, “Feature-space transformation improves supervised segmentation across scanners,” in *Machine Learning Meets Medical Imaging*, 2015, pp. 85–93.