UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

PhD thesis

Silas Nyboe Ørting

Assessing emphysema in CT scans of the lungs

Using machine learning, crowdsourcing and visual similarity

Advisors: Marleen de Bruijne, Jens Petersen

Handed in: March 12, 2019

Contents

Abstract

Emphysema is a pathology in chronic obstructive pulmonary disease (COPD), a leading cause of death worldwide. Emphysema is characterized by destruction of lung tissue leading to reduced capacity for gas exchange in the lungs.

The extent and appearance of emphysema can be assessed in CT scans of the lungs. Current recommendations for assessing emphysema in CT scans is to use a combination of densitometry and visual assessment. Densitometry is a quantitative method that estimates the amount of lung tissue affected by emphysema, by measuring the amount of voxels in the CT scans with attenuation below a certain threshold. Visual assessment of emphysema by experts provides an assessment of emphysema extent and patterns that has been found to be useful for lung cancer risk prediction. Current densitometry methods are vulnerable to variation in scanners, scan protocols and software implementations, and cannot characterize emphysema patterns. On the other hand, visual assessment requires expert knowledge, is difficult, time-consuming and commonly only provides semi-quantitative estimates of emphysema extent.

Machine learning methods that learn from visual assessment could combine the benefits of densitometry and visual assessment, to provide fully automated assessment of emphysema extent and patterns.

One of the main issues when applying supervised machine learning in medical image analysis is obtaining labels. Not only can the labeling procedure require medical expertise and be time-consuming and costly, it can also be very difficult, even for experts, to provide accurate labels.

This thesis investigates three approaches to reducing the need for labels when training machine learning methods to assess emphysema: weakly supervised learning, crowdsourcing and learning from visual similarity.

Weakly supervised machine learning aims at learning from global labels instead of local labels, for example learning from image labels instead of pixel labels. By learning from weak labels, we can reduce the need for medical expertise, reduce the cost of labeling and improve label quality because assigning global labels is generally easier and less costly than assigning local labels. The thesis investigates emphysema detection and quantification within two weakly supervised learning settings, multiple instance learning, where labels are binary, and learning with label proportions, where labels are proportions.

Crowdsourcing aims at reducing labeling costs by replacing expert annotators with non-experts. Assessing emphysema in a volumetric CT scan is a complex task requiring expert knowledge and experience. Adapting the task to the crowd setting could enable crowdsourced labels to be used when training machine learning methods. Thereby, allowing experts to focus on interpreting and validating trained models. The thesis provides a survey of how crowdsourcing is used in medical imaging, as well as an investigation into how emphysema assessment can be framed as a task that can be solved by non-expert crowd workers.

Learning from visual similarity aims at learning representations from relative comparisons of images. One of the reasons that labels are costly and require expertise, is that labels are often obtained for a specific task and new labels need to be acquired for additional tasks. Visual similarity assessments could provide a more general characterization of visual content in images, than labels obtained for a specific task. By learning from visual similarity it is possible that more general representations can be learned. Additionally, focusing on similarity could allow non-experts to replace experts, since comparing visual similarity requires less expertise than categorizing pathology patterns. The thesis investigates how visual similarity assessments can be obtained and used for training convolutional neural networks to learn representations of chest CT scans.

Resumé

Emfysem er en patologi i kronisk obstruktiv lungesygdom (KOL), en førende dødsårsag på verdensplan. Emfysem er karakteriseret ved ødelæggelse af lungevæv der medfører reduceret kapacitet til gasudveksling i lungerne.

Omfanget og udseendet af emfysem kan bedømmes fra CT skanninger af lungerne. De nuværende anbefalinger for bedømmelse af emfysem fra CT skanninger er at benytte en kombination af densitometri og visuel bedømmelse. Densitometri er en kvantitativ metode der estimerer omfanget af emfysem, ved at målet mængden af voxels i CT skanningerne med værdier under en vis tærskel. Visuel bedømmelse af emfysem af eksperter, giver bedømmelse af emfysem omfang og mønstre, der har vist sig at være brugbart ved lunge cancer prognoser, Nuværende densitometri metoder er sårbare overfor variation i skannere, skan protokoller og software implementeringer, og kan ikke karakterisere emfysem mønstre. På den anden side kræver visuel bedømmelse tid og ekspertviden, og giver oftest kun semi-kvantitative estimater af emfysem omfang.

Maskinlæringsmetoder der lærer fra visuel bedømmelse, kan kombinere fordelene ved densitometri og visuel bedømmelse til at give en automatisk bedømmelse af emfysem omfang og mønstre. En af hoved udfordringerne ved at anvende maskinlæring til medicinsk billedanalyse er at få annoteringer. Annoterings arbejdet kræver ikke kun ekspertviden, tid og penge. Det kan også være meget vanskeligt, selv for eksperter, at lave nøjagtige annoteringer. Denne afhandling undersøger tre tilgange til at mindske behovet for annoteringer når maskinlæringsmetoder trænes til at bedømme emfysem: læring fra svage annoteringer, crowdsourcing og læring fra visuel lighed.

Læring fra svage annoteringer sigter mod at lære fra globale annoteringer istedet for fra lokale annoteringer. F. eks fra billedeannoteringer istedet for fra pixel annoteringer. Ved at lære fra svage annoteringer mindskes både pris og behovet for medicinsk ekspertise, samtidigt med at kvaliteten af annoteringerne øges fordi det generelt er nemmere at lave globale annoteringer end lokale annoteringer. Denne afhandling undersøger detektering og kvantificering af emfysem indenfor to varianter af læring fra svage annoteringer, multiple instance learning, hvor annoteringer er binære, og learning with label proportions, hvor annoteringer er proportioner.

Crowdsourcing sigter mod at reducere annoterings omkostninger ved at erstatte eksperter med uerfarne bedømmere. Det er en kompleks opgave, der kræver ekspert viden og erfaring, at bedømme emfysem i en volumetrisk CT skanninger. Ved at tilpasse opgaven er det muligt at bruge crowdsourcing til at få annoteringer der kan bruges til at træne maskinlæringsmetoder. Derved kan eksperterne fokuserer på at fortolke og validere de trænede modeler. Denne afhandling undersøger hvordan crowdsourcing bliver brugt til medicinsk billedanalyse, samt hvordan bedømmelse af emfysem kan tilpasses så crowdsourcing kan benyttes til at få annoteringer.

Læring fra visuel lighed sigter mod at lære repræsentationer fra relative sammenligninger af billeder. En af grundene til at annoteringer er dyre og kræver ekspert viden, er at annoteringer ofte fokusere på et bestemt problem og nye annoteringer er nødvendige for hvert nyt problem. Bedømmelse af visuel lighed kan give en mere generel karakteristik af det visuelle indhold i billeder, end annoteringer fokuseret på et bestemt problem. Ved at lære fra visuel lighed er det muligt at mere generelle repræsentationer kan læres. Derudover, kan det gøre det muligt at erstatte eksperter med uerfarne bedømmere, eftersom det kræver mindre ekspertise at bedømme visuel lighed end at kategorisere patologiske mønstre. Denne afhandling undersøger hvordan visuel lighed kan bedømmes og bruges til at træne convolutional neural networks til at lære repræsentationer af lunge CT skanninger.

1 Introduction

> Chronic Obstructive Pulmonary Disease (COPD) is a common, preventable and treatable disease that is characterized by persistent respiratory symptoms and airflow limitation that is due to airway and/or alveolar abnormalities usually caused by significant exposure to noxious particles or gases. The chronic airflow limitation that is characteristic of COPD is caused by a mixture of small airways disease (e.g., obstructive bronchiolitis) and parenchymal destruction (emphysema), the relative contributions of which vary from person to person.
>
> *Global Initiative for Chronic Obstructive Lung Disease, 2019 report [14]*

The interplay between small airways disease, emphysema and factors such as genetics and blood biomarkers have been the focus of several large investigations [39, 8], with the aim of understanding how COPD develops in and impacts individuals. Several large lung cancer screening trials [44, 38, 25] have investigated CT-based lung cancer screening and found emphysema to be predictive of lung cancer risk. This increased focus on CT based analysis of emphysema has likely been a main contributing factor in recent years' increased interest in automating assessment of emphysema, primarily through the use of machine learning (ML) based methods. This thesis contributes to this line of research by investigating how machine learning methods can learn to assess emphysema.

Section 1.1 provides a brief overview of lung anatomy, emphysema and CT imaging of the lungs. Section 1.2 provides an overview of current practice for assessing emphysema, as well as outlining recent advances within machine learning based assessment. Section 2 provides an outline of the thesis and highlights the contributions. Sections 3,4 and 5 contain manuscripts included in the thesis. Section 6 discusses the contributions as well as perspectives for the future.

1.1 Lungs, emphysema and X-ray computed tomography

The following three paragraphs are largely based on Tortora and Derrickson [46], Flower [10] and Lynch et al. [22].

The human respiratory system is responsible for exchanging oxygen in the air with carbon-dioxide in the blood. Inhaled oxygen-rich air is spread throughout the lungs through the airway tree. Starting at the trachea, the airway branches multiple times, eventually culminating at the alveoli where the gas exchange occurs. Finally, the oxygen-poor air is exhaled. Emphysema is a pathology characterized by the destruction of alveoli. When air is inhaled it will also spread to areas with emphysema, however little or no gas exchange occurs in the affected regions.

X-ray computed tomography (CT) can be used for assessing the appearance, severity and distribution of emphysema in the lungs. When an X-ray passes through an object some of the energy of the ray is attenuated (absorbed and reflected) by the object. By measuring the amount of energy that is not attenuated, it is possible to measure the density of the object. Imagine a setup with an X-ray source, an object of interest and a detector on a line. Rotating the source and detector around the object forms a plane. From attenuation measurements 180 degrees around the object it is possible to compute a reconstruction of the part of the object intersected by the plane. By computing the reconstructing in small increments along the axis perpendicular to the plane, we obtain a tomographic reconstruction showing the 3D distribution of densities in the object. Such a reconstruction is referred to as a CT scan. Slices from a chest CT scan are shown in Figure 1. Medical CT scans are normalized with the Hounsfield unit (HU) scale such that the value of water is 0HU and the value of air is -1000HU. This normalization allows comparisons across subjects, scanners, and so forth.

Three main sub-types of emphysema have been defined based on the appearance and distribution of emphysema in CT. Centrilobular emphysema, characterized by areas of low attenuation surrounded by normal lung tissue; paraseptal emphysema, characterized by clearly defined areas of low attenuation at the boundary of the lungs and lobes; and panlobular emphysema, characterized by a more uniform decrease in attenuation across the affected area. In severe cases, the sub-types have similar appearance with large areas of the lung completely destroyed. Examples of centrilobular and paraseptal emphysema are shown in Figure 2. The data used in this thesis

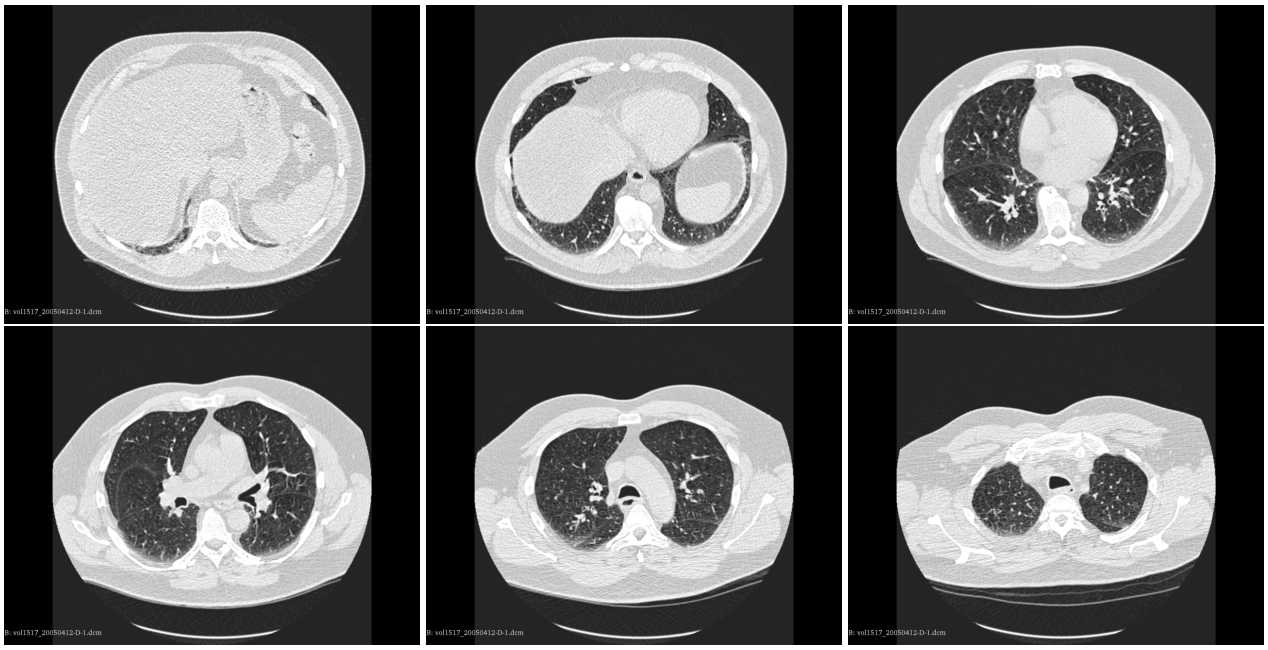

Figure 1: Examples of axial slices from a chest CT scans with a resolution of 0.78mm $\times$ 0.78mm. Moving from the stomach towards the head from top left to bottom right. The lung to the left in the images is the right lung. A typical chest CT scan with 1mm slice spacing contains around 300 slices.

is from the Danish Lung Cancer Screening Trial (DLCST) [38], where the prevalence of panlobular emphysema is very low. The panlobular emphysema pattern is thus not of particular interest in this thesis.

1.2 Assessment of emphysema

The current recommendation for assessing emphysema in CT scans of the lung is to use a combination of quantitative measures of lung density and visual assessment [22]. Recently, several machine learning methods have been proposed for quantifying and characterizing emphysema and COPD based on CT scans. Despite promising results, these methods have yet to impact clinical practice.

1.2.1 Density based assessment

Quantitative measures of lung density are based on the observation that air-filled areas in the lung will appear as areas of low attenuation in CT images. Early work by [26] showed that measuring the percentage of lung area with density below a fixed threshold (-910HU) correlated well with pathological findings in resected lung tissue. Later work by [13], however, found that only a threshold of -950HU resulted in measurements not significantly different from findings in resected lung tissue. Deciding which fixed threshold to use is challenging due to differences in acquisition and populations between studies.

An alternative to fixed thresholding is to consider the histogram of attenuation values in each scan. PD15, where the value at the 15'th percentile is used, was found to be the most sensitive when measuring progression in subjects with alpha 1-antitrypsin deficiency [37].

Regardless of method, when using density-based measures of emphysema it is necessary to consider the impact of variation in CT scans. There are many sources of variation in CT scans. Differences in scanners and protocols, inspiration level, study populations [16] and software implementations [49] all contribute to variation in CT scans and derived quantitative measurements. Some of the variation can be reduced by careful normalization and standardization [40, 21, 11]. However, even when variation is accounted for, density based measurements cannot currently characterize the appearance of emphysema patterns. The difference in appearance of centrilobular and paraseptal emphysema can be striking, and there is evidence that paraseptal emphysema is

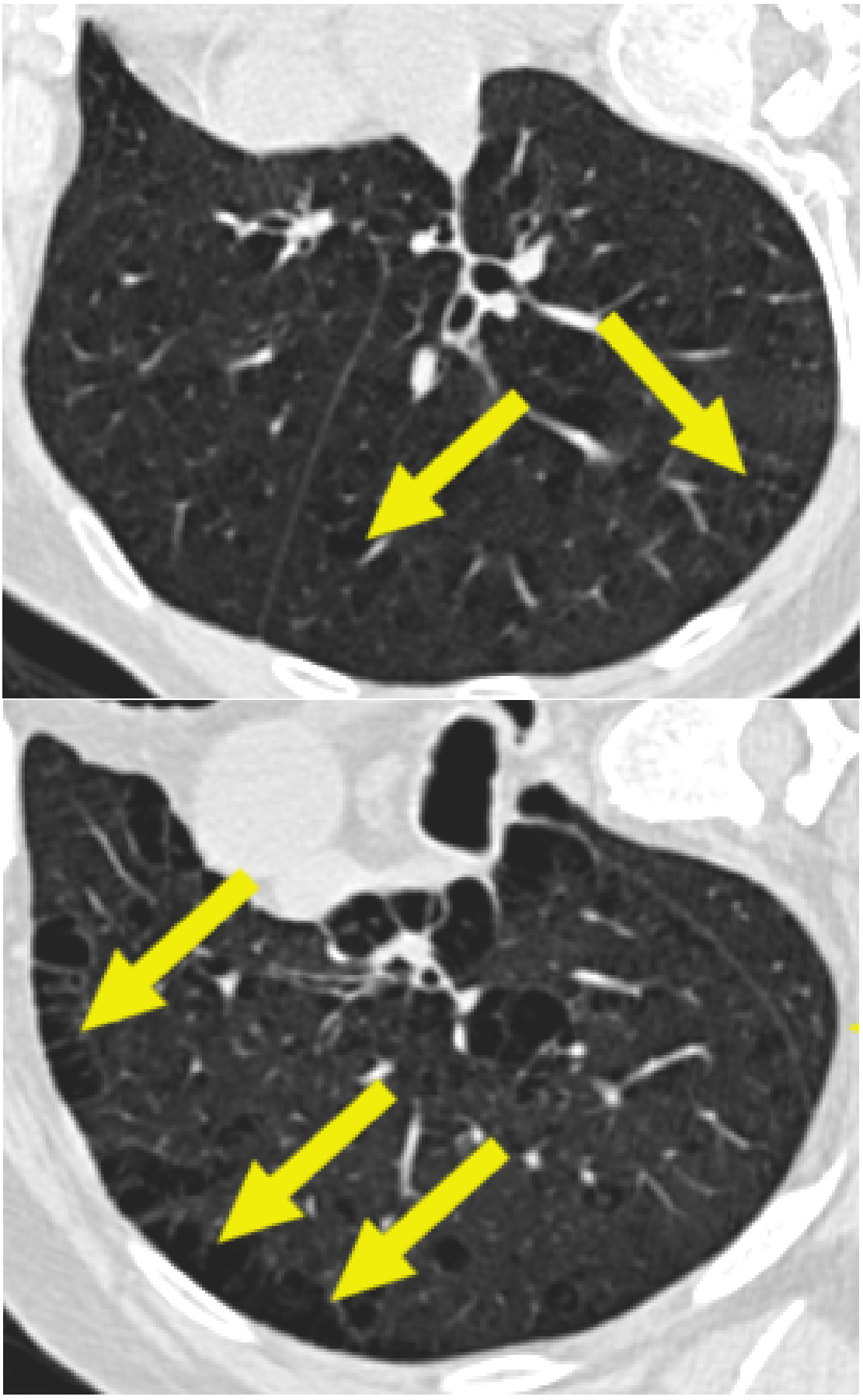

Figure 2: Appearance of emphysema in CT lung scans. Dark regions have low attenuation (black is -1000HU) and bright region have high attenuation. Arrows indicate some areas with emphysema. Top: Centrilobular emphysema. Bottom: Paraseptal emphysema.

less associated with symptoms than centrilobular emphysema [22]. Further, panlobular emphysema primarily occurs in subjects with a genetic disorder causing alpha 1-antitrypsin deficiency, indicating that the appearance of emphysema patterns is important for patient-specific prognosis and treatment.

1.2.2 Visual assessment

Visual assessment of emphysema dates back to the early studies of CT lung imaging. As technology has changed so have protocols for visual assessment, from grid-based assessment of hard-copy printouts in the eighties [26] to region based assessment using interactive views today [7]. The main benefit of visual assessment is a combined assessment of appearance, distribution and severity of emphysema that is lacking from density-based methods. Additionally, visual similarity is less affected by differences in scanners and inspiration level. There is indication that visual assessment of emphysema can improve lung cancer risk prediction [51] and overall mortality risk prediction [23] over density based assessment. The main limitations of visual assessment are cost and substantial inter-rater variation [7]. Additionally, the common practice of using a six-point scale for scoring emphysema extent is relatively coarse compared to density based assessment.

1.2.3 Machine learning based assessment

By learning from visually assessed emphysema, machine learning could overcome the limitations of visual assessment and provide a characterization of emphysema patterns lacking from density based methods. Machine learning in medical imaging is rapidly evolving. New methods and variations are constantly proposed and applied to a variety of tasks, including COPD and emphysema assessment. This section is intended as an overview of trends in the context of COPD and emphysema, not as a comprehensive review of machine learning based assessment of COPD and emphysema.

The majority of machine learning methods for assessing emphysema are based on supervised learning using visually assessed emphysema presence, extent or patterns as labels. The supervised machine learning methods can be grouped into weak supervision and strong supervision. There are degrees of weak and strong supervision. In this context we consider strong supervision to be supervision at pixel or patch level. Pixel level supervision refers to cases where areas containing a single pattern are contoured. We consider weak supervision to be supervision at image or region level. Region level supervision refers to cases where anatomically defined lung regions are assigned a single label, as is common in visual assessment of emphysema.

Several strongly supervised methods have been proposed. Sørensen et al. extracted 2D patches from the lungs and assigned each patch one of four emphysema labels. A k-nearest-neighbor classifier was then used to predict emphysema categories for new patches using a histogram representation of patches. A similar approach was used in [4] with six emphysema categories and a large set of images. More recently, CNN based methods have been trained to detect emphysema as one of several interstitial lung disease patterns [1, 12, 48]. These three studies used data from [9] where experts have annotated lung tissue patterns by tracing the contours of affected regions.

A common approach to weakly supervised learning is multiple instance learning (MIL). In MIL we have unlabeled samples grouped into labeled bags. When working with images, the common approach is to sample patches from images and view patches from the same image as a bag of samples. In the standard MIL setting, bag labels are binary and indicate that at least one sample satisfy the bag label. However, it is increasingly common to consider other relations between instances and bag labels, for example that a bag label is the majority of instance labels. In the context of COPD, bag labels could be COPD stage.

MIL approaches have been used for texture-based COPD prediction in [42, 5], where patches are represented by scale-space texture descriptors and MIL models trained to predict COPD diagnosis. More recently, González et al. proposed a CNN model for predicting COPD diagnosis as well as acute respiratory disease events. Four slices were extracted at predefined anatomical landmarks to reduce the dimensionality of the CT scan and allow the network to focus on the most important parts of the images. A variation on this approach was proposed in [17] where the resulting COPD predictions were shown to be useful for lung cancer risk prediction.

The traditional MIL setting can be extended in various ways, for example learning with label proportions (LLP) where bag labels are proportions indicating how many samples satisfy the label. When learning from emphysema extent scores, the LLP setting matches the learning task better than MIL methods and can potentially provide better results by improved utilization of label information.

MIL and LLP methods have also been used for emphysema classification and quantification. Hofmanninger and Langs proposed to learn so called "semantic profiles" linking super-voxel texture descriptors to image level labels indicating presence of five interstitial lung disease abnormalities, including emphysema. The results showed good correspondence between voxel labels predicted from semantic profiles and expert defined voxel classifications. More recently, several CNN methods have been proposed to predict visually assessed emphysema presence [17] and extent [3, 19]. Bortsova et al. showed that learning from emphysema extent scores (LLP) improved quantification and localization over learning from emphysema presence (MIL) alone.

A few approaches use unsupervised learning to search for patterns that can characterize emphysema. Häme et al. propose to use a two stage algorithm. Firstly, texture prototypes are generated by locating key-points using a standard blob detector (difference of Gaussians) and clustering texture descriptors extracted at the key points. Secondly, texture prototypes are grouped based on spatial proximity in lungs to obtain lung texture patterns. This idea is extended in [52], where spatial features are integrated to account for variation in emphysema across lung regions. Binder et al. propose to jointly model patient clusters and disease subtypes using a generative model. Explicitly modeling patient clusters allows the model to learn distinct patterns that capture a large part of the variation in emphysema subtypes.

2 Aim, outline and contributions

2.1 Aim

The aim of this thesis has been to investigate methods for assessing emphysema in chest CT scans, with emphasis on learning from visual assessment of texture patterns.

2.2 Outline

There are three tracks in the thesis, weakly supervised learning, crowdsourcing and learning from visual similarity.

2.2.1 Weakly supervised learning

One of the main issues when applying supervised machine learning in medical image analysis is obtaining labels. Not only can the labeling procedure require medical expertise and be time-consuming and costly, it can also be very difficult, even for experts, to provide accurate labels. By learning from weak labels, we can reduce the need for medical expertise, reduce the cost of labeling and improve label quality because assigning global labels is generally easier and less time-consuming than assigning local labels. Section 3 investigates emphysema detection and quantification using multiple instance learning (MIL) and learning with label proportions (LLP) methods that learn from region or scan level labels.

2.2.2 Crowdsourcing

A complementary approach to reduce labeling cost is to use crowdsourcing. Although there is a risk that label quality worsens when replacing experts with crowd workers, the increased quantity can make up for it. By adapting the labeling task so it is easier for crowd workers to do and combining labels from multiple workers, it is possible to obtain expert level quality [24]. Combining crowdsourcing and weakly supervised learning could reduce the need for experts when developing methods, and allow them to focus on interpreting and validating trained models. Section 4 provides a survey of how crowdsourcing is used in medical imaging, as well as an investigation into how emphysema assessment can be framed as a task that can be solved by non-expert crowd

workers. Instead of asking crowd workers to assess emphysema extent and sub-type, we ask them to decide which images are visually most similar.

2.2.3 Learning from visual similarity annotations

One of the reasons obtaining labels is costly and require expertise, is that labels are often obtained for a specific task, e.g. tumor segmentation or emphysema quantification, and new labels need to be acquired for additional tasks. By focusing on a more general characterization of the visual content in images it is possible that more general representations can be learned. Assessing visual similarity is an approach for characterizing image content based on the similarity between images. For three images a, b, c we could ask if image a is more similar to image b than to image c. The answer to this question yields a similarity triplet that can be used as a constraint for learning a representation of the images. A potential benefit of assessing visual similarity is that being able to recognize and distinguish pathologies is not necessary as long as the rater can decide if patterns are similar or not, thus reducing the reliance on experts for annotation tasks. Section 5 investigates how convolutional neural networks (CNNs) can learn representations from similarity triplets.

2.3 Contributions

The contributions in this thesis falls into three tracks, weakly supervised learning, crowdsourcing and learning from visual similarity.

2.3.1 Weakly supervised learning

A study of MIL and LLP methods for detecting and quantifying emphysema. The study is presented in the three papers [31, 34, 35] included in Section 3.

The paper [31] investigates how the LLP method Cluster Model Selection (CMS) [43] can be used for emphysema quantification. CMS is based on clustering patches into a small set of clusters and subsequently labeling the clusters to obtain patch labels. A weighting of features is learned using an evolutionary strategy in order to adapt the feature space to the task at hand. [31] proposes a new loss function to avoid an issue with singularities in the objective. as well as replacing the simple evolutionary strategy used in [43] with Covariance Matrix Adaptation - Evolutionary Strategy (CMA-ES), state-of-the-art in evolutionary optimization. [31] extends work produced as part of my Masters thesis [30]. The method achieved intra-class correlation (ICC) of 0.73 with a gold standard, compared to inter-rater ICC of 0.83. While MIL had previously been used for predicting COPD and content-based retrieval of scans with emphysema, this work was the first to investigate the use of proportion labels in order to learn from visual assessment of regional emphysema extent.

The paper [34] investigates how a "Simple MIL" method can be used for regional emphysema detection. This work presented the first comparison of learning from scan-level labels versus learning from region-level labels in the context of emphysema assessment. In the Simple MIL approach, samples inherit bag labels and the problem is treated as a standard supervised learning problem. In [34] a logistic regression model was trained with image level labels and compared to the the same model trained with region-level labels. The models achieved similar performance, with ROC AUC scores of .80 for scan level detection, and 0.76–0.89 for regional detection.

Finally, [35] combines and extends [31] and [34] to investigate if MIL methods that learn from binary labels can predict emphysema extent as well as LLP methods that learn from proportion labels.[35] compares nine different MIL and LLP methods and finds that the best MIL and the best LLP methods have similar performance, suggesting that emphysema presence labels, which are less costly to obtain, could be enough to learn to quantify emphysema. The comparison includes previously published methods as well as new variations of these methods.

A limitation of the work in these three papers, is that all methods use a predefined set of scale space features instead of learning the features. Jointly learning features and classifier using CNNs was explored in collaborations [3, 45] and are not included in this thesis.

2.3.2 Crowdsourcing

A study of crowdsourcing as an approach to obtaining labels for medical image analysis tasks, in particular for emphysema assessment. The study is presented in the papers [32, 29] included in Section 4.

The paper [32] investigates how emphysema assessment can be posed as a task that can be solved by crowd workers. This was the first work to investigate crowdsourced assessment of emphysema[1] . Instead of assessing presence or severity of emphysema, crowd workers are asked to assess how similar images are. Each task shows coronal slices from the upper right lung region of three subjects, and asks crowd workers to assess which slices are visually most similar. This results in a set of similarity triplets of the form "image a is more similar to image b than to image c". These similarity triplets are then used to find an embedding of the images that satisfy the triplets. Untrained crowd workers were recruited from the Amazon Mechanical Turk platform[2] and the collected similarity triplets used to find an embedding using t-distributed stochastic triplet embedding (tSTE) [47]. Although the quality of the crowdsourced annotations varied, we found that crowd workers' visual assessment was informative of emphysema sub-type patterns.

The paper [29] is a survey of crowdsourcing in medical imaging. It is a collaboration arising from the Lorentz workshop "Crowdsourcing in Medical Imaging" [3]. The survey summarize 55 papers on crowdsourcing in medical imaging covering a large range of modalities and tasks. The two most important conclusions of the survey are (1) crowdsourcing is a viable approach to label medical images and (2) reporting of experiments is lacking and must be improved in order to realize the full potential of crowdsourcing in medical image analysis.

2.3.3 Visual similarity

A study of how CNNs can learn expressive representations using visual similarity assessments of chest CT slices. The study is presented in [33, 36] included in Section 5.

This work builds on [32] where tSTE was used to embed images. tSTE finds an embedding that satisfies similarity triplets, but does not provide a method for mapping unseen images into the embedding. By using a triplet CNN it is possible to learn a mapping from image space to embedding space that satisfies the triplets, and can map unseen images into the embedding. Thus allowing unseen images to be represented in the learned space, and enabling classifiers built on the representation to classify unseen images.

The paper [33] investigates how CNNs can learn a representation of emphysema in chest CT scans using similarity measurements derived from scan labels. We define a "similarity oracle" using visually assessed emphysema extent scores and show that the oracle can be used to train a triplet CNN to learn an emphysema sensitive representation of CT slices. The CNN is trained by picking three images from a set of images and using the oracle to decide which are most similar. Since the emphysema extent scores are on a six point scale, there is a high likelihood of picking three images with the same extent score. In this case the oracle cannot provide any information. We show that although using only informative triplets (at least two images have different labels) is optimal, having a substantial amount of uninformative triplets only leads to minor decrease in performance. This indicates that the triplet CNN is robust to noisy labels.

The paper [36] investigates how CNNs can learn a representation of emphysema in chest CT scans using visual similarity measurements from a non-expert rater. We demonstrate how a large set of visual similarity triplets, 180,000 triplets for 300 images, can be obtained and find that the approach yields reasonable levels of inter- and intra-rater agreement. We use the similarity triplets to train a CNN model to learn an eight dimensional representation of the images, and show that this representation is useful for detecting both emphysema and interstitial lung disease. When assessing visual similarity for a large set of images, it is likely that the measure of similarity will not be the same for all images. We propose a method to incorporate multiple notions of similarity in the learning process. Although the approach does not improve performance, it pushes the features of the representation to be less correlated, which could be useful for interpreting the learned representation.

[1] Crowdsourcing annotations of interstitial lung disease patterns was investigated in [28] presented at the same workshop as [32]

[2] https://www.mturk.com

[3] https://www.lorentzcenter.nl/lc/web/2018/967/info.php3?wsid=967&venue=Snellius

3 Emphysema quantification using Multiple Instance Learning and Learning with Label Proportions

This section is based on three manuscripts. The first two investigate emphysema quantification using a learning with label proportions (LLP) method [31] and a multiple instance learning (MIL) method [34]. The last manuscript [35] investigates what labels are needed for learning to quantify emphysema extent by comparing nine different MIL and LLP methods.

[31] Silas Nyboe Ørting, Jens Petersen, Mathilde M W Wille, Laura H Thomsen, and Marleen de Bruijne. *Quantifying emphysema extent from weakly labeled CT scans of the lungs using label proportions learning*, The Sixth International Workshop on Pulmonary Image Analysis. 2016.

[34] Silas Nyboe Ørting, Jens Petersen, Laura H Thomsen, Mathilde M W Wille, and Marleen de Bruijne. *Detecting emphysema using multiple instance learning*. In 2018 IEEE 15th International Symposium on Biomedical Imaging. 2018.

[35] Silas Nyboe Ørting, Jens Petersen, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. *Learning to quantify emphysema extent: What labels do we need?* arXiv preprint arXiv:1810.07433. 2018.

Quantifying Emphysema Extent from Weakly Labeled CT Scans of the Lungs using Label Proportions Learning

Silas Nyboe Ørting[1], Jens Petersen[1], Mathilde M W Wille[2], Laura H. Thomsen[3], and Marleen de Bruijne[1,4]

[1] Department of Computer Science, University of Copenhagen, Denmark, silas@di.ku.dk
[2] Department of Diagnostic Imaging, Section of Radiology, Nordsjællands Hospital, Denmark
[3] Department of Respiratory Medicine, Gentofte Hospital, Denmark
[4] Department of Radiology and Medical Informatics, Erasmus MC Rotterdam, The Netherlands

Abstract. Quantification of emphysema extent is important in diagnosing and monitoring patients with chronic obstructive pulmonary disease (COPD). Several studies have shown that emphysema quantification by supervised texture classification is more robust and accurate than traditional densitometry. Current techniques require highly time consuming manual annotations of patches or use only weak labels indicating overall disease status (e.g, COPD or healthy). We show how visual scoring of regional emphysema extent can be exploited in a learning with label proportions (LLP) framework to both predict presence of emphysema in smaller patches and estimate regional extent. We evaluate performance on 195 visually scored CT scans and achieve an intraclass correlation of 0.72 (0.65–0.78) between predicted region extent and expert raters. To our knowledge this is the first time that LLP methods have been applied to medical imaging data.

1 Introduction

Emphysema is a central structural abnormality in patients suffering from chronic obstructive pulmonary disease (COPD), a leading cause of death worldwide. Emphysema is characterized by destruction of lung tissue and entrapment of air in affected regions. Quantifying emphysema extent is useful for monitoring progression [11] and in the search for genetic associations with COPD [1].

Emphysema is visible in chest CT scans and standard methods for CT-based assessment of emphysema are densitometry and visual scoring by experts. Densitometry provides an objective measure of emphysema, but is vulnerable to noise and cannot be used to distinguish emphysema sub-types. Visual scoring provide information about emphysema sub-type along with estimates of emphysema extent, but suffers from inter-observer variability and is time consuming. A recent machine learning approach used expert annotations of CT patches for

predicting emphysema sub-type and severity [2]. Region based visual scoring is less time-consuming than annotating patches and more clinically relevant [11], making it more realistic to obtain large data sets.

In this work we classify patches of CT scans by learning emphysema patterns from visual scoring of regional emphysema extent. In this type of visual scoring, the lungs are divided into six regions, the upper, middle and lower regions of the right and left lungs, and each region is assigned a percentage interval indicating the extent of emphysema in the region.

We view this learning problem as an instance of learning with label proportions (LLP). LLP is a relatively new learning setting first introduced by [6] as an extension of multiple instance learning (MIL) to proportion labels. In both MIL and LLP we are concerned with bags of instances, e.g. a collection of patches from a CT scan, and we wish to predict unknown instance labels from known bag labels. The difference between MIL and LLP is that MIL learns from binary bag labels, e.g. COPD versus no-COPD as in [3, 10] and LLP learns from proportion labels that indicate the proportion of instances in a bag with a certain label. Bag proportion labels provide more information about instance labels than binary bag labels and LLP methods attempt to use the extra information to improve performance.

Several LLP methods have been proposed, Kück and de Freitas [6] develop a graphical model where both instance labels and true bag proportions are treated as unknowns; Yu et al. [12] adapt support vector machines to LLP, and present a method for iteratively optimizing instance and bag loss; Patrini et al. [7] present Laplacian Mean Map and show that aggregate statistics can be sufficient for optimizing a large class of loss functions.

In this work we adapt cluster model selection (CMS) [9] to the problem of learning from visual scoring of emphysema. CMS searches for a clustering of patches that match known region labels. A part of the search is reshaping the feature space to improve clustering, and this feature space optimization together with the fact that no assumptions are made for the bag loss makes CMS attractive. We reformulate the CMS problem so it is straightforward to use a non-standard bag loss and contribute an interval bag loss for visually scored intervals of emphysema extent. We replace the feature weight optimization method with CMA-ES, a state-of-the-art method for black-box optimization and evaluate the method on visually scored CT scans. To our knowledge this is the first time that regional visual scoring of emphysema has been used to train a classifier, and the first time that LLP has been applied to medical image data.

2 Methods

Based on previous work by [10] and [3] for predicting COPD from CT scans, we take a texture-analysis approach to characterizing emphysema patterns. Each patch is represented by a collection of histograms of filter responses. The filters are multi-scale Gaussians and combinations of derivatives of Gaussians. A summary of the used filters is given in Table 1 and a thorough description of the

filters can be found in [10]. The filters are applied at scales $\sigma \in \{1.2, 2.4, 4.8\}$ mm, a subset of those used in [10] chosen as a compromise between feature space dimension and expressiveness.

Table 1. Multi-scale filters for analyzing lung texture. I is an image and G_σ is a Gaussian with scale σ. The asterisk $*$ indicates convolution. The Hessian is the matrix of second order partial derivatives of I, where the partial derivatives are computed by convolution with a corresponding partial derivative of a Gaussian

Feature name	Definition	Feature name	Definition
Gaussian blur	$G_\sigma * I$	Laplacian of Gaussian	$\sum_{i=1}^{3} \lambda_i$
Gradient magnitude	$\|\|\nabla G_\sigma * I\|\|$	Gaussian curvature	$\prod_{i=1}^{3} \lambda_i$
Eigenvalues of the Hessian	$\|\lambda_1\| \geq \|\lambda_2\| \geq \|\lambda_3\|$	Frobenius norm	$\sqrt{\sum_{i=1}^{3} \lambda_i^2}$

2.1 Cluster Model Selection

Cluster model selection (CMS) introduced by [9] is a machine learning method for learning from label proportions (LLP). Let $\mathcal{X}^d$ be a d-dimensional feature space, in our case it is the d filter responses, and $x \in \mathcal{X}^d$ an instance or patch. A bag $G_i \in \mathcal{X}^{l \times d}$ is a set of l patches from a lung region and $Y_G^i \in \mathcal{Y}$ is a bag label indicating the extent of emphysema in the region. Here $\mathcal{Y} = \{[I_{low}, I_{high}] | I_{low} < I_{high},\ I_{low}, I_{high} \in [0,1]\}$ is the set of closed intervals on the closed unit interval $[0,1]$. In LLP we have a set of m bags $G = \{G_1, G_2, \dots G_m\}$ with associated bag labels $Y_G = \{Y_G^1, Y_G^2, \dots Y_G^m\}$ and we want to predict a binary label for each patch indicating if emphysema is present.

Cluster model selection is a data-driven approach based on clustering. A cluster model in this context is a partitioning of $X = \{G_1 \cup G_2 \cup \dots G_m\}$ into k clusters $S = \{S_1, S_2, \dots S_k\}$ with a cluster labeling $Y_S \in \{0,1\}^k$ indicating if a cluster is an emphysema cluster. An instance $x \in S_i$ inherits the label of S_i and a bag label can be estimated as the mean instance label over all instances in the bag. The cluster model problem can be defined as

$$\arg\min_{w, \tilde{Y}_S} \frac{1}{m} \sum_{i=1}^{m} L(Y_G^i, \tilde{Y}_G^i) \ , \tag{1}$$

where $w \in [0,1]^d$ is a weighting of features and $\tilde{Y}_G$ the estimated bag labels derived from the cluster labeling $\tilde{Y}_S$. L is a bag loss function that measures the loss incurred by predicting $\tilde{Y}_G^i$ when the real bag label is Y_G^i.

Optimizing (1) is done by splitting it in smaller steps. For a given feature weight vector w we find a clustering S^w by minimizing the within-cluster distance

to the cluster center

$$S^w = \arg\min_S \sum_i^k \sum_{x \in S_i} d_P(x, \mu_i | w) \ , \tag{2}$$

where μ_i is the mean of instances in cluster S_i and d_P is a weighted patch distance defined by

$$d_P(x, y|w) = \sum_{i=1}^{d} w_i d_H(x_i, y_i) \ , \tag{3}$$

where d_H is a histogram distance function. Following [10] we use the earth movers' distance to measure histogram distance. Minimizing (2) is NP-hard and we use the k-means algorithm to find an approximate solution.

Cluster Labeling. The original CMS formulation considers real valued label proportions and uses a loss function with potentially[5] multiple "sub-optimal" global minima. The problem is that several terms are combined as a product, so if any term is zero the other terms can be arbitrarily large. While the loss function cannot distinguish between the cases where all terms are zero and one term is zero, it is unreasonable to consider the two cases equally good solutions. Here we contribute an interval bag loss more suitable for our purpose, and while it also has potential for multiple global minima, due to the interval bag labels, but all the global minima are "equally optimal" from the definition of the loss function.

For a clustering S we search for the cluster labeling that minimizes the bag loss L. Let $I = [I_{low}, I_{high}]$ be the known interval label and $p \in [0, 1]$ the predicted label. We define the bag loss

$$L(I^i, p_i) = \begin{cases} I^i_{low} - p_i & \text{if } p_i < I^i_{low} \\ p_i - I^i_{high} & \text{if } p_i > I^i_{high} \\ 0 & \text{otherwise} \end{cases} \ . \tag{4}$$

$L(I^i, p_i)$ is zero when p_i is inside the interval and equal to the shortest absolute distance from the interval otherwise.

The instances from each bag G_i are distributed over the clustering S and we define a matrix M that maps cluster labels to bag labels, such that M_{ij} is the proportion of instances from G_i that belongs to cluster S_j. This allows us to formulate the labeling problem as

$$\arg\min_{Y_S} \sum_i^m L(I^i, (MY_S)_i), \ s.t. \forall j \in [1:k]. \ 0 \leq Y^j_S \leq 1 \ . \tag{5}$$

Solving (5) is NP-hard for binary cluster labels, so we use a greedy heuristic. We start by assigning all clusters label zero, then we search for the best labeling

[5] It is potentially, because it depends on the clustering - some clusterings have a unique global minima

when only one cluster is labeled one. From a cluster labeling with i clusters labeled one, we search for the best labeling with $i+1$ clusters labeled one. The labeling is stopped when there is no longer an improvement in (5).

Feature Weight Optimization Clustering and cluster labeling is wrapped in a black-box optimization over w. The original formulation of CMS uses a simple genetic algorithm which we have replaced with state-of-the in black-box optimization, CMA-ES. Originally proposed in [5], CMA-ES is a genetic algorithm that works by generating a set of candidate weight vectors W from a multivariate Gaussian distribution with mean m and co-variance C. For each $w' \in W$ we evaluate the fitness of w' by optimizing (1) with $w = w'$. The candidate weights are then ranked and used to update m and C before a new set of candidates are generated. The process is iterated until convergence or a maximum number of iterations is reached.

3 Experiments and Results

The method is evaluated on low-dose CT scans from the Danish Lung Cancer Screening Trial [8]. Visual scoring of emphysema is performed by two raters using the method described in [11]. Each rater assigns one of seven labels to the upper, middle and lower regions of each lung. The labels {0%, 1–5%, 6–25%, 26–50%, 51–75%, 76–100%} indicate the percentage of the region affected by emphysema. Three data sets have been defined A_{train}, A_{validate}, A_{test} with respective sizes of 193, 195, 195 scans. Each data set was initially 200 scans, matching the data sets defined in [10], but some scans were excluded because they were not visually scored. A set of 50 patches with a size of approximately $21 \times 21 \times 21\text{mm}^3$ were sampled from each region of the lungs and aggregated into bags. Emphysema is commonly characterized by the appearance of tissue destruction in lobules, which are about 10–25mm in diameter [4], and the patch size has been chosen to approximately match the size of lobules. Each bag was labeled by combining the extent of both raters, such that the combined interval is the smallest interval containing the interval of both raters. We assume that the extent labels can be interpreted as the proportion of patches containing emphysema.

Model training is a two-step procedure, in the first step we train several models on A_{train} and use predictions on A_{validate} to choose parameters. In the second step we train on A_{train} combined with A_{validate} using the selected parameters and use predictions on A_{test} to estimate the performance of the model.

Choosing Parameters. A separate classifier was trained on each of the six regions and the number of clusters was set to $k = [5, 10, 15, 20, 25, 30]$ for each classifier, giving a total of 36 models. The performance of each model was estimated on A_{validate} by calculating mean absolute error (MAE) from the reference intervals and intraclass correlation (ICC). To calculate ICC we converted CMS predictions to interval midpoints and used the average interval midpoint of the

raters. MAE stabilized around 0.01 for all regions for $k \geq 20$. ICC was highest in the upper regions and values for the right and left upper regions are given in Table 2. ICC was poor in the lower (≤ 0.24) and middle (≤ 0.31) regions for all values of k. Prevalence of emphysema and rater agreement is generally highest in the upper regions (Average prevalence in upper, middle, lower: 26%, 20%, 12%. Average ICC in upper, middle, lower: 0.81, 0.65, 0.51). We focus on the upper regions in the following analysis because learning from the lower prevalence and rater agreement in the middle and lower regions appear to be a much harder problem, which we leave for future work.

Table 2. Intraclass correlation for parameter selection. The best values are shown along with the number of clusters in the model. ICC is calculated with a two-way model and measures consistency. Avg refers to the average of R1 and R2

Region	Number of clusters	Raters	ICC (CI)
Right upper	20	R1/R2	0.83 (0.79–0.87)
		Avg/CMS	0.62 (0.53–0.70)
Left upper	15	R1/R2	0.78 (0.72–0.83)
		Avg/CMS	0.53 (0.43–0.63)

Region Prediction. We use the selected parameters to train two new models on the combined data $A_{\text{combined}} = A_{\text{train}} \cup A_{\text{validate}}$. Performance of the four models, two trained on A_{train} and two trained on A_{combined}, is evaluated on A_{test} by calculating ICC, using the same procedure for converting predictions as for parameter selection. Performance scores are summarized in Table 3, and we see that ICC in the upper right region improves when training on the larger data set, while ICC decrease in the upper left region. We also note that performance in upper left on A_{test} is much lower than on A_{validate} indicating overfitting in the parameter selection.

Reduced data set for training A potential issue when applying CMS to this data is that the proportion of non-emphysema bags is large ($> 70\%$) and only very few bags have a label proportion larger than 25%. This gives a highly skewed data set where less than 10% of instances contain emphysema. It is possible that the skewed data makes it difficult to identify emphysema clusters because all clusters will contain mostly non-emphysema instances.

To investigate this hypothesis we re-run the above experiment, but use only bags with emphysema for training. This gives a less skewed data set, but the proportion of emphysema instances is still less than 25% . First we train on a reduced version of A_{train} and use performance on the full version of A_{validate} for parameter selection. Then we train a new model using the selected parameter

Table 3. Agreement between raters and model predictions on A_{test}. 95% confidence intervals are shown for ICC. ICC measures consistency and is calculated with a two-way model

Region	Raters	ICC on A_{test}	
		A_{train}	A_{combined}
Right upper	R1/R2	0.82 (0.76–0.86)	0.82 (0.76–0.86)
	R1/CMS	0.67 (0.59–0.74)	0.71 (0.63–0.77)
	R2/CMS	0.54 (0.44–0.64)	0.58 (0.48–0.66)
	Avg/CMS	0.64 (0.55–0.71)	0.67 (0.59–0.74)
Left upper	R1/R2	0.81 (0.75–0.85)	0.81 (0.75–0.85)
	R1/CMS	0.38 (0.25–0.49)	0.38 (0.25–0.49)
	R2/CMS	0.37 (0.24–0.49)	0.31 (0.18–0.43)
	Avg/CMS	0.40 (0.27–0.51)	0.36 (0.23–0.48)

on the reduced version of A_{combined} and measure performance on the full version of A_{test}.

Performance on A_{validate} is summarized in table 4 where we again see best performance in the upper right region. Performance on A_{test} is summarized in table 5 and again we see indication of overfitting in the parameter selection. Training on the reduced A_{combined} result in large improvements over training on the reduced A_{train}, beating performance when training on the full data.

Table 4. Intraclass correlation for parameter selection using reduced training data. Best values are shown with the number of clusters in the model. ICC is calculated with a two-way model and measures consistency. Avg refers to the average of R1 and R2

Region	Number of clusters	Raters	ICC (CI)
Right upper	30	R1/R2	0.83 (0.79–0.87)
		Avg/CMS	0.73 (0.65–0.79)
Left upper	20	R1/R2	0.78 (0.72–0.83)
		Avg/CMS	0.56 (0.46–0.65)

It is interesting to note that ICC between CMS and Avg is larger than ICC between CMS and any of the raters when training on the reduced data. Training on the full data shows highest ICC between CMS and R1 in three out of four cases. Estimates from R1 is generally a bit lower than from R2, so underestimating emphysema should give a better ICC with R1 than with R2 and Avg. This indicates that training on the reduced data overcomes a problem of underestimation present when training on the full data.

Table 5. Agreement between raters and model predictions on A_{test} using reduced training data. 95% confidence intervals are shown for ICC. ICC measures consistency and is calculated with a two-way model

Region	Raters	ICC on A_{test}	
		A_{train}	A_{combined}
Right upper	R1/R2	0.82 (0.76–0.86)	0.82 (0.76–0.86)
	R1/CMS	0.61 (0.52–0.69)	0.68 (0.60–0.75)
	R2/CMS	0.64 (0.55–0.71)	0.69 (0.61–0.75)
	Avg/CMS	0.66 (0.57–0.73)	0.72 (0.65–0.78)
Left upper	R1/R2	0.81 (0.75–0.85)	0.81 (0.75–0.85)
	R1/CMS	0.45 (0.33–0.56)	0.59 (0.49–0.67)
	R2/CMS	0.45 (0.33–0.55)	0.60 (0.50–0.68)
	Avg/CMS	0.47 (0.36–0.58)	0.63 (0.53–0.70)

Patch Prediction. We inspected patch predictions visually. Figure 1 shows slices and patch predictions for two subjects. Top row shows a case where raters and prediction agree and bottom row shows a case where prediction is larger than raters. In the case with agreement we see that patches classified as not emphysema contain little to no emphysema, while patches classified as emphysema contain large areas with clear tissue destruction. It appears that emphysema patches are in an area with a large degree of paraseptal emphysema, while not-emphysema patches are in an area with a small degree of centrilobular emphysema. In the case of larger predicted extent it appears that there is a small decrease in density in the upper part of the region compared to the lower part. The patches predicted as emphysema are in the upper part and appear to contain some tissue destruction.

4 Discussion and Conclusion

The agreement in the upper right region shows that CMS can estimate emphysema extent, which is clinically more relevant than predicting COPD presence considered in [10] and [3].

The performance improvement when training only on emphysema bags indicates that subsampling training data to achieve a more balanced data set is beneficial for CMS. The tendency to overfit, suggested by the performance decrease from A_{validate} to A_{test}, indicate that removing all non-emphysema bags is detrimental to performance. Future work could investigate how to determine the optimal mix of bags. It is possible that performance in the middle and lower regions could be improved in the same manner, but the very low prevalence in the lower regions could result in overfitting because the amount of training data is too small to be representative of the full data set. Another approach is to train on data from several regions, either by combining a couple of regions or using all six regions.

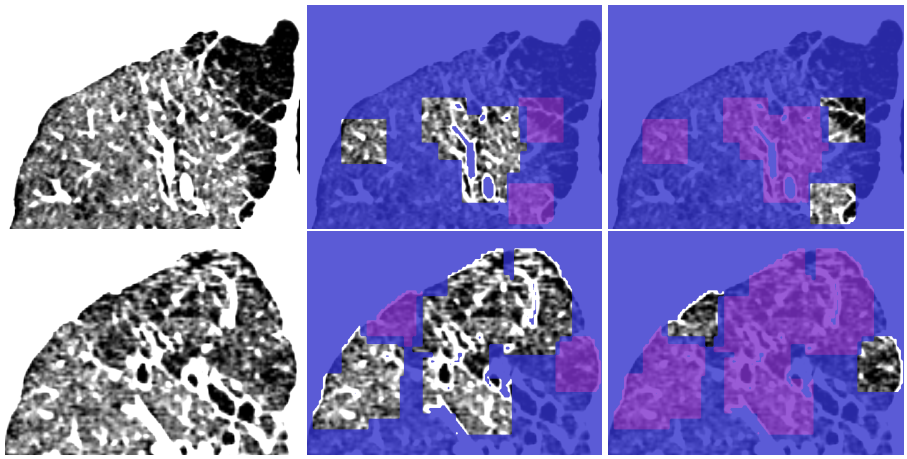

Fig. 1. Patch prediction in upper right region. Top: Rated as 26-50% extent, predicted as 26% extent. Bottom: Rated as 0% extent, predicted as 10% extent. Left: Intensity rescaled coronal slice. Center: Blue regions are not labeled. Purple patches are labeled as emphysema and non-colored patches as not emphysema. Right: Blue regions are not labeled. Purple patches are labeled as not emphysema and non-colored patches as emphysema

The increased performance when training on A_{combined} versus training on A_{train} indicates that improving performance could be a matter of increasing the amount of training data. However increasing the amount of training runs counter to one of the primary objectives of weak label learning, that of reducing the burden of labeling training data, and future work should consider the trade off between labeling burden and performance.

The inspected patch predictions show that patches with severe emphysema are likely to be labeled emphysema, while regions with mild emphysema tend to be labeled not emphysema. It is unlikely that we can account for the heterogeneity of emphysema with binary patch labels alone, and an alternative is to assign continuous labels indicating emphysema extent in the patch. This would allow us to rank patches and could be interesting as a tool for studying progression of emphysema. An interesting possibility suggested by the patch predictions for the region assessed as having 0% emphysema, is that the method is more sensitive to some mild cases of emphysema than the raters. If this is true, the approach could become a valuable tool for early detection of emphysema.

In this work we have focused on predicting emphysema extent without considering emphysema sub-type. Sub-type information is clinically interesting and a model that simultaneously predicts extent and sub-type is a future goal. Emphysema sub-types appear differently in CT scans, centrilobular emphysema is diffuse with small holes spread out over the affected area and paraseptal emphysema is more clearly defined with large bounded regions of complete tissue destruction. Simply extending cluster labels to {no-emphysema, centrilobular,

paraseptal} could improve performance, because it is likely that some patches with centrilobular emphysema are more similar to patches without emphysema than to patches with paraseptal emphysema.

There is, to our knowledge, no previous work that attempts to learn from the kind of visual assessment we consider here. The patch-based classifier from [2] uses a different labeling scheme with six classes (three severities of centrilobular, one panlobular, one pleural-based and one non-emphysema), and the evaluation metrics are also different making it difficult to compare. It appears that the biggest problem for [2] is distinguishing mild and severe cases of centrilobular emphysema. This suggests that replacing binary labels indicating presence with categorical labels indicating severity might not be enough to model emphysema severity and a continuous severity score could be the way forward.

In conclusion, we show that visual scoring of emphysema extent in regions can be used for training an LLP method to predict both region extent and presence of emphysema in patches. The results also show that predictions correlate poorly with raters when training on data where emphysema prevalence is very low and rater agreement is low to moderate.

References

1. Castaldi, P.J., Cho, M.H., San José Estépar, R., McDonald, M.L.N., Laird, N., Beaty, T.H., Washko, G., Crapo, J.D., Silverman, E.K.: Genome-wide association identifies regulatory Loci associated with distinct local histogram emphysema patterns. AM J RESP CRIT CARE 190(4), 399–409 (2014)
2. Castaldi, P.J., San José Estépar, R., Mendoza, C.S., Hersh, C.P., Laird, N., Crapo, J.D., Lynch, D.A., Silverman, E.K., Washko, G.R.: Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers. AM J RESP CRIT CARE 188(9), 1083–1090 (Aug 2013)
3. Cheplygina, V., Sørensen, L., Tax, D., Pedersen, J., Loog, M., de Bruijne, M.: Classification of COPD with Multiple Instance Learning. INT C PATT RECOG. pp. 1508–1513 (Aug 2014)
4. Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Müller, N.L., Remy, J.: Fleischner society: Glossary of terms for thoracic imaging. Radiology 246(3), 697–722 (2008)
5. Hansen, N., Ostermeier, A.: Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In: IEEE C EVOL COMPUTAT on. pp. 312–317. IEEE (1996)
6. Kuck, H., de Freitas, N.: Learning about individuals from group statistics. UNCERTAIN ARTIF INTELL. pp. 332–339. AUAI Press, Arlington, Virginia (2005)
7. Patrini, G., Nock, R., Caetano, T., Rivera, P.: (almost) no label no cry. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) ADV NEUR IN 27, pp. 190–198. Curran Associates, Inc. (2014)
8. Pedersen, J.H., Ashraf, H., Dirksen, A., Bach, K., Hansen, H., Toennesen, P., Thorsen, H., Brodersen, J., Skov, B.G., Døssing, M., Mortensen, J., Richter, K., Clementsen, P., Seersholm, N.: The Danish randomized lung cancer CT screening trial–overall design and results of the prevalence round. J THORAC ONCOL 4(5) (2009)

9. Stolpe, M., Morik, K.: Learning from label proportions by optimizing cluster model selection. LECT NOTES ARTIF INT, vol. 6913, pp. 349–364. Springer Berlin Heidelberg (2011)
10. Sørensen, L., Nielsen, M., Lo, P., Ashraf, H., Pedersen, J., de Bruijne, M.: Texture-based analysis of COPD: A data-driven approach. IEEE T MED IMAGING 31(1), 70–78 (Jan 2012)
11. Wille, M.M., Thomsen, L.H., Dirksen, A., Petersen, J., Pedersen, J.H., Shaker, S.B.: Emphysema progression is visually detectable in low-dose CT in continuous but not in former smokers. Eur Radiol 24(11), 2692–2699 (Nov 2014)
12. Yu, F.X., Liu, D., Kumar, S., Jebara, T., Chang, S.: $\propto$SVM for learning with label proportions. CoRR abs/1306.0886 (2013), http://arxiv.org/abs/1306.0886

2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

DETECTING EMPHYSEMA WITH MULTIPLE INSTANCE LEARNING

Silas Nyboe Ørting[1] *Jens Petersen*[1] *Laura H. Thomsen*[2] *Mathilde M W Wille*[3] *Marleen de Bruijne*[1,4]

[1]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[2]Department of Respiratory Medicine, Gentofte Hospital, Hellerup, Denmark
[3]Department of Diagnostic Imaging, Bispebjerg Hospital, Copenhagen, Denmark
[4]Biomedical Imaging Group Rotterdam, Departments of Radiology and Medical Informatics, Erasmus MC - University Medical Center Rotterdam, The Netherlands

ABSTRACT

Emphysema is part of chronic obstructive pulmonary disease, a leading cause of mortality worldwide. Visual assessment of emphysema presence is useful for identifying subjects at risk and for research into disease development. We train a machine learning method to predict emphysema from visually assessed expert labels. We use a multiple instance learning approach to predict both scan-level and region-level emphysema presence. We evaluate performance on 600 low-dose CT scans from the Danish Lung Cancer Screening Study and achieve an AUC of 0.82 for scan-level prediction and AUCs between 0.76 and 0.88 for region-level prediction.

Index Terms— Weak supervision, Emphysema, Multiple Instance Learning

1. INTRODUCTION

Emphysema is a lung pathology characterized by destruction of lung tissue and enlargement of airspaces in the lung. Emphysema is part of chronic obstructive pulmonary disease (COPD) a leading cause of mortality and morbidity world-wide [1]. Standard practice for assessment of emphysema is based on CT densitometry, however, visual assessment using CT may be more sensitive to emphysema development[2] and is a part of the PanCan model for lung cancer risk prediction[3]. Visual assessment of emphysema is time-consuming and suffers from inter-rater variability[2]. Automatic assessment of emphysema presence could provide more stable predictions at a much lower cost and thus be a valuable replacement of expert assessments.

In this work we present a multiple instance learning (MIL) algorithm to detect emphysema. The proposed method is trained to predict emphysema presence at scan level, as well as in six regions of the lungs. We investigate if information about emphysema at the scan-level is enough to train the method to predict region-level emphysema presence.

This study was financially supported by the Danish Council for Independent Research (DFF) and the Netherlands Organization for Scientific Research (NWO).

2. RELATED WORK

Multiple Instance Learning methods have previously been successfully used for COPD prediction[4] where simple MIL methods perform similarly to more complex methods. We have previously used a learning with label proportions approach for predicting emphysema extent[5] with some success. However training required a lot of extent annotations, which are laborious to produce. Here we focus on emphysema detection, which is simpler to annotate than extent. A method for emphysema detection was proposed by [6], using a bullae-score computed from connected regions of low attenuation. Several recent papers have proposed to train a convolutional neural network (CNN) to classify lung tissue patterns[7, 8]. However, these works focus on several interstitial lung disease patterns and emphysema is either not considered[7] or is only a small part of the data set[8]. Other works have focused on unsupervised discovery of emphysema subtypes[9, 10] which is more relevant for analysis of detected emphysema than for initial detection of emphysema.

3. MATERIALS & METHODS

3.1. METHOD

We consider regional emphysema detection as a multiple instance learning (MIL) problem. MIL is a weakly supervised machine learning setting where the goal is to predict the label of individual samples by learning from grouped samples with group labels. In MIL we refer to samples as instances and sets of samples as bags. In our context, an instance is a small volumetric patch of a chest CT scan and patches from the same scan are grouped into bags. The bags are given a binary label indicating if emphysema is visible in the scan and the objective is to predict both emphysema presence in the scans and which patches contain emphysema.

More formally, let $\mathcal{X}$ be an instance space, $\mathcal{Y}$ an instance label space, $\mathcal{Z}$ a bag label space and $\mathbf{b} = (\mathbf{x} \subseteq \mathcal{X}, z \in \mathcal{Z})$ a labeled bag of instances. We refer to the instances in $\mathbf{b}$ with $\mathbf{b}^{\mathbf{x}}$, the label of $\mathbf{b}$ with $\mathbf{b}^{z}$ and the unknown instance labels

978-1-5386-3636-7/18/$31.00

with $\mathbf{b}^{\mathbf{y}}$. For a set of m bags $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m\}$ the learning problem is

$$\arg\max_{\mathbf{Y},h,\Theta} \mathrm{P}(\mathbf{Y}, h, \Theta|\mathbf{B}), \quad (1)$$

where $\mathbf{Y} = \cup_{i=1}^{m} \mathbf{b}_i^{\mathbf{y}}$ is a labeling of instances, $\Theta : \mathcal{Y} \mapsto \mathcal{Z}$ is a bag labeling function relating $\mathbf{b}_i^{\mathbf{y}}$ to $\mathbf{b}_i^{z}$ and $h : \mathcal{X} \mapsto \mathcal{Y}$ is a hypothesis relating $\mathbf{b}_i^{\mathbf{x}}$ to $\mathbf{b}_i^{\mathbf{y}}$.

In this work we restrict our attention to logistic regression hypotheses

$$h(\mathbf{b}^{\mathbf{x}}) = \sigma(\mathbf{w}^T\mathbf{b}^{\mathbf{x}}) = (1 + \exp(-\mathbf{w}^T\mathbf{b}^{\mathbf{x}}))^{-1}, \quad (2)$$

and the mean bag label function

$$\Theta(\mathbf{b}^{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i, \quad (3)$$

where n is the number of instances in $\mathbf{b}$. The learning problem now becomes

$$\arg\max_{\mathbf{Y},\mathbf{w}} \mathrm{P}(\mathbf{Y}, \mathbf{w}|\mathbf{B}, \Theta). \quad (4)$$

Optimizing (4) is hard since it depends on both $\mathbf{Y}$ and $\mathbf{w}$. We use the simple MIL approach where instances are labeled with the corresponding bag label, i.e. $\mathbf{b}^{\mathbf{y}} = \mathbf{b}^{z} \cdot \mathbf{1}^n$. The learning problem is now a standard logistic regression problem.

3.2. DATA

We used data collected in the Danish Lung Cancer Screening Trial (DLCST) [11]. The screening arm of the study enrolled 2052 participants for annual low dose CT screening. Scan parameters described in [11] are reproduced below

> All CT scans of the study were performed on a MDCT scanner (16 rows Philips Mx 8000, Philips Medical Systems, Eindhoven, The Netherlands). Scans were performed supine after full inspiration with caudocranial scan direction including the entire ribcage and upper abdomen with a low dose technique, 120kV and 40 mAs. Scans were performed with spiral data acquisition with the following acquisition parameters: Section collimation 16 $\times$ 0.75 mm, pitch 1.5, rotation time 0.5 second[11].

We obtained visual assessment of emphysema from [2], where screening participants with at least two CT scans where selected for visual assessment (n=1990). Two experts assessed whether signs of emphysema were visible in the top, middle and lower regions of each left and right lung. The regions were defined as above carina, between carina and lower pulmonary vein, and below carina. We measured rater agreement on label c as the proportion of cases both raters assigned c, out of the number of cases at least one rater assigned c. Agreement on regional emphysema presence on the train set are summarized in table 1. We note that both prevalence and agreement increase as we move upwards in the lungs.

	LL	LM	LU	RL	RM	RU	Scan
Present	59	73	81	62	74	81	81
Absent	96	95	95	96	94	94	93
Prevalence	12	21	27	13	22	29	32

Table 1. Mean prevalence and rater agreement on regional emphysema presence in the three train sets. All numbers are percentages. LL=Left Lower, LM=Left Middle, LU=Left Upper, RL=Right Lower, RM=Right Middle, RU=Right Upper.

3.2.1. Data representation

We represented a lung CT scan as a set of 3D patches sampled from the six regions. A fixed patch size of approximately $11mm^3$ was used to match the size of the secondary lobule [12]. We sampled 100, possibly overlapping, patches randomly from each of the six regions. For each patch we extracted a set of multi-scale filter responses and used equalized histograms of the filter responses as the final representation of the patch. The filters used were Gaussian blur, gradient magnitude, eigenvalues of the Hessian, Laplacian of Gaussian, Gaussian curvature and the Frobenius norm of the Hessian. All filters were calculated at scales $1mm$, $2mm$ and $4mm$ resulting in 24 histograms with 13 bins each. The filters have previously been used successfully for COPD texture analysis [13], and further details can be found there.

The prevalence and appearance of emphysema varies between regions so we included an extra feature encoding which region an instance is sampled from.

4. EXPERIMENTS & RESULTS

We selected a test set of 600 subjects and a train set of 1200 subjects. The train set was further split into three sets of 400 subjects and we trained the models separately on each of the three datasets. This was done primarily to provide estimates of variance due to changes in training data and to estimate stability of predicted instance and bag labels, but we also combine them to create more stable ensemble predictions. In the following we refer to the three training sets as replications.

We derived reference region labels from visual scoring by assigning presence labels to a region if at least one rater indicated emphysema was present. Scans were assigned presence labels if at least one region had a presence label.

We used three variations of the proposed method: a model trained with scan labels; a model trained with region labels; and a model trained for each region separately with region labels. The last variation thus discarded the region feature. We refer to the three variations as "global" (G), "region" (R) and "separate" (S). We validated the predictive power of the method on scan and region predictions. Instance reference labels are not available, so it is not possible to directly validate instance predictions.

	LL	LM	LU	RL	RM	RU	Scan
G	78	81	83	76	82	88	80
R	80	81	83	76	82	87	80
S	74	81	84	76	82	87	82
G	80	83	85	77	83	89	82
R	80	82	84	77	83	88	81
S	78	83	84	77	83	87	82

Table 2. AUC ($\times 100$) of emphysema prediction . First three rows show mean AUC across replications. Last three rows show AUC for ensemble predictions.

	LL	LM	LU	RL	RM	RU	Scan
G	78	82	87	78	84	91	83
R	71	80	88	69	85	90	88
S	62	80	87	72	86	89	87

Table 3. AUC ($\times 100$) of ensemble predictions using instance threshold (0.5) prior to aggregating bag labels.

4.1. EMPHYSEMA DETECTION

Table 2 summarize the area under the receiver operating characteristic curve (AUC) of the three methods. The first three rows show mean AUC of the three replications and the last three rows show the AUC when the replications are used as an ensemble. We see that all three methods have similar performance on all regions and scan prediction, with most difference in the lower left (LL) region with AUC from 0.74 to 0.80. We also see that AUC correlates with lower, middle, and upper region and that the best AUC is achieved in the upper right region (0.89). Scan-level AUC is lower than the best region-level AUCs and seems to be negatively affected by the performance in the lower regions. We note that using the ensemble is beneficial in all cases, although improvement is only 1-2 percentage points. Using densitometry (RA950) to detect emphysema achieves an AUC of 0.67.

The mapping from instance labels to bag labels does not influence the model training. However, it can have a large influence on bag predictions. If for example, we alternatively convert instance predictions to binary labels by thresholding at 0.5 prior to aggregating instance labels into bag labels, we obtain the AUC values in table 3 (only ensemble predictions shown). We see that this results in more variation and more extreme values across regions (0.62 – 0.91), but also a large increase in scan AUC for R and S.

4.2. REGION FEATURES

We included region as feature for G and R and we are interested in how this feature is weighted. Based on the prevalence in table 1 we expect that an instance from a lower region is less likely to contain emphysema. Table 4 summarize the region weights across replications. The right upper (RU) region is the reference. We see that both G and R seem to capture the relationship between prevalence and upper, middle, lower region. The fact that G, without access to region labels, appears to capture the difference in prevalence across upper, middle, lower regions indicates the model is learning to discriminate emphysema and non-emphysema tissue.

model	LL	LM	LU	RL	RM	RU
G	-0.11	-0.04	0.07	-0.10	-0.05	0
G	-0.16	-0.12	0.01	-0.14	-0.10	0
G	-0.15	-0.09	0.03	-0.12	-0.09	0
R	-1.36	-0.53	-0.07	-1.23	-0.41	0
R	-1.20	-0.53	-0.12	-1.16	-0.48	0
R	-1.40	-0.61	-0.09	-1.23	-0.50	0

Table 4. Fitted weights of region features. One row for each replication of G and R.

4.3. STABILITY OF PREDICTIONS

We assess the variation in predictions arising from differences in training data. Bag predictions are continuous values in the interval $(0, 1)$ and we measure stability with intraclass correlation (ICC, two-way model, agreement) and median absolute deviation (MAD). ICC is calculated between each pair of replications and we report the mean. MAD is calculated for each triplet of replicated predictions and we report the 90'th percentile. The 90'th percentile is chosen to provide an upper bound on the most common deviation. The ICC measure informs us about the linearity of predictions and accounts for systematic differences between replications. The MAD measure informs us about the magnitude of the difference between replications. We can interpret the numerical value of the MAD measure as *for 90% of the bags, the largest deviation between any two replications is at most 2 $\times$ MAD* . Table 5 shows the measures. All three models have increasing ICC going from lower over middle to upper regions which could be related to the prevalence of emphysema in the regions. R has higher ICC and lower mad than G in all cases except LL. R has similar or higher ICC and similar or lower mad as S in all cases. G has higher ICC than S in three cases and S has higher ICC in four cases. The mad measure for G and S is similar in all cases. This suggests that having access to region labels and training on all regions simultaneously yields a more stable model. This is probably because, training a single model on all six regions versus training a model on each region increases the amount of training data and decreases the number of parameters by a factor of six.

5. DISCUSSION & CONCLUSION

We proposed to use a MIL approach to predict regional and scan-level emphysema presence. We achieved an AUC of 0.82 at scan-level, which is comparable to the bullae-score

	LL	LM	LU	RL	RM	RU	Scan
G	.81	.83	.87	.82	.82	.90	.84
R	.74	.84	.93	.85	.89	.95	.90
S	.45	.75	.94	.74	.87	.96	.91
G	.026	.028	.033	.026	.027	.032	.029
R	.017	.027	.024	.015	.023	.022	.022
S	.027	.025	.030	.020	.025	.035	.028

Table 5. Stability of bag predictions. First three rows show ICC, last three rows show mad.

approach [6] that achieved AUC of 0.82 on a dataset with substantially higher emphysema prevalence. We showed that scan-level AUC can be improved from 0.82 to 0.88 by making a binary classification of instances before aggregating into bag labels. However, this comes at the cost of increased difference in performance between regions. A possible explanation is that the low prevalence in the lower regions yields a larger ratio of spurious to actual detections in the region, leading to decreased scan-level performance. We saw that learned region feature weights correlate with prevalence, so it is likely that predictions in the lower regions are generally lower. Imposing a threshold will then suppress more detections in the lower regions, thereby improving the overall detection performance at the cost of missed detections in the lower regions.

We have shown that learning from visual scoring of emphysema is feasible and can produce good predictions of both scan-level and region-level emphysema presence. We further found that training on scan-level labels achieves performance similar to training on region-level labels. This could substantially reduce the burden of producing training data.

6. REFERENCES

[1] "Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2015," 2015.

[2] M. M. Wille, L. H. Thomsen, A. Dirksen, J. Petersen, J. H. Pedersen, and S. B. Shaker, "Emphysema progression is visually detectable in low-dose CT in continuous but not in former smokers," *Eur Radiol*, vol. 24, no. 11, pp. 2692–2699, Nov 2014.

[3] A McWilliams, M C. Tammemagi, J R. Mayo, H Roberts, G Liu, K Soghrati, K Yasufuku, S Martel, F Laberge, M Gingras, S Atkar-Khattra, C D. Berg, K Evans, R Finley, J Yee, J English, P Nasute, J Goffin, S Puksa, L Stewart, S Tsai, M R. Johnston, D Manos, G Nicholas, G D. Goss, J M. Seely, K Amjadi, A Tremblay, P Burrowes, P MacEachern, R Bhatia, M Tsao, and S Lam, "Probability of cancer in pulmonary nodules detected on first screening ct," *New England Journal of Medicine*, vol. 369, no. 10, pp. 910–919, 2013.

[4] V. Cheplygina, L. Sørensen, D. M. J. Tax, J. H. Pedersen, M. Loog, and M. d. Bruijne, "Classification of copd with multiple instance learning," in *International Conference on Pattern Recognition*, 2014, pp. 1508–1513.

[5] S N Ørting, J Petersen, M MW Wille, L H Thomsen, and M de Bruijne, "Quantifying emphysema extent from weakly labeled ct scans of the lungs using label proportions learning," *MICCAI PIA*, pp. 31–42, 2016.

[6] R Wiemker, M Sevenster, H MacMahon, F Li, S Dalal, A Tahmasebi, and T Klinder, "Automated assessment of imaging biomarkers for the pancan lung cancer risk prediction model with validation on nlst data," in *Proc.SPIE*, 2017, vol. 10134.

[7] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE TMI*, vol. 35, no. 5, pp. 1207–1216, May 2016.

[8] Q. Wang, Y. Zheng, g. yang, W. Jin, X. Chen, and y. yin, "Multi-scale rotation-invariant convolutional neural networks for lung texture classification," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, no. 99, pp. 1–1, 2017.

[9] P Binder, N K. Batmanghelich, R S J Estepar, and P Golland, *Unsupervised Discovery of Emphysema Subtypes in a Large Clinical Cohort*, pp. 180–187, Springer International Publishing, Cham, 2016.

[10] 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*, pp. 116–124, Springer International Publishing, Cham, 2017.

[11] J H. Pedersen, H Ashraf, A Dirksen, K Bach, H Hansen, P Toennesen, H Thorsen, J Brodersen, B G Skov, M Døssing, J Mortensen, K Richter, P Clementsen, and N Seersholm, "The Danish randomized lung cancer CT screening trial–overall design and results of the prevalence round," *Journal of Thoracic Oncology*, vol. 4, no. 5, 2009.

[12] D M. Hansell, A A. Bankier, H MacMahon, T C. McLoud, N L. Müller, and J Remy, "Fleischner society: Glossary of terms for thoracic imaging," *Radiology*, vol. 246, no. 3, pp. 697–722, 2008.

[13] L. Sørensen, M. Nielsen, Pechin Lo, H. Ashraf, J.H. Pedersen, and M. de Bruijne, "Texture-based analysis of COPD: A data-driven approach," *IEEE TMI*, vol. 31, no. 1, pp. 70–78, 2012.

Learning to quantify emphysema extent: What labels do we need?

Silas Nyboe Ørting, Jens Petersen, Laura H. Thomsen, Mathilde M. W. Wille and Marleen de Bruijne *Member, IEEE,*

***Abstract*—Accurate assessment of pulmonary emphysema is crucial to assess disease severity and subtype, to monitor disease progression and to predict lung cancer risk. However, visual assessment is time-consuming and subject to substantial inter-rater variability and standard densitometry approaches to quantify emphysema remain inferior to visual scoring. We explore if machine learning methods that learn from a large dataset of visually assessed CT scans can provide accurate estimates of emphysema extent. We further investigate if machine learning algorithms that learn from a scoring of emphysema extent can outperform algorithms that learn only from a scoring of emphysema presence. We compare four Multiple Instance Learning classifiers that are trained on emphysema presence labels, and five Learning with Label Proportions classifiers that are trained on emphysema extent labels. We evaluate performance on 600 low-dose CT scans from the Danish Lung Cancer Screening Trial and find that learning from emphysema presence labels, which are much easier to obtain, gives equally good performance to learning from emphysema extent labels. The best classifiers achieve intra-class correlation coefficients around 0.90 and average overall agreement with raters of 78% and 79% on six emphysema extent classes versus inter-rater agreement of 83%.**

I. Introduction

EMPHYSEMA is a lung pathology characterized by destruction of lung tissue and enlargement of airspaces in the lung, causing shortness of breath. It is a main component of chronic obstructive pulmonary disease (COPD), a leading cause of mortality and morbidity world-wide [1]. Emphysema can be assessed on chest CT scans and its extent quantified by densitometry, where the amount of tissue affected by emphysema is estimated by measuring the percentage of lung volume with attenuation below a specific threshold. Although densitometry is simple and provides a single interpretable measurement of emphysema extent, it is also highly dependent on scanner hardware, reconstruction parameters [2] and software used for analysis [3].

This study was financially supported by the Danish Council for Independent Research (DFF) and the Netherlands Organization for Scientific Research (NWO). The sponsors had no involvement in the work.

S. Ørting and J. Petersen are with Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

L. Thomsen is with Department of Internal Medicine, Hvidovre Hospital, Copenhagen Denmark

M. M. W. Wille is with Department of Diagnostic Imaging, Bispebjerg Hospital, Copenhagen, Denmark

M. de Bruijne is with Department of Computer Science, University of Copenhagen, Copenhagen, Denmark and Biomedical Imaging Group Rotterdam, Departments of Radiology and Medical Informatics, Erasmus MC - University Medical Center Rotterdam, The Netherlands

Manuscript received January 21st, 2019

An alternative to densitometry is visual assessment that can quantify extent and characterize emphysema subtype. The COPDGene CT Workshop Group [4] proposed a standard for visual assessment of COPD based on the characterization of emphysema appearance from the Fleischner society [5]. A slightly modified version of the standard was used for visual assessment in the Danish Lung Cancer Screening Trial (DLCST), where it was shown to be predictive of lung cancer [6]. A similar classification scheme defined in [7] was used in [8] where it was shown that visual presence and severity of emphysema is associated with increased mortality independent of densitometric measures of emphysema severity. The downside of visual assessment is that it is time-consuming and subject to inter-rater variability [4], [9].

Automated approaches based on the appearance of emphysema could provide fast and reproducible assessment of emphysema extent, location and sub-type, thus combining the superior disease characterization of visual assessment with the ease of densitometry. For instance [10] has shown that a shape-model of bullae-like structures can be used for emphysema detection. We have previously used machine learning algorithms based on texture features to predict regional emphysema presence [11] and emphysema extent [12]. Other learning based approaches have focused on discovery of emphysema patterns using supervised [13] and unsupervised [14], [15] learning, COPD detection and staging [16], [17] and emphysema detection in the more general context of interstitial lung disease classification [18], [19].

Multiple Instance Learning (MIL) has been used with success in a number of the prior works on emphysema and COPD detection [11], [16], [17] and for many related medical image analysis tasks as reviewed in [20]. MIL is a learning setting where the objects of interest are represented by a collection of samples. Each collection has a binary label and the goal is to learn which samples in a collection are "responsible" for the label. MIL has been very succesful at detecting presence of abnormalities. However, visual assessment systems for lung disease, such as those developed for COPD [4], give estimates of affected lung tissue that is better captured by proportion labels. Label Proportions Learning (LLP) is the natural extension of MIL to cases where labels are proportions, but despite the success of MIL, LLP has seen almost no usage in medical imaging.

In this work we present the largest comparison yet of machine learning methods for assessing emphysema extent, extending our previous work on emphysema presence prediction [11], where a MIL method was used for regional

emphysema detection, and our work on extent prediction [12], where the LLP method Cluster Model Selection was used for regional emphysema extent prediction. We compare four MIL methods, of which three have not been used for emphysema detection before, and five LLP methods, of which four have not been used for emphysema detection or in medical imaging before. We investigate if learning from emphysema extent labels improves performance over learning from emphysema presence labels. Knowing what can be achieved by learning from labels of different quality and cost is paramount for cost-effective development and application of machine learning methods for clinical decision making.

II. Materials and methods

We view emphysema extent prediction as a bag learning problem. Bag learning is a machine learning setting where we are given a set of instances, a partition of the instances into bags and a labeling of the bags. The objective is to learn to predict both instance and bag labels for unseen data. In this work we view a region of the lung as a bag and patches sampled from the region as instances. The bag labels are regional emphysema extent scores, corresponding to estimated percentage of affected lung volume, and we wish to predict which patches contain emphysema, as well as the extent of emphysema in the region. Representing a scan as a set of patches provides a representation of local patterns in the lungs. By controlling the patch size we can focus on the scale at which patterns are expected to be distinct.

More formally, let $\mathcal{X}$ be an instance space, $\mathcal{Y}$ an instance label space, $\mathcal{Z}$ a bag label space and $\mathbf{b} = (\mathbf{x} \subseteq \mathcal{X}, z \in \mathcal{Z})$ a labeled bag of instances. We use superscripts to refer to the label ($\mathbf{b}^z$), instances ($\mathbf{b}^\mathbf{x}$) and instance labels ($\mathbf{b}^\mathbf{y}$) associated with a bag $\mathbf{b}$. For a set of m bags $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m\}$, $\mathbf{b}_i^\mathbf{x}$ are the instances in the i'th bag and $\mathbf{b}_{ij}^\mathbf{x}$ is the j'th instance in the i'th bag. We define the learning problem as

$$\arg\max_{\mathbf{Y},h,\Theta} \mathrm{P}(\mathbf{Y}, h, \Theta|\mathbf{B}), \tag{1}$$

where $\mathbf{Y} = \cup_{i=1}^m \mathbf{b}_i^\mathbf{y}$ is a labeling of instances, $\Theta : \mathcal{Y} \mapsto \mathcal{Z}$ is a bag labeling function relating $\mathbf{b}_i^\mathbf{y}$ to $\mathbf{b}_i^z$ and $h : \mathcal{X} \mapsto \mathcal{Y}$ is a hypothesis relating the instances $\mathbf{b}_i^\mathbf{x}$ to the corresponding instance labels $\mathbf{b}_i^\mathbf{y}$, i.e. h is a method for predicting $\mathbf{b}_i^\mathbf{y}$ from $\mathbf{b}_i^\mathbf{x}$.

Two well known bag learning settings are multiple instance learning (MIL) and learning with label proportions (LLP). In the standard MIL setting bag labels are binary, instance labels are binary and bag labels are related to instance labels by the max rule, i.e. a bag is positive if at least one instance is positive

$$\mathbf{b}_i^z = \Theta_{\max}(\mathbf{b}_i^\mathbf{y}) = \max_j \mathbf{b}_{ij}^\mathbf{y}. \tag{2}$$

This MIL setting is powerful because it allow us to learn about instance labels when only little information about the relation between instance and bag labels is available. A potential issue with the max rule is that it focuses on the single most discriminative instance. This could lead to a situation with good bag-level detection but poor localization and extent prediction. Including information about the proportion of positive instances could improve localization and extent prediction. In the standard LLP setting, bag labels are proportions, instance labels are binary and bag labels are related to instance labels by the mean rule, i.e. the bag label is the proportion of positive instances

$$\mathbf{b}_i^z = \Theta_{\mathrm{mean}}(\mathbf{b}_i^\mathbf{y}) = \frac{1}{|\mathbf{b}_i|} \sum_j^{|\mathbf{b}_i|} \mathbf{b}_{ij}^\mathbf{y}. \tag{3}$$

Although MIL methods require binary labels for training, i.e. $\Theta : \mathcal{Y} \mapsto \{0, 1\}$, we can use Θ_{mean} at test time to obtain proportion estimates of emphysema extent.

A. Methods

We compared four MIL methods (logistic, SVM, mi-logistic, mi-SVM) and five LLP methods (beta, Cluster Model Selection, $\propto$-SVM, $\propto$-logistic, Laplacian Mean Map). The methods can be grouped into three distinct strategies used to solve the bag learning problem: the simple strategy, the relabeling strategy and the mean strategy. Some methods have previously been successfully applied to emphysema and COPD prediction, logistic, SVM and mi-SVM in [17], [11] and Cluster Model Selection in [12]. The LLP methods, $\propto$-SVM [21] and Laplacian Mean Map [22], have been shown to perform well on a variety of datasets. The beta method [23] can be seen as an LLP version of logistic and the mi-logistic and $\propto$-logistic methods are logistic regression versions of their SVM counterparts.

a) Simple strategy: In the simple strategy the bag learning problem is solved by ignoring intra-bag dependencies. We assign each instance the label of the bag it came from, i.e $\mathbf{b}^\mathbf{y}{}_{ij} = \mathbf{b}_i^z$, and train a standard supervised method on the instance labels. Labels for unseen bags are predicted by predicting instance labels and using Θ_{mean} to derive a bag label. The learning problem now becomes

$$\arg\max_{\phi} \mathrm{P}(h_\phi|\mathbf{Y}, \mathbf{X}), \tag{4}$$

where $\mathbf{X} = \cup_{i=1}^m \mathbf{b}^\mathbf{x}{}_i$ is the set of instances and h is a model parameterized by ϕ. We consider two simple MIL models, logistic regression (log) and a support vector machine (svm); and one simple LLP model, beta regression [23] (beta). Beta regression is a generalized linear model where the outcome $\mathbf{Y}$ follows a beta distribution allowing us to perform regression with proportion outcomes. Note that bag labels are only used for the initial instance labeling, so Θ plays no role in the simple strategy.

b) Relabeling strategy: In the relabeling strategy the bag learning problem is solved by splitting it into two sub problems that are solved separately, a standard learning problem (5) and an instance labeling problem (6),

$$\arg\max_{\phi} \mathrm{P}(h_\phi|\mathbf{Y}, \mathbf{X}) \tag{5}$$

$$\arg\max_{\mathbf{Y}} \mathrm{P}(\mathbf{Y}|h_\phi, \Theta, \mathbf{Z}), \tag{6}$$

where $\mathbf{Z} = \cup_{i=1}^m \{\mathbf{b}_i^z\}$ is the set of bag labels and $\Theta = \Theta_{\max}$ for MIL and $\Theta = \Theta_{\mathrm{mean}}$ for LLP. The two sub problems are iterated until convergence, with the result of (5) being

used for (6) and the result of (6) being used for (5). We consider two relabeling MIL methods, mi-SVM [24] (misvm) and mi-logistic (milog); and three relabeling LLP methods, $\propto$-SVM [21] (psvm), $\propto$-logistic (plog) and Cluster Model Selection [25] (cms). The methods milog and plog have not previously been published, they are however very similar to their svm counterparts and we do not include the derivation here. Details can be found in Appendix B. The cms algorithm differs from the other relabeling methods in that it solves (5) by unsupervised clustering. We use a version of cms previously described in [12].

c) Mean strategy: In the mean strategy the bag learning problem is solved by replacing the direct dependence on instance labels with a dependence on a mean statistic $\boldsymbol{\mu}$ calculated over all instances

$$\arg\max_{\phi} \mathrm{P}(h_\phi|\boldsymbol{\mu}, \mathbf{X}). \tag{7}$$

$\boldsymbol{\mu}$ is defined as

$$\boldsymbol{\mu} = \frac{1}{n}\sum_i \mathbf{Y}_i\mathbf{X}_i \tag{8}$$

where $\mathbf{Y}_i \in \{-1, 1\}$ and n is the number of instances. Knowing $\boldsymbol{\mu}$ allow us to minimize the expected risk of a large class of loss functions. However, since the instance labels $\mathbf{Y}$ are still unknown $\boldsymbol{\mu}$ must be estimated. The basic idea for the mean strategy is to express $\boldsymbol{\mu}$ in terms of bag-wise averages and solve for these bag-wise averages

$$\boldsymbol{\mu} = \sum_{i=1}^{m} \frac{|\mathbf{b}_i|}{n}\boldsymbol{\mu}_i \tag{9}$$

$$\boldsymbol{\mu}_i = \mathbf{b}_i^z\boldsymbol{\mu}_i^+ - (1 - \mathbf{b}_i^z)\boldsymbol{\mu}_i^- \tag{10}$$

where $|\mathbf{b}_i|$ the number of instances in bag i and $\boldsymbol{\mu}_i, \boldsymbol{\mu}_i^+, \boldsymbol{\mu}_i^-$, are the unknown mean instance, mean positive instance and mean negative instance of bag i, respectively. Equation (10) yields an underdetermined system of equations. We consider a single mean LLP method, Laplacian Mean Map [22] (lmm), that solves the system of equations by regularizing with a bag similarity term. We refer to [22] for further details.

B. Measures

We measure agreement in the following way. Let n_k be the number of ratings for case k and $n_{c,k}$ the number of times label c is assigned to case k. Agreement on label c over all cases is defined as

$$\frac{\sum_k n_{c,k}\left(n_{c,k} - 1\right)}{\sum_k n_{c,k}\left(n_k - 1\right)}. \tag{11}$$

Overall agreement across labels is defined as

$$\frac{\sum_{c,k} n_{c,k}\left(n_{c,k} - 1\right)}{\sum_k n_k\left(n_k - 1\right)}. \tag{12}$$

When all cases have two ratings Equation 11 corresponds to the Jaccard similarity and Equation 12 corresponds to multi-class accuracy. For multiple raters these measures ensure that partial agreement, e.g. two out of three, is counted appropriately. We measure prevalence of label c as the proportion of times a case is assigned label c out of all assignments.

$$\frac{\sum_k n_{c,k}}{\sum_{c,k} n_{c,k}} \tag{13}$$

C. Data

Examples of the appearance of emphysema in CT scans are provided in Appendix A.

1) Study population, CT scanning & visual assessment: We used data collected in the Danish Lung Cancer Screening Trial (DLCST) [26]. The screening arm of the study enrolled 2052 participants for annual low dose CT screening. Scan parameters are reproduced below verbatim from [26].

> All CT scans of the study were performed on a MDCT scanner (16 rows Philips Mx 8000, Philips Medical Systems, Eindhoven, The Netherlands). Scans were performed supine after full inspiration with caudocranial scan direction including the entire ribcage and upper abdomen with a low dose technique, 120kV and 40 mAs. Scans were performed with spiral data acquisition with the following acquisition parameters: Section collimation 16 $\times$ 0.75 mm, pitch 1.5, rotation time 0.5 second.

We used a 1mm reconstruction with pixel size of 0.78mm $\times$ 0.78mm.

We obtained visual assessment of emphysema from [9], where screening participants with at least two CT scans were selected for visual assessment (n=1990). The visual assessment used a slight modification of the assessment sheets from [4]. Baseline and final followup scan was assessed by two experts. Emphysema extent was assessed for the top, middle and lower regions of each lung. The regions were defined as above carina, between carina and lower pulmonary vein, and below lower pulmonary vein. Each region was assigned a score of 0%, 1-5%, 6-25%, 26-50%, 51-75% or 76-100% indicating the extent of emphysema in the region.

In general, prevalence was highest and rater agreement best in the upper regions. Prevalence and agreement for the upper right region are summarized in Table I. Prevalence for emphysema extent above 26% is low ($\approx$ 36 of 1200 subjects). Agreement on the five categories indicating emphysema presence was around 50%. Using only two categories (0%, $\geq$ 1%) improves agreement to 82% on the emphysema category. Although the original six categories provide more information than presence/absence labels, they are noisier and likely harder to learn from.

2) Patches: We represented a lung region as a collection of 3D patches sampled from the region. Sampling was done by choosing patch center locations uniformly at random within the region. We used a fixed patch size of approximately $11\mathrm{mm}^3$ to match the size of the secondary lobule [5] and allowed overlapping patches. For each patch we extracted a set of multi-scale filter responses and used equalized histograms of the filter responses as the final representation of the patch.

All			Presence		
Extent	Agreement	Prev	Extent	Agreement	Prev
0%	94 (93–95)	75.2	0%	94 (93–95)	75.2
1-5%	54 (47–60)	14.7			
6-25%	44 (34–53)	7.0			
26-50%	45 (26–61)	2.0	$\geq 1\%$	81 (78–85)	24.8
51-75%	57 (26–80)	0.9			
$\geq 76\%$	67 (00–99)	0.1			
Overall	83 (81–85)			91 (89–92)	

TABLE I: Agreement and mean prevalence in the upper right region of the training data. Numbers are percentages. First three columns are for all six categories, last three columns are for presence/absence. 95% confidence intervals for agreement estimated by bootstrapping are given in parenthesis.

The filters used were Gaussian blur, gradient magnitude, eigenvalues of the Hessian, Laplacian of Gaussian, Gaussian curvature and the Frobenius norm of the Hessian. All filters were calculated at scales 1mm, 2mm and 4mm. The filters and the patch sampling strategy have previously been used successfully for COPD texture analysis in [16].

III. Experiments and results

We created a set of 1800 bags by sampling patches from the upper right region of 1800 subjects, such that each bag corresponds to one unique subject. We chose the upper right region because it has the highest prevalence and agreement. Results in [11] indicate that although absolute performance decreases when training on regions with lower prevalence and agreement, this decrease is relatively smaller than the decrease in rater agreement and prevalence.

Each bag contained 100 patches from a single subject. The bags were split into three non-overlapping datasets of 400 training and 200 test bags. Each experiment was run on all three datasets. In each split, we used two-fold cross validation on the training bags for parameter tuning. The three separate sets of classifiers were finally trained on all 400 training bags and performance estimated on the corresponding 200 test bags.

All classifiers provide posterior instance label probabilities which were converted to binary predictions using a classifier specific instance threshold fitted on the training bags. Parameters are summarized in Appendix C.

To train and evaluate we derived point estimates of emphysema extent by converting visually assessed extent intervals to interval midpoints and taking the mean over both raters. As an example, for a region with ratings 6-25% and 1-5%, the ratings are converted to 15.5% and 3% and combined into 9.25%. The point estimates where used directly for training LLP classifiers and thresholded at zero to obtain binary labels for training MIL classifiers.

A. Extent prediction accuracy

The prediction performance of the nine classifiers is illustrated with correlation plots in Fig. 1. The numbers in the title of each plot are intra-class correlation coefficients (ICC, two-way model, agreement) for each replication. The average ICC coefficients over the three replications are shown in Table II. We see clear positive correlation between reference and predicted extent for all classifiers. It appears that plog and cms tend to underestimate extent, whereas psvm tends to overestimate for cases with low extent but seems to perform very well for larger extent. Most classifiers show the largest variation for 15% reference extent. For extent larger than 15% we see very few cases with 0% extent predicted. The ICC values across replications, also seen in Fig. 1, illustrate that the performance of some classifiers varies a lot, with a difference of 0.25 in the worst case (cms). The most stable ICC performance is seen for lmm, which also has the highest average ICC.

log	svm	milog	misvm	beta	plog	psvm	cms	lmm
0.88	0.86	0.90	0.89	0.89	0.69	0.71	0.78	0.91

TABLE II: Average ICC of of emphysema extent over the three replications. MIL on the left, LLP on the right.

	0	1-5	6-25	26-50	51-75	76-100	Overall
log	88	35	54	38	17	00	74
svm	89	37	49	27	12	00	74
milog	85	35	50	36	29	00	71
misvm	91	39	58	36	31	00	79
beta	91	35	54	24	47	00	78
plog	72	26	57	45	00	00	58
psvm	27	15	21	35	51	17	24
cms	62	22	49	28	37	00	49
lmm	81	31	49	30	44	17	66
Rater	95	49	53	47	32	00	83

TABLE III: Agreement percentages between classifiers and raters averaged over replications and raters. First four columns show MIL classifiers, next five columns show LLP classifiers, last column shows rater agreement.

B. Replacing a rater

The ICC of predicted extent and average rater extent provides an overall measure of performance and a validation that the classifiers have learned what they are trained to do. We are also interested in how the classifiers compare against each rater on the original rater task, i.e assign one of six intervals of emphysema extent. We converted predicted extent into the six extent intervals and calculated agreement with each rater. Agreement was calculated as described in Section II-B and is reported in Table III as an average over raters and replications. The final column in Table III provides inter-rater agreement averaged over replications. We see that beta and misvm have the highest overall agreement (78 and 79), which is not far from the overall rater agreement of 83. The agreement pattern of misvm and beta also seem to match that of the raters to a large degree, with a large agreement on 0% extent cases. It is interesting that psvm has the worst overall performance yet seems to outperform the other classifiers and raters for 51-75% extent. However, we cannot rule out that this is just a random coincidence given the low prevalence of that class. Another interesting observation is that the best results relative to inter-rater agreement is seen for 6-25% and 51-75%, with four classifiers having better agreement scores than inter-rater agreement.

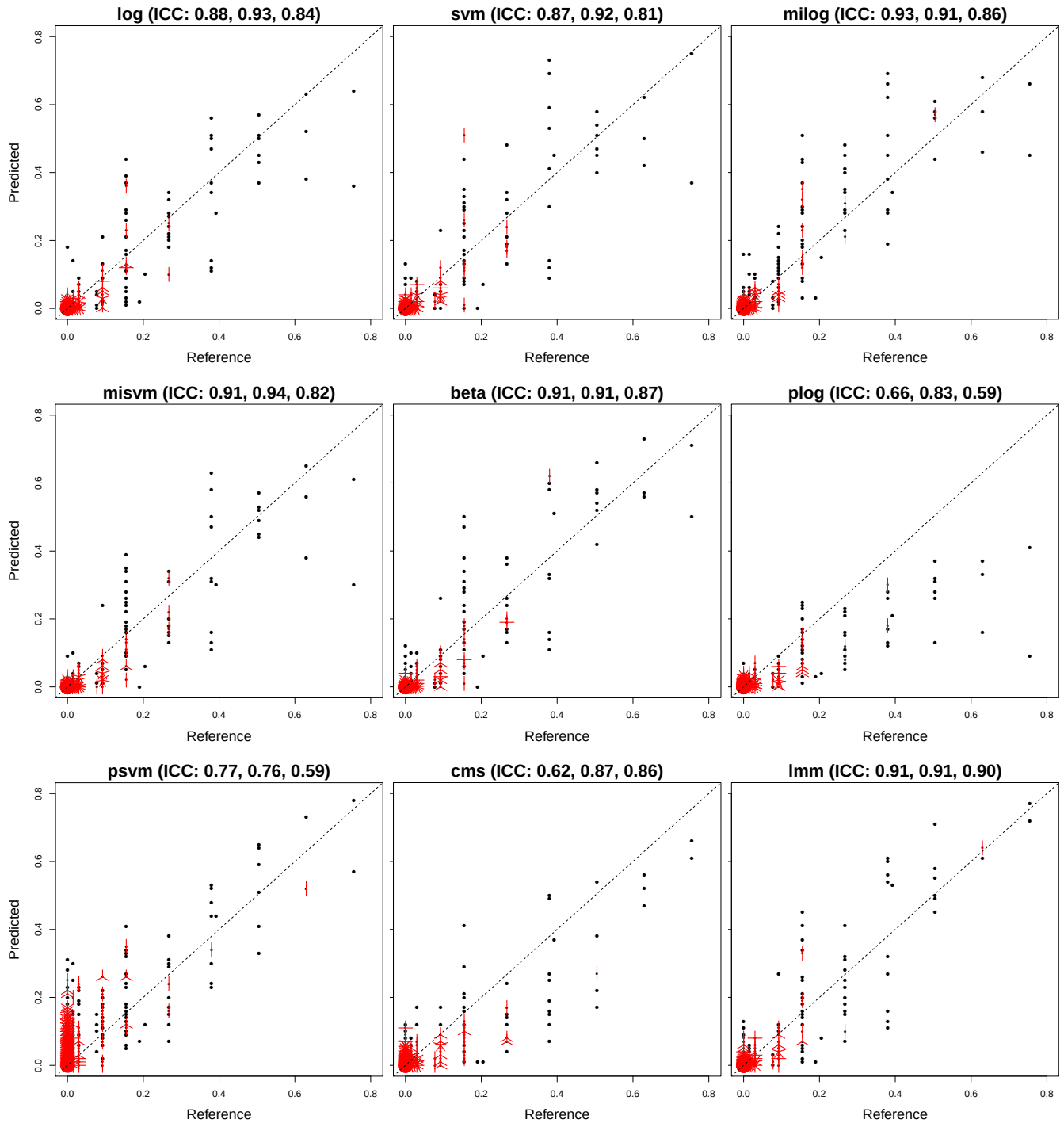

Fig. 1: Correlation between predicted and reference extent of emphysema. The x-axis is reference extent and the y-axis is predicted extent. The amount of red "petals" at a coordinate indicates the amount of coincident points. Plot titles show ICC coefficients for each replication.

C. *Ranking classifiers*

We use Friedmans and Nemenyis test for comparing classifiers as suggested in [27]. We test the hypothesis H_0 : *All classifiers are equal* using Friedmans test and significance level $\alpha = 0.05$. This test is based on the rank of the classifiers for each sample prediction. We use the absolute distance from predicted extent to reference extent to assign ranks. In all three replications we get $p < 0.001$ for the Friedman test and reject the hypothesis that all classifiers have equal performance. We then test the pairwise hypothesis H_0 : *The classifiers are equal* for all pairs of classifiers using the Nemenyi test. The results of the Nemenyi tests are summarized in Fig. 2. Columns are

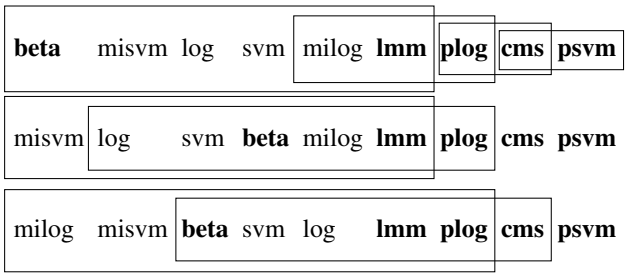

Fig. 2: Grouping of classifiers based on difference in extent prediction performance as decided by the Nemenyi test. Classifiers in the same box are not significantly different ($\alpha = 0.05$). Columns are sorted by mean rank over all test samples in descending order. **Bold** typeface indicates LLP methods.

sorted by average ranks and H_0 is rejected for classifiers that are not in the same box. We see that the LLP methods plog, cms and psvm are consistently ranked low, confirming the low ICC in Table II and the low overall agreement in Table III. Even though lmm is never significantly different from the best classifier, it is consistently ranked low. We also saw in Table III that lmm had low overall agreement with raters yet achieved the best average ICC. It is also interesting that misvm is consistently ranked in the top-2.

1) Label stability: We investigate label stability under changes in training data by predicting all test data with the trained models from each replication. For each classifier we got three sets of predictions of 60,000 instances and 600 bags. We converted bag predictions to the six extent intervals and measured agreement between replications for predicted bag and instance labels for each classifier. The stability results are summarized in Table IV. For bag labels, most classifiers have best agreement on 0% extent followed by 6-25%. Overall, beta and misvm are the most stable classifiers for both bag and instance labels, whereas milog is the most stable classifier on 6-25%, 26-50% and 51-75%. The missing scores for 51-75% and 76-100% are because there are no predictions of these classes in any of the replications. The inter-rater agreement on bag labels is included in the last row of Table IV. We see that misvm and beta always have equal or better agreement than the raters, and most methods have better agreement than raters on all non-zero extent scores.

IV. Discussion & Conclusion

We have focused on comparing MIL methods, which have previously shown promising results for COPD and emphysema detection, with LLP methods that can learn directly from proportion labels. While end-to-end learning using CNNs have shown promising results for medical imaging tasks, and have just recently been used for emphysema quantification [28], we decided to use classic scale space features to focus on the aspects of learning from binary versus proportion labels, and to establish performance of classic feature engineering approaches.

Using the average rater as reference, the best classifiers achieve ICC coefficients around 0.9. Average overall agreement between the best classifiers and each rater on six emphysema extent intervals is close to the inter-rater agreement (78-79% vs 83%). For some extent intervals the classifiers are better than the inter-rater agreement. These results show that that the presented approach to automatic emphysema extent prediction is viable and could be useful for routine assessment of emphysema extent.

The four best performing classifiers, beta, misvm, milog and lmm, have very similar ICC coefficients, with lmm being slightly more consistent across replications. However, beta and misvm show superior overall prediction of extent intervals with a much better discrimination of CT scans without visible emphysema compared to milog and lmm. Overall stability of beta and misvm is also superior to milog and lmm, although milog shows more stable predictions for the lower prevalence extent intervals 6-25%, 26-50% and 51-75%. Learning from scores indicating emphysema extent did not appear to be advantageous for extent prediction compared to learning based on emphysema presence alone. The MIL classifiers, misvm and milog, and the LLP classifiers, beta and lmm, show comparable performance.

One possible explanation for the lack of improved performance when training on extent labels, is that the extent labels are too noisy, as the relatively large disagreement between observers suggests. Obtaining more accurate and precise extent labels is costly and it is not clear if it is possible to improve the label quality significantly. In this work we have combined the emphysema estimates of two raters by simple averaging of point estimates. In [12] we showed that performance of the cms classifier improved when learning from labels incorporating rater uncertainty over learning from averaged point estimates. The approach used in [12] is not directly applicable to the other methods used here and we have used point estimates to keep the comparison fair. Recent work on classification of retinal images with a CNN-based method [29] show that modeling individual raters can improve performance over simple averaging of multiple raters. Although more than 30 raters were used in [29] it is possible that a more complex model of rater annotations could also improve performance when only two raters are used.

Another possible factor is that the model of proportion labels is too simple to exploit the additional information in the labels. The results in [28] indicate that learning from proportion labels can help more complex models based on CNNs to converge faster and to a better optima than learning from binary labels. A possible explanation for this is that explicitly modeling proportion labels has a regularizing effect on the feature learning part of CNNs, which would also explain why we do not see improved performance for LLP methods when features are fixed.

We considered three strategies for learning from bag labels, the simple strategy, the relabeling strategy and the mean strategy. For the MIL classifiers it appears that the relabeling strategy is best, whereas for the LLP classifiers it appears that the simple and mean strategies are best. One reason the simple strategy works better for LLP than for MIL could be that a proportion instance label as used in simple LLP is interpreted as a probability of emphysema in the patch, whereas the binary instance labels as used in simple MIL are

	Bag							Instance	
	0%	1-5%	6-25%	26-50%	51-75%	76-100%	Overall	E	NE
log	89	60	68	58	57	–	79	35	97
svm	89	57	60	51	58	–	78	44	97
milog	84	60	**81**	**82**	**77**	–	78	44	96
misvm	95	70	77	70	74	–	88	52	**98**
beta	**96**	**71**	72	58	62	**50**	**89**	**56**	**98**
plog	65	54	73	78	–	–	61	07	95
psvm	24	40	47	47	60	0	41	12	82
cms	65	60	59	45	12	–	61	05	93
lmm	82	60	68	58	67	17	73	43	97
Rater	95	49	53	47	32	0	83	–	–

TABLE IV: Label stability. Agreement percentages between predictions from each replication. Instance columns are for binary instance predictions (**E**mphysema / **N**o **E**mphysema). A dash (–) indicates no predictions in that category.

interpreted as the probability the patch came from a CT scan with emphysema. In this sense, the proportion instance labels match the intended objective, predicting the proportion of patches with emphysema, much better than the binary instance labels.

A limitation of this study is that we have only trained and validated the classifiers on the upper right region of the lung. Due to the lower prevalence and agreement of visual scoring in the remaining five regions, we expect some decrease in extent prediction accuracy for these regions, similar to what was observed in [11] for regional emphysema detection. Investigating the performance over all regions should be considered in future work. However, the results in [11] show that a simple MIL classifier trained on subject-level presence/absence labels can provide the same performance as a classifier trained on region-level presence/absence labels. In light of the results here, this suggests that a MIL classifier, such as misvm, could provide accurate regional emphysema extent estimates even when trained only on subject-level presence of emphysema.

In conclusion, the best performing classifiers have close to human-level performance and are promising candidates for automatic quantification of emphysema extent. Furthermore, MIL classifiers having access to only emphysema presence labels perform just as well as LLP classifiers with access to emphysema extent labels. Reducing the labeling task from estimating emphysema extent to indicating presence, reduces the cost of training and makes it more feasible to implement in new settings.

V. References

References

[1] "Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2017," 2017. [Online]. Available: http://www.goldcopd.org/

[2] J. Sieren, J. Newell, P. Judy, D. Lynch, K. Chan, J. Guo, and E. Hoffman, "Reference standard and statistical model for intersite and temporal comparisons of CT attenuation in a multicenter quantitative lung study," *Medical physics*, vol. 39, no. 9, pp. 5757–5767, 2012.

[3] M. O. Wielpütz, D. Bardarova, O. Weinheimer, H.-U. Kauczor, M. Eichinger, B. J. Jobst, R. Eberhardt, M. Koenigkam-Santos, M. Puderbach, and C. P. Heussel, "Variation of densitometry on computed tomography in COPD–influence of different software tools," *PloS one*, vol. 9, no. 11, p. e112898, 2014.

[4] COPDGene CT Workshop Group:, R. G. Barr, E. A. Berkowitz, F. Bigazzi, F. Bode, J. Bon, R. P. Bowler, C. Chiles, J. D. Crapo, G. J. Criner, J. L. Curtis, C. Dass, A. Dirksen, M. T. Dransfield, G. Edula, L. Erikkson, A. Friedlander, M. Galperin-Aizenberg, W. B. Gefter, D. S. Gierada, P. A. Grenier, J. Goldin, M. K. Han, N. A. Hanania, N. N. Hansel, F. L. Jacobson, H.-U. Kauczor, V. L. Kinnula, D. A. Lipson, D. A. Lynch, W. MacNee, B. J. Make, A. J. Mamary, H. Mann, N. Marchetti, M. Mascalchi, G. McLennan, J. R. Murphy, D. Naidich, H. Nath, J. D. N. Jr., M. Pistolesi, E. A. Regan, J. J. Reilly, R. Sandhaus, J. D. Schroeder, F. Sciurba, S. Shaker, A. Sharafkhaneh, E. K. Silverman, R. M. Steiner, C. Strange, N. Sverzellati, J. H. Tashjian, E. J. van Beek, L. Washington, G. R. Washko, G. Westney, S. A. Wood, and P. G. Woodruff, "A combined pulmonary-radiology workshop for visual evaluation of COPD: Study design, chest CT findings and concordance with quantitative evaluation," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 9, no. 2, pp. 151–159, 2012.

[5] D. M. Hansell, A. A. Bankier, H. MacMahon, T. C. McLoud, N. L. Müller, and J. Remy, "Fleischner Society: Glossary of terms for thoracic imaging," *Radiology*, vol. 246, no. 3, pp. 697–722, 2008.

[6] M. M. W. Wille, L. H. Thomsen, J. Petersen, M. de Bruijne, A. Dirksen, J. H. Pedersen, and S. B. Shaker, "Visual assessment of early emphysema and interstitial abnormalities on CT is useful in lung cancer risk analysis," *European Radiology*, pp. 1–8, 2015.

[7] D. A. Lynch, J. H. Austin, J. C. Hogg, P. A. Grenier, H.-U. Kauczor, A. A. Bankier, R. G. Barr, T. V. Colby, J. R. Galvin, P. A. Gevenois *et al.*, "CT-definable subtypes of chronic obstructive pulmonary disease: a statement of the Fleischner Society," *Radiology*, vol. 277, no. 1, pp. 192–205, 2015.

[8] D. A. Lynch, C. M. Moore, C. Wilson, D. Nevrekar, T. Jennermann, S. M. Humphries, J. H. Austin, P. A. Grenier, H.-U. Kauczor, M. K. Han *et al.*, "CT-based visual classification of emphysema: Association with mortality in the COPDGene study," *Radiology*, p. 172294, 2018.

[9] M. M. W. Wille, L. H. Thomsen, A. Dirksen, J. Petersen, J. H. Pedersen, and S. B. Shaker, "Emphysema progression is visually detectable in low-dose CT in continuous but not in former smokers," *European Radiology*, vol. 24, no. 11, pp. 2692–2699, Nov 2014.

[10] R. Wiemker, M. Sevenster, H. MacMahon, F. Li, S. Dalal, A. Tahmasebi, and T. Klinder, "Automated assessment of imaging biomarkers for the PanCan lung cancer risk prediction model with validation on NLST data," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134. International Society for Optics and Photonics, 2017, p. 1013421.

[11] S. N. Ørting, J. Petersen, L. H. Thomsen, M. M. W. Wille, and M. de Bruijne, "Detecting emphysema using multiple instance learning," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018.

[12] S. N. Ørting, J. Petersen, M. M. Wille, L. H. Thomsen, and M. de Bruijne, "Quantifying emphysema extent from weakly labeled CT scans of the lungs using label proportions learning," *The Sixth International Workshop on Pulmonary Image Analysis*, pp. 31–42, 2016.

[13] P. J. Castaldi, R. S. J. Estépar, C. S. Mendoza, C. P. Hersh, N. Laird, J. D. Crapo, D. A. Lynch, E. K. Silverman, and G. R. Washko, "Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers," *American Journal of Respiratory and Critical Care Medicine*, vol. 188, no. 9, pp. 1083–1090, 2013.

[14] P. Binder, N. K. Batmanghelich, R. S. J. Estepar, and P. Golland, "Unsupervised discovery of emphysema subtypes in a large clinical cohort," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2016, pp. 180–187.

[15] J. Yang, E. D. Angelini, P. P. Balte, E. A. Hoffman, J. H. 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," in *International Conference on Medical Image*

Computing and Computer-Assisted Intervention. Springer, 2017, pp. 116–124.

[16] 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, Jan 2012.

[17] 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 *International Conference on Pattern Recognition*, 2014, pp. 1508–1513.

[18] Q. Wang, Y. Zheng, G. Yang, W. Jin, X. Chen, and Y. Yin, "Multi-scale rotation-invariant convolutional neural networks for lung texture classification," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, no. 99, pp. 1–1, 2017.

[19] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers, Z. Xu, and D. J. Mollura, "Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 1, pp. 1–6, 2018.

[20] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE reviews in biomedical engineering*, vol. 10, pp. 213–234, 2017.

[21] F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S. Chang, "$\propto$SVM for learning with label proportions," *Proccedings of International Conference on Machine Learning*, 2013.

[22] G. Patrini, R. Nock, T. Caetano, and P. Rivera, "(Almost) no label no cry," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 190–198.

[23] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *Journal of Applied Statistics*, vol. 31, no. 7, pp. 799–815, 2004.

[24] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2003, pp. 577–584.

[25] M. Stolpe and K. Morik, "Learning from label proportions by optimizing cluster model selection," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. Springer Berlin Heidelberg, 2011, vol. 6913, pp. 349–364.

[26] J. H. Pedersen, H. Ashraf, A. Dirksen, K. Bach, H. Hansen, P. Toennesen, H. Thorsen, J. Brodersen, B. G. Skov, M. Døssing, J. Mortensen, K. Richter, P. Clementsen, and N. Seersholm, "The Danish randomized lung cancer CT screening trial–overall design and results of the prevalence round," *Journal of Thoracic Oncology*, vol. 4, no. 5, 2009.

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

[28] G. Bortsova, F. Dubost, S. Ørting, I. Katramados, L. Hogeweg, L. Thomsen, M. Wille, and M. de Bruijne, "Deep learning from label proportions for emphysema quantification," *arXiv preprint arXiv:1807.08601*, 2018.

[29] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, "Who said what: Modeling individual labelers improves classification," *Proceedings of AAAI Conference*, 2018.

[30] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine learning*, vol. 68, no. 3, pp. 267–276, 2007.

APPENDIX A
EMPHYSEMA

Fig. 3 shows slices from the upper right region of three CT scans. Background and airways have been masked. The left image is assessed as having no visible emphysema extent. The center image as having 6-25% and the right image as having 51-75% emphysema extent. For the center image, emphysema is predominately visible at the boundary of the lung, whereas it is distributed throughout the region in the right image.

APPENDIX B
METHODS

A. Notation

Let $\mathcal{X}$ be an instance space, $\mathcal{Y}$ an instance label space, $\mathcal{Z}$ a bag label space and $\mathbf{b} = (\mathbf{x} \subseteq \mathcal{X}, z \in \mathcal{Z})$ a labeled bag of instances. We use superscripts to refer to the label ($\mathbf{b}^z$), instances ($\mathbf{b}^\mathbf{x}$) and instance labels ($\mathbf{b}^\mathbf{y}$) associated with a bag $\mathbf{b}$. For a set of m bags $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m\}$, $\mathbf{b}_i^\mathbf{x}$ are the instances in the i'th bag and $\mathbf{b}_{ij}^\mathbf{x}$ is the j'th instance in the i'th bag. For the set of all instances we use $\mathbf{X} = \cup_{i=1}^m \mathbf{b}_i^\mathbf{x}$, for all instance labels we use $\mathbf{Y} = \cup_{i=1}^m \mathbf{b}_i^\mathbf{y}$ and for all bag labels we use $\mathbf{Z} = \cup_{i=1}^m \{\mathbf{b}_i^z\}$.

B. mi-logistic

The bag learning problem for mi-logistic is a constrained optimization problem over model weights and unknown instance labels

$$\max_{\mathbf{w},\mathbf{Y}} \prod_{i,j} p(\mathbf{b}_{i,j}^\mathbf{y} \,|\, \mathbf{b}_{i,j}^\mathbf{x}, \mathbf{w}) \tag{14}$$

$$\text{s.t.} \forall i : \Theta_{\max}(\mathbf{b}_i^\mathbf{y}) = \mathbf{b}_i^z \in \{0, 1\}. \tag{15}$$

We use the heuristic for solving the mi-SVM problem from [24]. Initially, fix instance labels by setting them to bag labels, $\mathbf{b}_{i,j}^\mathbf{y} = \mathbf{b}_i^z \forall i, j$. For fixed instance labels (14) reduces to standard logistic regression. Let $h(\cdot) = \sigma(\mathbf{w}^T \cdot)$ denote the fitted model. Instance labels are predicted as

$$\tilde{\mathbf{b}}_{i,j}^\mathbf{y} = \mathbb{1}\{h(\mathbf{b}_{i,j}^\mathbf{x}) > 0.5\} \tag{16}$$

and bag labels are predicted as

$$\tilde{\mathbf{b}}_i^z = \Theta_{\max}(\tilde{\mathbf{b}}_i^\mathbf{y}). \tag{17}$$

Instance labels are then updated according to

$$\mathbf{b}_{i,j}^\mathbf{y} = \begin{cases} 0 & \text{if } \mathbf{b}_i^z = 0 \\ 1 & \text{if } \mathbf{b}_i^z = 1. \tilde{\mathbf{b}}_i^z = 0. h(\mathbf{b}_{i,j}^\mathbf{x}) > h(\mathbf{b}_{i,k}^\mathbf{x}) \forall k \neq j \\ \tilde{\mathbf{b}}_{i,j}^\mathbf{y} & \text{if otherwise} \end{cases} \tag{18}$$

The first clause ensures that instances from negative bags are always labeled negative. The second clause ensures that a positive bag predicted as negative will always have one positive instance by labeling the "most" positive instance as positive, and the third clause ensures all other instances in positive bags are relabeled to match the predicted class.

C. $\propto$-logistic

The $\propto$-logistic model can be derived by considering the joint probability over instances $\mathbf{X}$, bag labels $\mathbf{Z}$ and instance labels $\mathbf{Y}$

$$P(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) = P(\mathbf{Z}|\mathbf{Y}, \mathbf{X}) P(\mathbf{Y}, \mathbf{X}) \tag{19}$$

$$= P(\mathbf{Z}|\mathbf{Y}) P(\mathbf{Y}, \mathbf{X}) \qquad \mathbf{Z} \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{Y} \tag{20}$$

$$\propto P(\mathbf{Z}|\mathbf{Y}) P(\mathbf{Y}|\mathbf{X}) \qquad P(\mathbf{X}) = \text{Constant} \tag{21}$$

We use a logistic model for instance labels and a binomial model for bag labels.

$$P(\mathbf{Y} \,|\, \mathbf{X}) = \prod_{i,j} P(\mathbf{b}_{i,j}^\mathbf{y} | \mathbf{b}_{i,j}^\mathbf{x}, \mathbf{w}) \tag{22}$$

$$= \prod_{i,j} \sigma(\mathbf{w}^T \mathbf{b}_{i,j}^\mathbf{x})^{\mathbf{b}_{i,j}^\mathbf{y}} (1 - \sigma(\mathbf{w}^T \mathbf{b}_{i,j}^\mathbf{x}))^{1 - \mathbf{b}_{i,j}^\mathbf{y}} \tag{23}$$

$$P(\mathbf{Z} \,|\, \mathbf{Y}) = \prod_i P(\mathbf{b}_i^z | \mathbf{b}_i^\mathbf{y}) = \tag{24}$$

$$\prod_i \binom{|\mathbf{b}_i|}{|\mathbf{b}_i|\mathbf{b}_i^z} \Theta_{\text{mean}}(\mathbf{b}_i^\mathbf{y})^{|\mathbf{b}_i|\mathbf{b}_i^z} (1 - \Theta_{\text{mean}}(\mathbf{b}_i^\mathbf{y})^{|\mathbf{b}_i| - |\mathbf{b}_i|\mathbf{b}_i^z} \tag{25}$$

substituting into (21) gives us

$$P(\mathbf{Z}|\mathbf{Y}) P(\mathbf{Y}|\mathbf{X}) = \prod_i P(\mathbf{b}_i^z | \mathbf{b}_i^\mathbf{y}) \prod_j P(\mathbf{b}_{i,j}^\mathbf{y} | \mathbf{b}_{i,j}^\mathbf{x}, \mathbf{w}) \tag{26}$$

We want to find the $\mathbf{Y}$ and $\mathbf{w}$ that maximize (26)

$$\arg\max_{\mathbf{Y},\mathbf{w}} \prod_i P(\mathbf{b}_i^z | \mathbf{b}_i^\mathbf{y}) \prod_j P(\mathbf{b}_{i,j}^\mathbf{y} \,|\, \mathbf{b}_{i,j}^\mathbf{x}, \mathbf{w}) \tag{27}$$

We do this by fixing $\mathbf{Y}$ and $\mathbf{w}$ iteratively. For fixed $\mathbf{Y}$ we get standard logistic regression. For fixed $\mathbf{w}$ we can optimize over each bag individually

$$\arg\max_{\mathbf{b}_i^\mathbf{y}} P(\mathbf{b}_i^z \,|\, \mathbf{b}_i^\mathbf{y}) \prod_{j=1} P(\mathbf{b}_{i,j}^\mathbf{y} \,|\, \mathbf{b}_{i,j}^\mathbf{x}, \mathbf{w}). \tag{28}$$

This can be done with the same greedy method used for $\propto$-SVM in [21].

APPENDIX C
PARAMETERS

All classifiers provide probability estimates of instance labels and a classifier-specific instance threshold was fitted on the training data by trying all thresholds in the range $[0, 0.01, 0.02, \ldots, 0.99, 1]$. Fitted thresholds are reported in Table V. There is a large variation in fitted instance thresholds across classifiers, and for some classifiers there is a large variation across replications. Variation across replications is an indication that the classifier has learned substantially different decision rules for each replication. Variation between classifiers could just be a scaling issue, but is at least an indication that interpreting instance predictions as probability estimates is problematic.

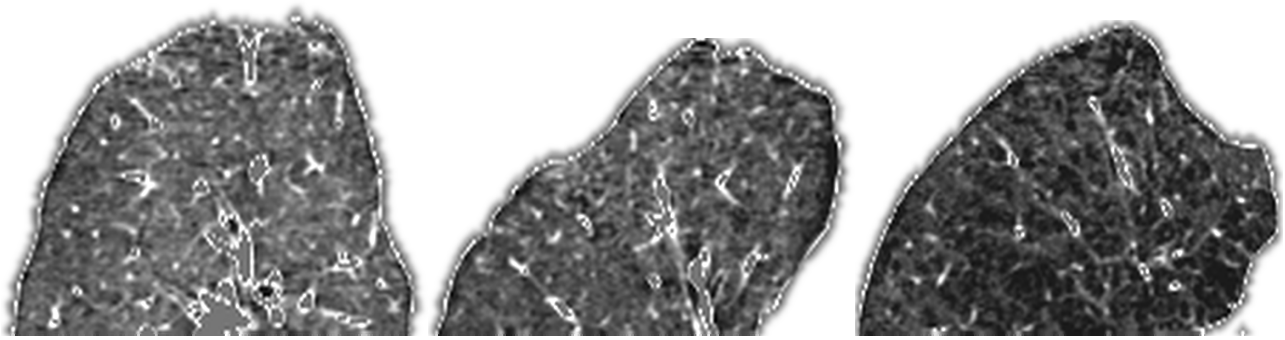

Fig. 3: Example slices. From left, visually assessed emphysema extent is 0%, 6-25% and 51-75%. Window level -780HU, window width 560HU.

Classifier	D1	D2	D3
log	0.78	0.79	0.60
beta	0.09	0.09	0.09
svm	0.77	0.76	0.70
misvm	0.85	0.94	0.78
milog	0.99	0.99	0.99
psvm	0.97	0.99	0.14
plog	0.01	0.01	0.01
cms	0.86	0.68	0.99
lmm	0.75	0.62	0.73

TABLE V: Fitted instance thresholds for each classifier and replication

A. *beta*

The implementation of beta regression requires uncorrelated features and we used the PCA algorithm to decorrelate features. We tried dimensionality reduction (only keep principal components with standard deviation ≥ 1). We tried two optimization methods, maximum likelihood estimation (ML) and bias correction (BC).

Fitted parameters

beta	D1	no dimensionality reduction, ML
	D2	no dimensionality reduction, BC
	D3	no dimensionality reduction, ML

B. *svm, misvm, psvm*

For all three classifiers we tried both linear and RBF kernels. In both cases we tried $C \in \{0.1, 1, 10, 100\}$. For psvm we tried $C_2 \in \{1, 10, 100, 1000\}$. For the RBF kernels we tried $\gamma \in \{0.1, 1\}$. We used Platt calibration [30] to obtain probability estimates from all three SVMs.

Fitted parameters

svm	D1	linear kernel, $C = 1$
	D2	linear kernel, $C = 0.1$
	D3	linear kernel, $C = 10$
misvm	D1	linear kernel, $C = 0.1$
	D2	linear kernel, $C = 0.1$
	D3	linear kernel, $C = 0.1$
psvm	D1	rbf kernel, $C = 1, C_2 = 1, \gamma = 0.1$
	D2	rbf kernel, $C = 0.1, C_2 = 1, \gamma = 1$
	D3	rbf kernel, $C = 1, C_2 = 1, \gamma = 0.1$

C. *log, milog, plog*

We tried dimensionality reduction using PCA for log. We did not use dimensionality reduction for milog and plog. We ran milog and plog until convergence of instance labels or for 20 iterations, whichever came first.

Fitted parameters

log	D1	no dimensionality reduction
	D2	no dimensionality reduction
	D3	dimensionality reduction

D. *cms*

We used the following fixed parameters, branching = 2, number of k-means iterations = 25, maximum iterations of CMA-ES = 1000, $\lambda = 13$. We tried number of clusters $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$.

Fitted parameters

cms	D1	$k = 70$
	D2	$k = 50$
	D3	$k = 100$

E. *lmm*

We tried

$$\lambda \in \{0, 1, 10, 100\}$$
$$\gamma \in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 1\}$$
$$\sigma \in \{0.001, 0.01, 0.1, 0.125, 0.25, 0.5, 1.0\}$$

Fitted parameters

lmm	D1	$\lambda = 1, \gamma = 0.01, \sigma = 0.1$
	D2	$\lambda = 1, \gamma = 0.01, \sigma = 0.1$
	D3	$\lambda = 10, \gamma = 0.00001, \sigma = 0.25$

4 Crowdsourcing for medical imaging

This section is based on two manuscripts. The first [32] investigates how emphysema assessment can be framed as a task that can be solved by non-expert crowd workers. The second [29] provides a review of crowdsourcing in medical image analysis.

[32] Silas Nyboe Ørting, Veronika Cheplygina, Jens Petersen, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. *Crowdsourced emphysema assessment*. In Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. 2017.

[29] Silas Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R. Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina. *A survey of crowdsourcing in medical image analysis*. arXiv e-prints, art. arXiv:1902.09159. Feb 2019.

Crowdsourced Emphysema Assessment

Silas Nyboe Ørting[1( )], Veronika Cheplygina[2,5], Jens Petersen[1], Laura H. Thomsen[3], Mathilde M.W. Wille[4], and Marleen de Bruijne[1,5]

[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
silas@di.ku.dk
[2] Medical Image Analysis (IMAG/e), Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[3] Department of Respiratory Medicine, Gentofte Hospital, Hellerup, Denmark
[4] Department of Diagnostic Imaging, Bispebjerg Hospital, Copenhagen, Denmark
[5] Biomedical Imaging Group Rotterdam, Departments of Radiology and Medical Informatics, Erasmus MC - University Medical Center Rotterdam, Rotterdam, The Netherlands

Abstract. Classification of emphysema patterns is believed to be useful for improved diagnosis and prognosis of chronic obstructive pulmonary disease. Emphysema patterns can be assessed visually on lung CT scans. Visual assessment is a complex and time-consuming task performed by experts, making it unsuitable for obtaining large amounts of labeled data. We investigate if visual assessment of emphysema can be framed as an image similarity task that does not require expert. Substituting untrained annotators for experts makes it possible to label data sets much faster and at a lower cost. We use crowd annotators to gather similarity triplets and use t-distributed stochastic triplet embedding to learn an embedding. The quality of the embedding is evaluated by predicting expert assessed emphysema patterns. We find that although performance varies due to low quality triplets and randomness in the embedding, we still achieve a median F_1 score of 0.58 for prediction of four patterns.

Keywords: Crowdsourcing · Emphysema · Similarity learning

1 Introduction

Emphysema is a lung pathology common to chronic obstructive pulmonary disease that is a major cause of morbidity and mortality world wide [3]. Emphysema is characterized by destruction of lung tissue. Lung CT scans can reveal emphysema and visual scoring can be used to rate the extent and type of emphysema in the lungs [14]. Visual scores can be used for training classifiers to automatically assess presence and extent of emphysema [9,11]. However, visual scoring of emphysema by experts is both expensive and prone to high rater disagreement [14]. Instead of performing a full visual scoring, which requires expert knowledge

M.J. Cardoso et al. (Eds.): CVII-STENT/LABELS 2017, LNCS 10552, pp. 126–135, 2017.
DOI: 10.1007/978-3-319-67534-3_14

of the lungs, we investigate whether it is possible to reduce emphysema assessment to a simpler task that can be performed by untrained raters, or crowds.

In fields such as computer vision, crowdsourcing - outsourcing simple tasks to a crowd of online users, often without any specific training - has been used successfully to gather labels for training and validation of classifiers [4]. Most of this research focuses on collecting labels that directly characterize the content of the image, for instance presence of an object or indicating regions of interest. Motivated by the fact that some categorization tasks may be difficult for non-experts, a few others instead focus on collecting assessments of similarities between images. For example, Wah et al. [13] collect similarities between images of different bird species, which most people do not know by name, but can easily assess their visual similarity. The similarities can then be used to learn an embedding that can aid classification.

Due to the success of crowdsourcing in computer vision, there have also been several efforts to apply it to medical imaging [1,2,6,8]. Similar to methods from the computer vision field, these works focus on collecting labels for images, targeting classification or segmentation tasks. For example, the crowd can be asked to grade retinal images as normal or abnormal [8] or to segment airways in 2D slices of chest CT images [2]. To the best of our knowledge, this work is the first to gather crowdsourced similarities for medical images, as well as to apply a crowdsourcing approach to classification of emphysema patterns.

2 Materials and Methods

2.1 Data

We used 40 chest CT scans from the a national lung cancer screening trial [10] and visual assessment of emphysema from [14]. Visual assessment is performed by considering the full 3D volume and splitting each lung in three regions. The top, middle and lower regions are defined as above carina, between carina and inferior pulmonary vein, and below inferior pulmonary vein. The volume is assigned a label indicating the predominant emphysema pattern and each region is assigned an estimate of the extent of emphysema in the region. The 40 scans were selected amongst those where raters agreed on visual assessment of both predominant pattern and emphysema extent in the upper right region. We excluded scans with panlobular emphysema due to low prevalence. We grouped candidate scans based on predominant pattern: normal (N), centrilobular (C), paraseptal (P), mixed (M), and chose ten scans from each group. For the three emphysema groups (C, P, M) we chose the scans with highest extent, and for the normal group we chose ten scans at random. We used lung fields segmented from the scans obtained from [5].

We extracted nine coronal slices from the top region of the right lung of each scan. The slices were evenly spaced (10 mm) and located such that the center slice coincided with the center slice of the region. In this way we covered a depth of 80 mm and avoided slices at the very boundary of the lungs. An example of an extracted set of slices is given in Fig. 1. The slices are extracted from

a subject with a large extent of centrilobular emphysema. We see that while texture patterns vary a lot throughout the region, patterns are similar between neighboring slices. It is also clear that size and shape of the lung region varies with slice location. To avoid having workers focus on the differences in lung size and shape, we stratify slices by their location in the lung when sampling triplets.

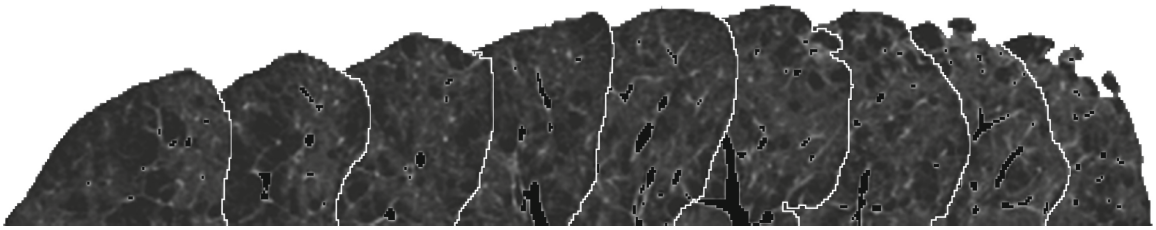

Fig. 1. Nine slices extracted from a single volume. There is a large extent of centrilobular emphysema. We can see that neighboring slices tend to have more similar texture patterns than slices that are far away from each other. White border added for clarity.

2.2 Crowdsourced Triplets

We used Amazon MTurk[1] to collect similarity triplets. MTurk centers on the concept of a human intelligence task (HIT), a self-contained task that can be solved by a worker. We designed our HIT as a set of three image triplets where the task is to provide similarity assessment of each of the three triplets. A screenshot showing part of a HIT is given in Fig. 2. We asked workers to choose one of two images on the right with the most similar disease patterns to the image on the left. We instructed workers to look for emphysema patterns, defined as areas of low intensity, and consider the distribution of patterns of these areas: scattered throughout the lung or concentrated. We emphasized that workers should ignore differences in size and shape of the lung. We asked three different workers to perform each HIT. We required workers that had at least 1000 previously approved HITs and a 95% approval rate. The reward for each task was $0.10.

We collected 9720 similarity triplets for 3240 unique image triplets. 150 different workers worked on the HITs, with a median number of HITs per worker of 6.5 (19.5 similarity triplets). The median work time per HIT was 55 s. The most productive worker submitted 131 HITs and the lowest work time for a HIT was 4 s. More than 92% of the HITs were finished within 30 min of the first HIT being available. The total cost was $388.80.

2.3 Similarity Embedding

We used t-distributed stochastic triplet embedding (t-STE) [12] to learn an n-dimensional Euclidean embedding from the similarity triplets. t-STE searches for an embedding X that maximizes the probability of observing the given triplets.

[1] https://www.mturk.com.

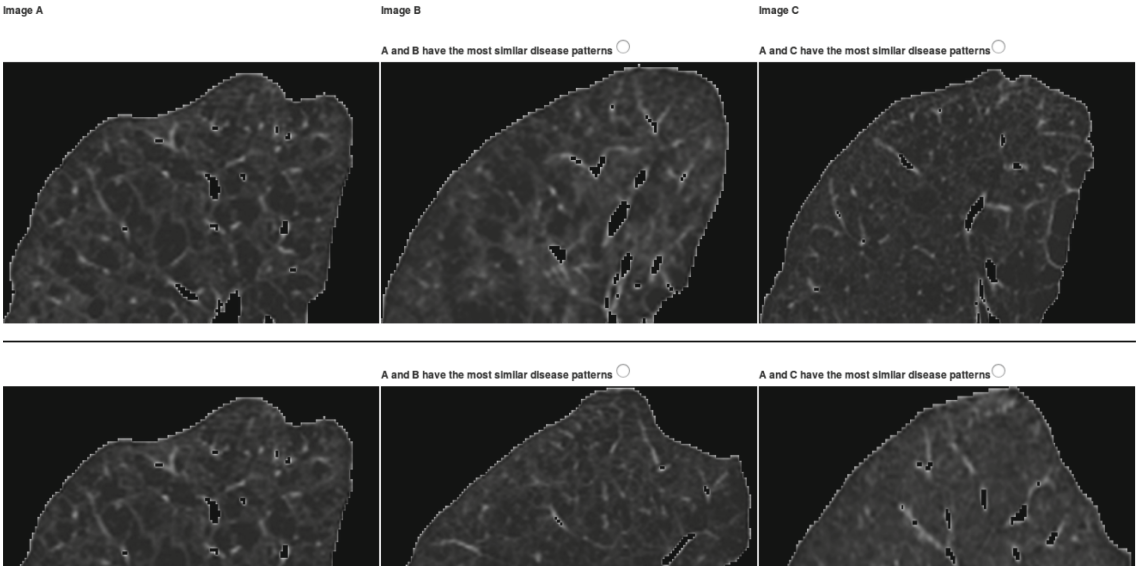

Fig. 2. Amazon MTurk user interface for collecting the similarity triplets

Let T be the set of known triplets and $ijl \in T$ a triplet indicating that $d(i,j) < d(i,l)$. The probability of ijl given $x_i, x_j, x_l \in X$ is

$$p_{ijl} = \frac{\left(1 + \frac{||x_i - x_j||_2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1 + \frac{||x_i - x_j||_2}{\alpha}\right)^{-\frac{\alpha+1}{2}} + \left(1 + \frac{||x_i - x_l||_2}{\alpha}\right)^{-\frac{\alpha+1}{2}}} \tag{1}$$

The optimization problem is

$$\min_X - \sum_{ijl \in T} \log p_{ijl} \tag{2}$$

which is solved with gradient descent using the implementation from Michael Wilber[2].

Crowdsourced similiarity triplets are very likely to contain inconsistent and redundant triplets. When multiple workers perform the same HIT this is definitely the case. McFee and Lanckriet [7] give empirical evidence that pruning triplets for consistency and redundancy reduces computation time without affecting performance. However, they compare against a baseline where directly disagreeing triplets are removed. Removing triplets where workers disagree removes information about the uncertainty of the triplets. We can implicitly model this uncertainty by keeping all triplets. It can be shown that for $x = x_i, x_j, x_l$ the conflicting triplets satisfy

$$\frac{\partial}{\partial x} p_{ijl} = -\frac{\partial}{\partial x} p_{ilj}, \tag{3}$$

and the sum of the derivatives becomes

$$\frac{\partial}{\partial x} \log p_{ijl} + \frac{\partial}{\partial x} \log p_{ilj} = \frac{\partial}{\partial x} p_{ijl} \left(\frac{1}{p_{ijl}} - \frac{1}{p_{ilj}} \right) \tag{4}$$

[2] https://github.com/gcr/cython_tste.

which will drive the triplets to become equally probable, i.e. $||x_i - x_j|| = ||x_i - x_l||$. In the case where ijl occur c_j times and ilj occur c_l the gradient will depend on both the ratio c_j/c_l and the distances $||x_i - x_j||, ||x_i - x_l||$. In this way workers uncertainty about triplets will be accounted for in the optimization.

We used k-fold cross-validation with a multinomial log-linear model to estimate the predictive performance of the obtained embeddings. We enforced that each test fold contained exactly one sample from each class. For four classes with ten scans each this resulted in 10-fold cross-validation. We used the predominant pattern from the expert visual scoring of the regions as class labels. The model was fitted as a neural network with one hidden layer using the multinom function from the `nnet` package[3].

3 Experiments and Results

3.1 Simulated Similarity Triplets

To estimate how many triplets are needed to reveal an underlying pattern we performed a simulation experiment. We defined a distance function that encodes a similarity hierarchy of visually assessed patterns and emphysema extent. Paraseptal emphysema often appear as a small number of large holes, whereas centrilobular emphysema often appear as a large number of small holes. We therefore expect most raters will consider normal and centrilobular patterns more similar than normal and paraseptal patterns. We also expect both centrilobular and paraseptal patterns to be considered more similar to the mixed pattern than to each other. For images with the same pattern class we used absolute distance on emphysema extent. This simple distance function does not account for variability in patterns and it is unlikely that image based similarity triplets will match the visual assessment perfectly. However, it does provide some insight into the amount of triplets necessary. We used three sets of randomly selected triplets with sizes of 120, 240, and 360. For each set of triplets we generated 100 2D embeddings and estimated the prediction performance of the embedding with the multinomial model described above. We used the F_1 score to measure performance

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \,. \tag{5}$$

The median F_1 score for 120 triplets was 0.8 and improved to 0.9 for 240 triplets and to 1.0 for 360 triplets. There was some variation in performance for 120 and 240 triplets, whereas almost all 360 triplet embeddings gave perfect prediction. Representative embeddings for 120 and 240 triplets are given in Fig. 3. We can see that the embedding matches the distance function quite well, with normal and paraseptal being furthest from each other and mixed in between centrilobular and paraseptal. We also see some class overlap for 120 triplets and almost no overlap for 240 triplets. We used these results to guide the crowdsourcing to gather relatively many triplets for a small number of scans.

[3] https://cran.r-project.org/web/packages/nnet.

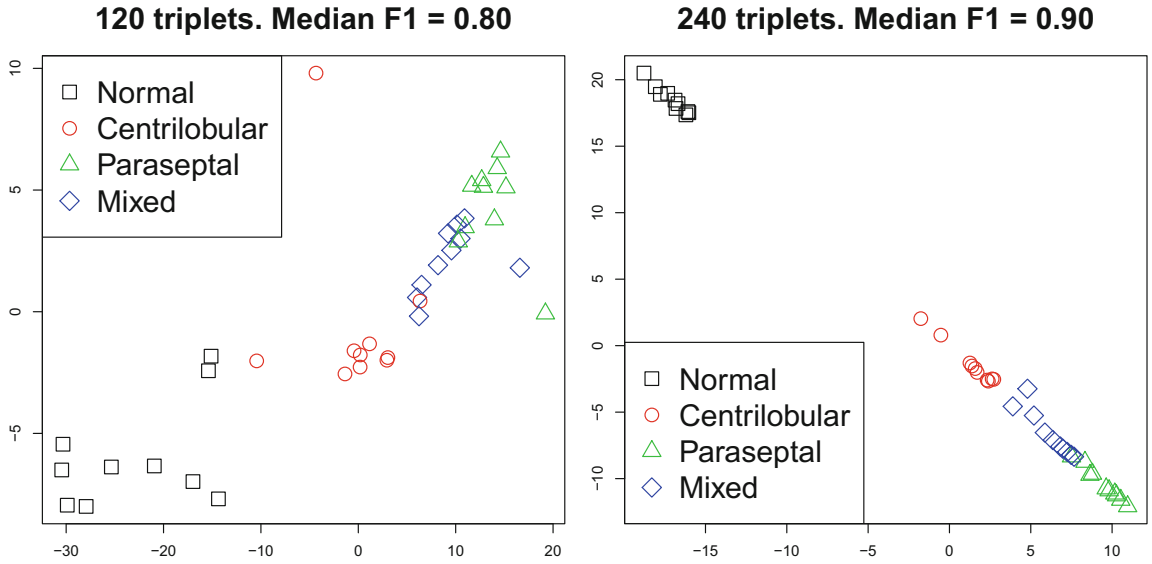

Fig. 3. Example embeddings from simulated triplets. Left: 120 triplets. Right: 240 triplets. While there is no overlap between emphysema and normal classes in both cases, there is some overlap between emphysema classes for 120 triplets.

3.2 Crowdsourced Similarity Triplets

We estimated the quality of the crowdsourced triplets by measuring the agreement with a small set of validation triplets. The validation triplets were labeled by one of the authors and consist of 52 triplets that the authors view as easy to reproduce. The overall agreement was 71% with a large variation between workers. We expected most workers to work on one or more validation triplets. However, due to the large number of workers only 41% of workers worked on a validation triplet and only 11% on more than two validation triplets. While agreement was lower than anticipated, and some workers had very poor agreement, we decided to include all triplets.

We varied the embedding dimensionality d from $1-10$. We set $\alpha = \max(d-1, 1)$ for all experiments and used a random initialization of t-STE. From the similarity triplets we learned an embedding of slices. Due to the stratification of triplets by slice location it is not meaningful to embed different slice locations simultaneously. We therefore concatenated the slice feature vectors to obtain a region embedding. We normalized each slice embedding to avoid that slice locations with numerically large distances dominated the region embedding. As an alternative to embedding each slice location separately we added triplets between slice locations and embedded all slice locations simultaneously. The extra triplets were derived by exploiting that neighboring slices in a region, in general, are more similar than slices further away from each other. This "neighbor similarity" was encoded with the distance function

$$d(slice_i, slice_{i+1}) < d(slice_i, slice_{i+3}),\ i \in [1:6],$$
$$d(slice_i, slice_{i-1}) < d(slice_i, slice_{i-3}),\ i \in [4:9],$$

and the corresponding triplets were added to T. We refer to the first approach as stratified and the second as combined. All embeddings were repeated 100 times to account for variability arising from the random initialization of t-STE.

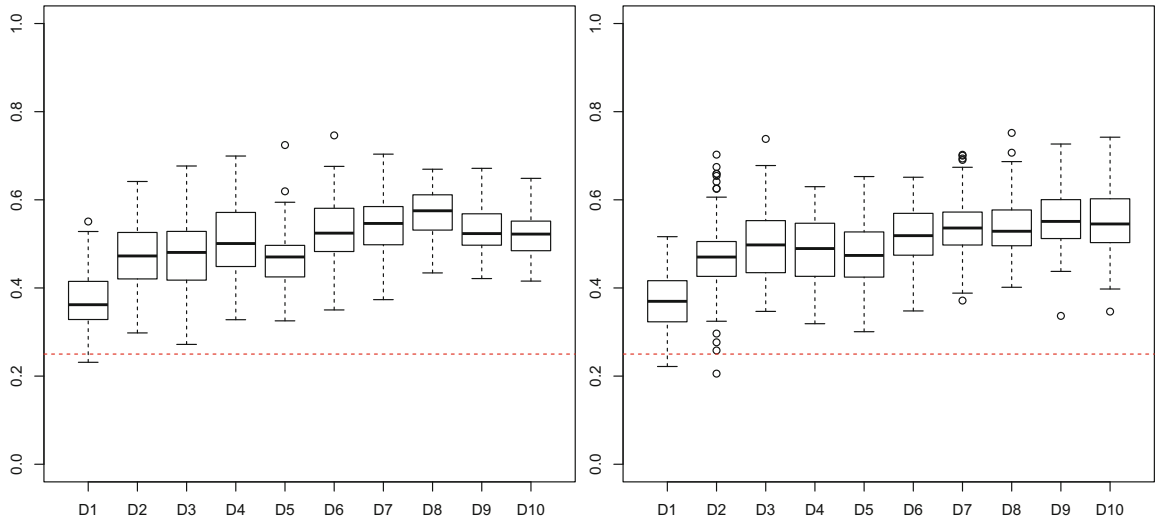

Fig. 4. Distribution of mean F_1 scores for classification of emphysema type. Left stratified, right combined. The dashed red line indicate random performance ($F_1 = 0.25$). (Color figure online)

Figure 4 shows the mean F_1 score over all classes for increasing embedding dimension for stratified and combined embeddings. Best median performance was achieved with D8 for stratified ($F_1 = 0.58$) and with D9 for combined ($F_1 = 0.55$). In both plots we see a large variation in performance. Adding the extra triplets for combined embedding seems to make performance more similar across dimensions, but does not decrease variation within each dimension. The direct source of the variation is the random initialization of t-STE. However, as the simulation showed, having a large consistent set of triplets will drive the variation in prediction performance to 0. The extra triplets for combined, that as subset is consistent, did not reduce variation, so the main underlying cause is likely having too many inconsistent triplets.

Figure 5 show performance by class. In all cases we see best performance on centrilobular and normal. For $D > 5$ we see consistently higher performance on centrilobular than on normal. Performance on centrilobular seems to be the main cause for the higher mean scores at D8 and D9. Treating mixed and paraseptal as one pattern makes the performance similar to performance on centrilobular (results not shown). This indicates that the main difficulty is in distinguishing paraseptal and mixed.

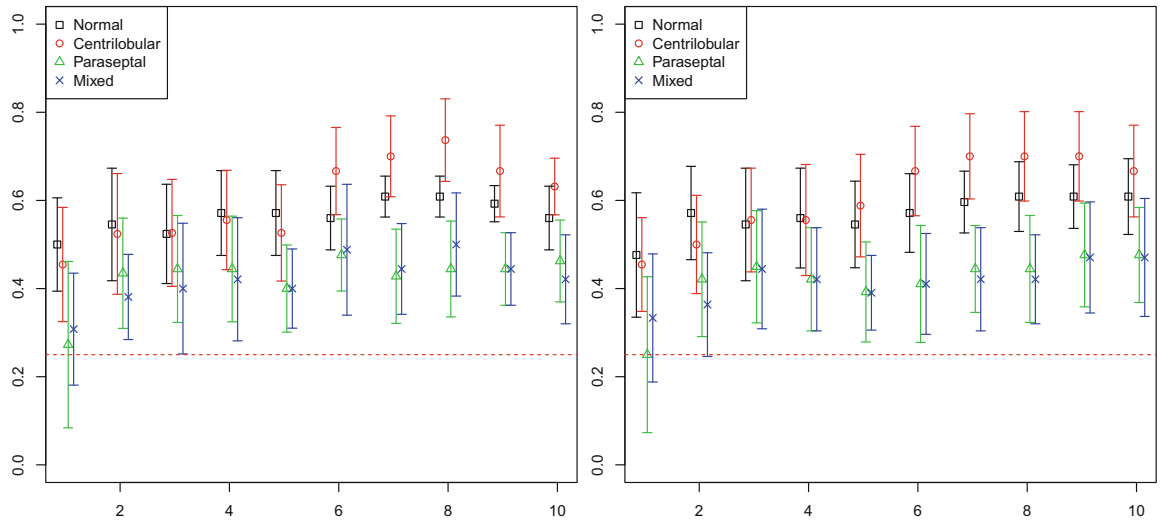

Fig. 5. F_1 scores for classification of emphysema type. Left for stratified, right for combined. The dashed red line indicate random performance ($F_1 = 0.25$). Symbols indicate median values and bars indicate ± 1 median absolute deviation. (Color figure online)

4 Discussion and Conclusion

Although there was large variation in prediction performance, it was in all but a few cases substantially better than random. The results from the simulation experiment show that more triplets improve median prediction performance and reduce variance. However, the simulation experiment uses triplets that perfectly encodes a distance function on patterns. While more crowdsourced triplets might improve performance and reduce variance, it is possible that higher quality set of triplets is needed to see significant gains.

Pruning triplets could improve quality. Directly inconsistent triplets, i.e. $ijl, ilj \in T$, can arise from poorly performing workers or difficult decisions. If we assume they represent difficult decisions, then they contain important information that we would like to keep. Pruning triplets is shown by [7] to be NP-hard and can only be solved approximately. Using the information from the direct inconsistencies to guide the pruning could be an interesting approach to improve the quality of the triplet set.

Direct inconsistencies due to poorly performing workers should not guide anything, but be removed. One approach is to rank workers and discard triplets from the least trustworthy workers. Ranking could be done by ensuring all workers perform tasks with a reference. Alternatively, it could be based on how well each worker agree with other workers. The first case requires expert labels and that each worker perform a minimum number of reference tasks. The second case requires that workers perform a large number of tasks and that tasks overlap with many different workers. In the future we intend to use one or both approaches to improve the quality of the triplet set.

An alternative to filtering triplets from poorly performing workers is to only enlist high performing workers. This could be done by splitting the tasks into many small sets and only allow the best performing workers to work on a new set. In this way the workforce would be trained to solve the tasks to our specification. Another option is recruiting workers that find the tasks worth doing beyond the financial gain. One worker expressed interest in working more on this type of tasks and asked "*Am I qualified to be a pulmonologist now?*". Compared to many other crowdsourcing tasks, medical image analysis seems like a good fit for community research, where people outside the traditional research community play an active part. It requires a larger degree of openness and communication about the research process but could be a tool to recruit high quality workers.

In this work we aimed at keeping HITs as simple as possible, hence the choice of collecting triplets. Instead of similarity triplets it is possible to ask workers to label the images. We believe that asking untrained workers to assess emphysema pattern and extent would be overly optimistic. However, focusing on a few simple questions might work well, for example "Are there dark holes in the lung?", "Are holes present in more than a third of the lung?", "Are the holes predominantly at the boundary of the lung?". These types of questions correspond to a model we have of emphysema and could be used to derive emphysema pattern and extent labels. The downside is that we need to know exactly what we want answered at the risk of missing important unknowns in the data.

Regardless of the high variance in performance, we conclude that untrained crowd workers can perform emphysema assessment when it is framed as a question of image similarity. No quality assurance, beyond requiring that workers had experience with MTurk, was performed. It is likely that large improvements can be gained by quality assurance of similarity triplets.

Acknowledgments. We would like to thank family, friends and coworkers at the University of Copenhagen, Erasmus MC - University Medical Center Rotterdam, Eindhoven University of Technology, and the start-up understand.ai for their help in testing prototype versions of the crowdsourcing tasks. This study was financially supported by the Danish Council for Independent Research (DFF) and the Netherlands Organization for Scientific Research (NWO).

References

1. Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N.: Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE TMI **35**(5), 1313–1321 (2016)
2. Cheplygina, V., Perez-Rovira, A., Kuo, W., Tiddens, H.A.W.M., de Bruijne, M.: Early experiences with crowdsourcing airway annotations in chest CT. In: Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J.M.R.S., Belagiannis, V., Papa, J.P., Nascimento, J.C., Loog, M., Lu, Z., Cardoso, J.S., Cornebise, J. (eds.) LABELS/DLMIA 2016. LNCS, vol. 10008, pp. 209–218. Springer, Cham (2016). doi:10.1007/978-3-319-46976-8_22
3. From the Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) (2015)

4. Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K.: Crowdsourcing in computer vision. Found. Trends. Comput. Graph. Vis. **10**(3), 177–243 (2016)
5. Lo, P., Sporring, J., Ashraf, H., Pedersen, J.J.H.: Vessel-guided airway tree segmentation: a voxel classification approach. Med. Image Anal. **14**(4), 527–538 (2010)
6. Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H.G., Eisenmann, M., Speidel, S.: Can masses of non-experts train highly accurate image classifiers? In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 438–445. Springer, Cham (2014). doi:10.1007/978-3-319-10470-6_55
7. McFee, B., Lanckriet, G.: Learning multi-modal similarity. J. Mach. Learn. Res. **12**, 491–523 (2011)
8. Mitry, D., Zutis, K., Dhillon, B., Peto, T., Hayat, S., Khaw, K.-T., Morgan, J.E., Moncur, W., Trucco, E., Foster, P.J.: The accuracy and reliability of crowdsource annotations of digital retinal images. Trans. Vis. Sci. Technol. **5**(5), 6–6 (2016)
9. Nishio, M., Nakane, K., Kubo, T., Yakami, M., Emoto, Y., Nishio, M., Togashi, K.: Automated prediction of emphysema visual score using homology-based quantification of low-attenuation lung region. PLOS ONE **12**(5), 1–12 (2017)
10. Pedersen, J.H., Ashraf, H., Dirksen, A., Bach, K., Hansen, H., Toennesen, P., Thorsen, H., Brodersen, J., Skov, B.G., Døssing, M., Mortensen, J., Richter, K., Clementsen, P., Seersholm, N.: The Danish randomized lung cancer CT screening trial-overall design and results of the prevalence round. J. Thorac. Oncol. **4**(5), 608–614 (2009)
11. Nyboe Ørting, S., Petersen, J., Wille, M., Thomsen, L., de Bruijne, M.: Quantifying emphysema extent from weakly labeled CT scans of the lungs using label proportions learning. In: MICCAI PIA, pp. 31–42. CreateSpace Independent Publishing Platform (2016)
12. van der Maaten, L., Weinberger, K.: Stochastic triplet embedding. In: IEEE MLSP, pp. 1–6 (2012)
13. Wah, C., Van Horn, G., Branson, S., Maji, S., Perona, P., Belongie, S.: Similarity comparisons for interactive fine-grained categorization. In: IEEE CVPR, pp. 859–866 (2014)
14. Wille, M.M., Thomsen, L.H., Dirksen, A., Petersen, J., Pedersen, J.H., Shaker, S.B.: Emphysema progression is visually detectable in low-dose CT in continuous but not in former smokers. Eur. Radiol. **24**(11), 2692–2699 (2014)

A Survey of Crowdsourcing in Medical Image Analysis

Silas Ørting[1]✉, Andrew Doyle[2,*], Matthias Hirth[3,*] Arno van Hilten[4,*], Oana Inel[5,7,*],
Christopher R. Madan[6,*], Panagiotis Mavridis[7,*], Helen Spiers[8,9,*], and Veronika Cheplygina[10]✉
[1] University of Copenhagen, Copenhagen, Denmark
[2] McGill Centre for Integrative Neuroscience, Montreal, Canada
[3] Technische Universität Ilmenau, Ilmenau, Germany
[4] Erasmus Medical Center, Rotterdam, The Netherlands
[5] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[6] University of Nottingham, Nottingham, United Kingdom
[7] Delft University of Technology, Delft, The Netherlands
[8] University of Oxford, Oxford, United Kingdom
[9] Zooniverse, University of Oxford, Oxford
[10] Eindhoven University of Technology, Eindhoven, The Netherlands
✉ *silas@di.ku.dk, v.cheplygina@tue.nl*
* *These authors contributed equally and are listed alphabetically by last name*

***Abstract*—Rapid advances in image processing capabilities have been seen across many domains, fostered by the application of machine learning algorithms to "big-data". However, within the realm of medical image analysis, advances have been curtailed, in part, due to the limited availability of large-scale, well-annotated datasets. One of the main reasons for this is the high cost often associated with producing large amounts of high-quality meta-data. Recently, there has been growing interest in the application of crowdsourcing for this purpose; a technique that has proven effective for creating large-scale datasets across a range of disciplines, from computer vision to astrophysics. Despite the growing popularity of this approach, there has not yet been a comprehensive literature review to provide guidance to researchers considering using crowdsourcing methodologies in their own medical imaging analysis. In this survey, we review studies applying crowdsourcing to the analysis of medical images, published prior to July 2018. We identify common approaches, challenges and considerations, providing guidance of utility to researchers adopting this approach. Finally, we discuss future opportunities for development within this emerging domain.**

***Index Terms*—Medical imaging, crowdsourcing, citizen science, machine learning**

I. Introduction

The limited availability and size of labeled datasets for training machine learning algorithms is a common problem in medical image analysis [Greenspan et al., 2016, Litjens et al., 2017, Cheplygina et al., 2018]. In several other fields, crowdsourcing - defined as the outsourcing of tasks to a crowd of individuals [Howe, 2006]- has been found effective for labeling large quantities of data. For example, in computer vision crowdsourcing has been used to annotate large datasets of images and videos with various tags [Kovashka et al., 2016].

Due to the success of crowdsourcing, several researchers have recently applied these techniques to the annotation of medical images. Although such images present specific challenges, including absence of expertise of the crowd, several early papers such as [Mitry et al., 2013, Mavandadi et al., 2012, Maier-Hein et al., 2014a] have demonstrated promising results. Despite the growing interest, there has not been an overview of the work in this field. In this paper we summarize existing literature on crowdsourcing in medical imaging.

This paper originated during the Lorentz workshop "Crowdsourcing in medical image analysis" in June 2018[1]. As participants of the workshop, we searched Google Scholar with the query "crowdsourcing AND (medical or biomedical)" and screened the results for papers focusing on the topic. Google Scholar was selected due to previous papers highlighting the poor indexing of the topic in databases and the high prevalence of crowdsourcing papers in conferences [Wazny, 2017]. Additional papers were identified for inclusion by examining the references and citations of selected papers. We only included papers where the crowd was involved in the analysis of medical or biomedical images, for example by annotating them. Our search strategy resulted in 55 papers. Key terms emerging from these studies are illustrated in Fig. 1. Five key dimensions were identified for discussion; the application involved, the type of interaction between the crowd and the images, the scale of the task (such as the number of images), the type of evaluation performed on the crowd annotations, and the results of the evaluation.

There are a number of surveys which are related to this work. However, they are quite different in scope:

- Ranard et al. [2014] survey crowdsourcing in health

[1] https://www.lorentzcenter.nl/lc/web/2018/967/info.php3?wsid=967&venue=Snellius

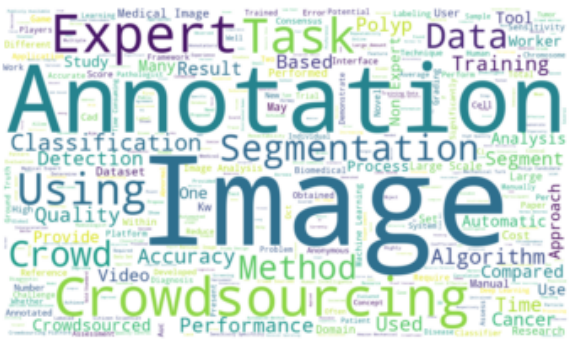

Fig. 1. A word cloud of the abstracts of the surveyed papers.

and medical research. They identify four tasks: problem solving, data processing, monitoring and surveying and cover 21 papers published until March 2013. In contrast, we only focus on papers where image analysis (i.e. data processing) is involved.

- Kovashka et al. [2016] survey crowdsourcing in computer vision. The surveyed papers focus on analysis of everyday/natural images. Only one of the 195 referenced papers (Gurari et al. [2015]) uses biomedical data.
- Wazny [2017] present a meta-review of crowdsourcing from 2006 to 2016. Similar to Ranard et al. [2014], they take a more broad view of crowdsourcing. They review existing review papers until August 2015, focusing on how each review categorizes the papers, for example by platform, size of crowd, and so forth.
- Alialy et al. [2018] is most similar to our survey, but only focuses on crowdsourcing in human pathology. They do a systematic search with several steps, excluding conference papers or abstracts, and summarize seven papers. The coverage of literature is therefore much more limited than in this work.

While this paper is a preprint, we welcome feedback from other researchers, which we will aim to incorporate in the journal version. Interested researchers can submit comments via https://goo.gl/forms/Qzr2yAJQjOnRCAF23.

The paper is organized as follows. The five dimensions according to which we analyzed the papers are described in more detail in Sections II to VI. We then discuss overall trends, limitations, and opportunities for future research in Section VII.

II. Applications

There are a variety of crowdsourcing applications to medical imaging data addressed in the papers surveyed in this work. We group these applications by (1) the type of task performed by the crowd, (2) the biomedical content of the image and (3) the dimensionality of the images.

A. Type of task

An important task in medical image analysis is classification, and 42% of the surveyed papers focus on this task. Classification can refer to assigning a label to an entire image, such as diagnosing whether a chest CT image contains any abnormalities. Classification can also refer to assigning a label to a part of the image, for example, the type of abnormality located in a particular region of interest. Other types of labels include non-diagnostic labels such as image modality [de Herrera et al., 2014], visual attributes [Cheplygina and Pluim, 2018], and assessing the quality of the image [Keshavan et al., 2018]. These three types of labels are based more on visual characteristics, and thus might be easier to provide than diagnostic labels without any medical training.

A further 38% of the papers focus on localization or segmentation. Typically the goal is to delineate the boundary of an entire healthy structure, or of an abnormality such as a lesion. The difference with how we define the classification task above is that instead of providing information about the image, the annotator has to modify the image, by providing positions or outlines. These tasks rely more on visual characteristics than classification tasks, and may be more easily explained to a non-expert crowd.

In 13% of the papers both classification and segmentation are addressed. Often this means that the annotator first has to indicate if the structure of interest is visible, and if yes, to locate it in the image.

Finally, 7% of papers request less standard tasks from their crowd. For example, Maier-Hein et al. [2015] focus on determining correspondence between pairs of images. Although this is a type of detection task, where the annotator has to locate points of interest in an image, it is also different since a point of reference is already provided. Another example is Ørting et al. [2017], where the annotator has to decide which image is more similar to a reference image. This is a type of classification problem, but again relying more on visual features than on prior knowledge.

B. Type of image

Medical images are acquired at vastly different scales and locations depending on the physiological measurement of interest. The imaging acquisition modality and strategy depends heavily on the scale of the anatomy of interest, and different technologies' expected contrast with surrounding tissues. Here we categorize the images by where in the body the image originates from, which narrows down the modality. We use the following categorization, also used in two recent surveys of medical imaging [Litjens et al., 2017, Cheplygina et al., 2018]: brain, eye, heart, breast, lung, abdomen, histology/microscopy, multiple, other.

We compare the distribution of applications surveyed in this work with the two other surveys in Table I. An interesting observation is that Litjens et al. [2017] and Cheplygina et al. [2018] have a similar distribution of applications despite surveying different topics: Litjens et al. covers deep learning, where a larger dataset is preferred, while Cheplygina et al. covers weakly supervised learning, where datasets are smaller in size. Given that crowdsourcing is often proposed as an alternative to weakly supervised learning, it is surprising that the current survey has a different distribution of papers.

Application	This survey	Cheplygina et al. [2018]	Litjens et al. [2017]
Brain	9%	21%	18%
Eye	15%	4%	5%
Lung	9%	13%	14%
Breast	0%	6%	7%
Heart	2%	4%	7%
Abdomen	22%	14%	9%
Histo/Micro	29%	17%	20%
Multiple	7%	12%	4%
Other	7%	10%	16%

TABLE I
COMPARISON OF THE DISTRIBUTION OF APPLICATIONS IN THIS SURVEY AND TWO OTHER RECENT SURVEYS IN MEDICAL IMAGE ANALYSIS. PERCENTAGES ARE ROUNDED TO THE NEAREST WHOLE NUMBER.

Many of the papers in this survey are aimed at 2D images. The most common application is histopathology/microscopy with 29% of all the papers, followed by retinal images with 15% of the papers. Both applications are over-represented compared to [Litjens et al., 2017] and Cheplygina et al. [2018]. This overrepresentation in crowdsourcing studies may be because many retinal and microscopic images are acquired in 2D, which might be easier to use in a crowdsourcing study than 3D images.

Breast and heart images, which were already not well represented in the other two surveys, are almost absent in crowdsourcing studies. Both applications can be aimed at 2D or 3D images. However, perhaps due to lack of datasets or perceived difficulty of assessing these images, these applications are almost never considered for crowdsourcing.

Several other papers address applications where images are often 3D, such as the brain (9%) and the lungs (9%). Compared to [Litjens et al., 2017, Cheplygina et al., 2018], brain and lung images are underrepresented in crowdsourcing. This could be due to complexity of images or limitations in interfaces. One approach for dealing with 3D images is to select 2D parts of the original 3D images. For example, [Ørting et al., 2017, O'Neil et al., 2017] select axial slices. [Cheplygina et al., 2016] shows patches of 2D projections in various directions in the image. Others circumvent the 3D problem by presenting a video to the users where the entire image is displayed as a sequence of 2D frames [Boorboor et al., 2018]. Only a few of the papers addressing 3D images, present images in 3D [Huang and Hamarneh, 2017, Sonabend et al., 2017].

The last type of data that is addressed is video, common for endoscopy and colonoscopy (both in the abdomen category). Several different approaches are used for presenting video data: 2D frames [Maier-Hein et al., 2014b, 2015, 2016, Heim, 2018, Roethlingshoefer et al., 2017], 3D renderings [Nguyen et al., 2012, McKenna et al., 2012], short video clips [Park et al., 2017], or longer videos that can be paused and annotated [Park et al., 2018].

Other applications of crowdsourcing include segmenting hip joints in 2D MRI [Chávez-Aragón et al., 2013], rating visual characteristics of dermatological images [Cheplygina and Pluim, 2018] and assessing surgical performance [Malpani et al., 2015, Holst et al., 2015]. Two papers [Foncubierta Rodríguez and Müller, 2012, de Herrera et al., 2014] look at multiple applications, where the task is classifying image modality, rather than segmentation or diagnosis. A few papers address segmentation in multiple modalities: [Gurari et al., 2016] focus on both natural and biomedical images, [Lejeune et al., 2017] address segmentation across four medical applications.

III. INTERACTION

An important aspect of crowdsourcing medical image annotations is task design. The interplay between the type of image data, the type of annotations that are needed and the available tools for annotation, needs to be considered to successfully crowdsource annotations. A major component of task design is choosing how workers interact with the task. The type of interaction influences time per task and the required level of expertise and training, which ultimately translates into cost and quality. We identified four categories of interaction tasks across the studies surveyed:

- Rate an entire image
- Draw shapes to identify regions of interest
- Click on specific locations
- Compare two or more images

Furthermore, we also observed that studies generally had crowds either (1) create entirely new annotations on unlabeled data, or (2) make responses based on pre-existing annotations, e.g., output from automated segmentation methods.

Rating entire images was the most common interaction and was the main task of 52% of the studies surveyed here. Ratings took many forms, identifying the presence/absence of specific visual features [Sonabend et al., 2017], counting number of cells [Smittenaar et al., 2018], assessing intensity of cell staining [dos Reis et al., 2015], or discriminating healthy samples from diseased [Mavandadi et al., 2012]. Most commonly, crowd worker were asked to create new annotations (89% of rating tasks). Less commonly, crowd workers were asked to validate pre-existing annotations (14%). One study involved both validating pre-existing annotations and creating new ones [Heim, 2018], so the percentages do not sum to 100%. Existing annotations were the output of automated methods [Roethlingshoefer et al., 2017, Ganz et al., 2017, Gur et al., 2017] half of the time, and the crowdsourced annotations were used to identify instances with errors to be corrected.

Drawing a shape was the second most common task, comprising 37% of studies. Here crowd workers were asked to draw bounding boxes or segment outlines of structures of interest. Sometimes, this was only after identifying if a structure was present in the image or not [Heim, 2018]. Similar to rating images, drawing shapes was used as an interaction for both creating new annotations (90% of drawing tasks) and validating existing annotations (15%). In the case of evaluating existing annotations, drawing was used as a means to indicate the location of errors in segmentations produced by automated methods [Roethlingshoefer et al., 2017, Ganz et al., 2017].

Clicking on specific locations was the third most used interaction, occurring in 26% of studies. Clicking was only used to create new annotations such as identifying the precise location of specific cells, abnormalities, or artifacts within

	Section II		Section III	Section IV			Section V			
Paper	Task	Domain	Interaction	I	Platform	Reward	Filtering	Aggregation	Comparison	Gold standard
Albarqouni et al. [2016a]	classify	histo	rate	M	custom	unknown	before/during	majority/weighted	indirect	multiple experts
Albarqouni et al. [2016b]	other	histo	rate	L	custom	volunteers	none	weighted	direct	multiple experts
Roethlingshoefer et al. [2017]	segment	abdomen	draw	L	none	none	none	none	na	na
Boorboor et al. [2018]	segment	lung	draw	S	paid	low	during	none	direct	multiple experts
Brady et al. [2014]	classify	eye	rate	S	paid	unknown	none	none	direct	?
Brady et al. [2017]	classify	eye	rate	M	paid	low	none	majority/weighted	direct	?
Bruggemann et al. [2018]	segment	histo	click	L	paid	low	none	none	direct	multiple experts
Cabrera-Bean et al. [2017]	segment	histo	click	XS	volunteer	volunteers	none	other	direct	?
Chávez-Aragón et al. [2013]	segment	other	draw	M	custom	volunteers	after	none	direct	other
Cheplygina et al. [2016]	segment	lung	draw	S	paid	unknown	after	average	direct	one expert
Cheplygina and Pluim [2018]	classify	other	rate	S	students	volunteers	none	average	indirect	?
Della Mea et al. [2014]	s+c	histo	click	S	paid	unknown	during	average	direct	one expert
dos Reis et al. [2015]	classify	histo	rate	L	volunteer	volunteers	during	average	direct	one expert
Eickhoff [2014]	classify	histo	rate	M	paid	low	none	majority	direct	one expert
Foncubierta Rodríguez and Müller [2012]	classify	multiple	rate	L	paid	unknown	during	none	direct	one expert
Ganz et al. [2017]	segment	brain	draw	S	paid	low	none	average	direct	one expert
Gur et al. [2017]	classify	heart	rate	M	custom	unknown	none	none	indirect	multiple experts
Gurari et al. [2016]	s+c	multiple	rate+draw	M	paid	low	before/after	majority	direct	multiple experts
Heim [2018]	s+c	abdomen	rate+draw	M	paid	low	before	majority/weighted	direct	multiple experts
Heller et al. [2017]	segment	abdomen	draw	XS	custom	unknown	none	none	direct	other
de Herrera et al. [2014]	classify	multiple	rate	L	paid	volunteers	none	none	indirect	other
Holst et al. [2015]	classify	other	rate	S	paid	low	before	other	direct	multiple experts
Huang and Hamarneh [2017]	classify	lung	click	?	custom	unknown	none	other	direct	?
Irshad et al. [2015]	segment	histo	click+draw	M	paid	unknown	before/during	none	direct	multiple experts
Irshad et al. [2017]	s+c	histo	rate+draw	M	paid	unknown	before/during	majority/weighted	direct	one expert
Keshavan et al. [2018]	classify	brain	rate	M	custom	volunteers	during	weighted	direct	multiple experts
Lawson et al. [2017]	classify	histo	click	S	volunteer	hourly	none	none	direct	multiple experts
Lee and Tufail [2014]	segment	eye	draw	S	paid	low	none	none	direct	na
Lee et al. [2016]	segment	eye	draw	S	paid	low	before	none	direct	na
Leifman et al. [2015]	s+c	eye	rate+draw+click	?	custom	volunteers	during	other	direct	multiple experts
Lejeune et al. [2017]	segment	multiple	click	S	experts	unknown	none	average	indirect	?
Luengo-Oroz et al. [2012]	segment	histo	click	S	volunteer	volunteers	during	other	direct	multiple experts
Maier-Hein et al. [2014a]	segment	abdomen	draw	S	paid	unknown	none	none	indirect	one expert
Maier-Hein et al. [2014b]	other	abdomen	click+compare	S	paid	unknown	none	other	direct	multiple experts
Maier-Hein et al. [2015]	other	abdomen	click+compare	S	paid	low	none	other	direct	?
Maier-Hein et al. [2016]	segment	abdomen	click+compare	M	paid	unknown	none	majority	indirect	?
Malpani et al. [2015]	classify	other	rate+compare	M	?	hourly	during	majority/weighted	direct	multiple experts
Mavandadi et al. [2012]	classify	histo	rate	M	none	unknown	before/during	none	direct	one expert
McKenna et al. [2012]	classify	abdomen	rate	M	paid	low	before	other	direct	?
Mitry et al. [2013]	classify	eye	rate	S	paid	low	none	none	direct	multiple experts
Mitry et al. [2015]	classify	eye	rate	M	paid	low	before/after	none	direct	multiple experts
Mitry et al. [2016]	s+c	eye	rate+draw	S	paid	low	before/after	majority	direct	multiple experts
Nguyen et al. [2012]	classify	abdomen	rate	M	paid	low	before	majority	direct	?
O'Neil et al. [2017]	segment	lung	draw	S	custom	volunteers	after	majority	direct	one expert
Ørting et al. [2017]	other	lung	compare	M	paid	low	before	other	indirect	multiple experts
Park et al. [2017]	classify	abdomen	rate	M	paid	unknown	none	majority	direct	one expert
Park et al. [2018]	segment	abdomen	click	M	paid	unknown	before/during	none	direct	?
Rajchl et al. [2016]	segment	brain	click	L	custom	volunteers	none	none	indirect	one expert
Rajchl et al. [2017]	segment	abdomen	?	M	none	none	none	none	na	na
Sameki et al. [2016]	segment	histo	draw	M	paid	low	none	other	direct	?
Sharma et al. [2017]	segment	histo	draw	S	paid	low	none	other	direct	na
Smittenaar et al. [2018]	classify	histo	rate	L	custom	volunteers	none	weighted	direct	multiple experts
Sonabend et al. [2017]	classify	brain	rate	?	experts	unknown	none	none	direct	na
Sullivan et al. [2018]	classify	histo	rate+draw	L	custom	volunteers	none	other	direct	?
Timmermans et al. [2016]	s+c	brain	rate+draw	M	custom	volunteers	none	none	na	na

TABLE II

OVERVIEW OF PAPERS WITH THE TASK (S+C = SEGMENT AND CLASSIFY), APPLICATION, INTERACTION, NUMBER OF IMAGES (XS=< 10, S=10 TO 100, M=100 TO 1000, L=> 1000), PLATFORM, REWARD, FILTERING, AGGREGATION, TYPE OF COMPARISON USED (NA = NOT APPLICABLE), SOURCE OF GOLD STANDARD. A QUESTION MARK "?" INDICATES THE INFORMATION WAS NOT FOUND IN THE PAPER.

an image. The use of multiple clicks to outline a structure was considered a "drawing a shape" interaction. Selecting points was also used in pairs of video frames to determine the stereotactic correspondence of two video streams for follow-up 3D reconstruction [Maier-Hein et al., 2014b, 2015, 2016].

Comparing two or more images was the least used interaction, occurring in only 5 (9%) of studies. In all cases, comparisons were used to create new annotations, such as marking corresponding points in two consecutive video frames [Maier-Hein et al., 2015, 2016] or to choose which of two images was more similar to a target image [Ørting et al., 2017].

Overall, crowds were more often used to create new annotations, than to make judgments on existing annotations. Ratings and drawing of shapes can be used to obtain more detailed annotations than information already present in datasets. Clicking interactions are sometimes used to identify specific image features, but more commonly used to create bounding boxes or draw object boundaries. Evaluating existing annotations is always done with rating or drawing interactions.

IV. Platform, Scale and Wages

In this section we summarize the main meta parameters and settings of crowdsourcing experiments. First, we classify the reviewed papers based on the type of platform used to perform the crowdsourcing experiments. Second, we report on the scale of the experiments where we consider 1) the number of images annotated and 2) the number of annotators per image. Finally, we summarize the wages paid to crowd workers.

A. Crowdsourcing platforms

A potentially important factor that varies across the surveyed papers is the choice of platform for conducting crowdsourcing experiments. We classify the platforms into six categories: paid commercial marketplaces such as Amazon Mechanical Turk[2] and FigureEight[3] (formerly known as CrowdFlower), volunteers such as Zooniverse[4] and Volunteer Science[5], custom recruitment/platforms, lab participants, experts and simulation or no experiment at all. The most common choice is a commercial platform (53%). The second most common choice is a custom platform (22%) followed by a volunteer platform (10%). The remaining 15% were almost equally divided into the other categories with around 7% of all papers reporting prototypes or simulation studies.

B. Scale

We summarize the scale of the crowdsourcing experiments in terms of number of images annotated and number of annotations per image.

[2]https://www.mturk.com
[3]https://www.figure-eight.com
[4]https://www.zooniverse.org
[5]https://volunteerscience.com

1) Number of images: We classify the number of images into four categories: very small (less than 10 annotated images), small (10 to 100 annotated images), medium (100 to 1000 annotated images) and large (more than 1000 annotated images). The large majority of reviewed papers, 71%, report small and medium scale experiments, while a smaller part report large experiments (21%) or very small experiments (5%). However, in around 3% of the reviewed papers, the scale of the experiments is not reported.

2) Number of annotations per image: We divide the number of annotations per image into two categories: a single annotator per image (5%) or multiple annotators per image (61%). Surprisingly, for 34% of surveyed papers the number of annotations per image is not reported nor can it be inferred.

Overall, the experiments using a single annotator per image involve either simulations or locally recruited, volunteer-based annotators that are not remunerated. The number of annotators per image for experiments using multiple annotators per image ranges from 2 to 5000. However, the majority (70%) of these experiments use between 5 to 25 annotators per image.

C. Annotators Wage

We classify the wage given to annotators into six different categories: a few dollars per hour, less than or equal to $0.10 per annotation, more than $0.10 per annotation, volunteers (no monetary incentive), not specified (if we have no information about compensation) and none (if no actual experiment or recruitment took place).

More than a third (35%) of papers did not specify anything about wage. In 32% of papers the wage was less than or equal to $0.10, in 25% of papers crowds where volunteers with no monetary incentive, in 5% of papers the wage was more than $0.10, and in 3% of papers the wage was an hourly payment of a few dollars per hour.

Overall, very few and mainly the papers that mention an hourly payment considered crowd worker wages in relation to minimum wage rules and regulations.

V. Evaluation

In this section we describe how the crowdsourced annotations are evaluated. This is done via two strategies:

- ensuring sufficient quality of annotations by preprocessing
- estimating the utility of the crowd annotations for the task at hand

Although the two strategies are closely related and should be considered jointly when designing crowdsourcing experiments, it is informative to consider them separately here.

The first strategy is closely related to the field of quality control in crowdsourcing. Numerous approaches exist to tackle this, starting from simple majority voting and worker filtering to sophisticated statistical and machine learning methods that consider workers' specific skills, task difficulty and clarity of task descriptions. The second strategy is more domain-specific, as different tasks may have different levels of tolerance for errors.

A. Preprocessing of annotations

Preprocessing of annotations broadly covers what is done to the crowdsourced annotations prior to using them for their intended purpose. It includes filtering individual annotations and/or aggregating annotations. Of the surveyed papers, 84% perform some form of preprocessing.

1) Filtering individuals: One way to filter annotations, is to remove annotations made by "poorly performing" annotators. Most crowdsourcing platforms offer a rating score for workers that provides an estimate of their performance, based on their percentage of previously approved tasks. This score is used in 15% of surveyed papers to filter workers prior to assigning tasks. A related approach, used in 13% of surveyed papers, is to exclude workers that fail a test task prior to the actual tasks. A refinement of this, used in 24% of surveyed papers, is to integrate separate test tasks in the tasks and exclude workers that fail the tests. One example is adding a smiley face to colonoscopy videos to ensure attention [Park et al., 2018].

Another common filtering approach for individual workers, used in 22% of surveyed papers, is comparing annotations to gold standard annotations. In this case, tasks with known gold standard annotations, are injected into the regular working process. A worker's correspondence with the gold standard can then be used to estimate individual worker performance. In contrast to platform scores and unrelated test tasks, this approach assesses worker performance on the specific task, allowing more fine-grained worker selection.

2) Aggregating results: One of the main benefits of crowdsourcing is the fast and cost-effective collection of a large number of annotations. This allows aggregating annotations to reduce noise in the individual annotations.

Majority voting is widely used due to its computational and conceptual simplicity and was found in 22% of the papers. In the context of medical image analysis, majority voting is applied to annotations, ratings, and also to aggregate slices of images. One example is presented in Heim [2018] where the authors used crowdsourcing for organ segmentation in computed tomography scan. Multiple organ outlines are collected via an online tool and pixel-wise majority voting is applied to improve the accuracy of the segmentation.

In the case of numerical ratings, mean and median statistics are also used in 13% of the papers to determine a final annotation. For example, Cheplygina et al. [2016] use median to aggregate the areas of the annotations created by individual workers.

A more sophisticated version of the majority vote uses additional information about the general quality of workers. This information can be derived if workers perform multiple tasks or if gold standard data is available. Weighted voting is used in 16% of surveyed papers, for example in [Keshavan et al., 2018], where the XGBoost algorithm was used to estimate annotator weights and in [Brady et al., 2017] where task difficulty is taken into account and annotator weights are estimated as the probability an annotator is correct conditioned on the difficulty of the task,

B. Evaluating annotations

Evaluating how well crowd annotations solve the intended purpose is most commonly (78% of surveyed papers) done by directly comparing crowdsourced annotations to a gold standard. In about 16% of surveyed papers crowd annotations are used for training a machine learning method, and the performance of the machine learning method used to indirectly evaluate annotations. The remaining 5% have no evaluation of how well annotations solve the intended purpose.

The gold standard originates from different sources. In about 20% of surveyed papers the gold standard is based on a single expert, in about 36% the gold standard is based on multiple experts and in the remaining papers the number of experts is not reported or no expert gold standard is used. Using a gold standard based on a single expert can be problematic since experts often disagree on all but the most trivial tasks. However, only 3 of 20 papers that use multiple experts consider how well experts agree.

Expert-based gold standards are generally not obtained from experts performing exactly the same task as the crowd. In several cases the only difference in expert and crowd tasks is due to differences in user interface, e.g. a clinical workstation for experts and a web interface for crowds. As long as the fundamental task is the same (e.g. count cells) and the user interface has not been dramatically changed we consider the expert and crowd tasks to be the same. Using this definition, about 40% of the papers use the same task and about 40% use a different task. In the remaining 20% it is either not reported or no expert gold standard is used.

There are several reasons for asking crowds to perform a different task than what experts have done for the gold standard. Some papers use a simplified version of the expert task in order to make the task easier or more suitable as a small self-contained task. For example, ranking relative performance in pairs of surgical videos instead of grading performance in each [Malpani et al., 2015]; assessing visual similarity of images instead of classifying disease patterns [Ørting et al., 2017]; refining segmentation proposals instead of performing a full segmentation [Maier-Hein et al., 2016]; annotating polyps in a single frame instead of in a full video [Park et al., 2017]; counting stained cells instead of classifying disease status [Irshad et al., 2017]. Other papers focus on changing the user interface, such as in [Lejeune et al., 2017] where an eye tracker is used for segmentation instead of a mouse, or in [Albarqouni et al., 2016b, Mavandadi et al., 2012] where the user interface is changed to support gamification strategies.

In a few papers, evaluation is focused on variation in annotations. For example, in [Lee and Tufail, 2014, Lee et al., 2016] where annotations are evaluated in terms of inter-rater reliability; and in [Heller et al., 2017, Huang and Hamarneh, 2017, Leifman et al., 2015, Sonabend et al., 2017] where individual annotations are compared to aggregated annotations. Measuring variability of annotations it not directly useful for evaluating the correctness of annotations. However, annotation variability is essential when evaluating how much the crowdsourced annotations can be trusted. Additionally, variation provides an indirect measure of correctness. Large variation

can indicate that annotations are often wrong, while small variation indicates that annotations are often correct or the task has been designed such that annotators are consistently wrong.

VI. Results and recommendations

Here, we provide an overview of the primary results and recommendations emerging from the papers examined in this review. Complementary to the topics discussed in Section V we consider (1) How effective is the application of crowdsourcing to medical image analysis? (2) Recommendations to ensure data quality.

A. How effective is the application of crowdsourcing to medical image analysis?

The vast majority of studies examined in this review found crowdsourcing to be a valid approach for data production. Crowdsourcing of medical image analysis was noted to be an accurate approach [Lawson et al., 2017], that can produce large quantities of annotations needed to solve high-throughput problems requiring human input [Irshad et al., 2015, dos Reis et al., 2015, Lee and Tufail, 2014, Maier-Hein et al., 2014b]. Crowdsourcing can be used to create new annotations or make existing data more robust, both cheaper and faster than annotation by medical experts [Rajchl et al., 2016, Holst et al., 2015, Gurari et al., 2016, Eickhoff, 2014, Park et al., 2017].

Although the relative efficacy of crowdsourcing applied to medical image analysis will be dependent on the complexity of the task, the papers examined here show crowdsourcing to be an effective methodology across a wide variety of applications, including objective assessment of surgical skill [Malpani et al., 2015], emphysema assessment [Ørting et al., 2017], polyp marking in virtual colonoscopy [Park et al., 2018], identification of chromosomes [Sharma et al., 2017] and biomarker discovery in immunohistochemistry data [Smittenaar et al., 2018]. Notably, only one project stated that crowdsourcing could not always be applied effectively to the studied task ("it is very difficult and maybe even impossible to entirely outsource the task of labelling mitotic figures in histology images to crowds" [Albarqouni et al., 2016a]).

Rather than comparing the absolute performance of the crowd to experts or to algorithms, it might be worth considering their relative benefits. For example, crowds were particularly useful for rare classes [Sullivan et al., 2018], which are often difficult cases for algorithms. Another situation where crowds can be useful is identifying data that is missing from the gold standard provided by experts, see for example [Luengo-Oroz et al., 2012]. Benefits of combining crowds with algorithms were also demonstrated by [Albarqouni et al., 2016a, Sharma et al., 2017].

B. Recommendations to ensure data quality

The papers examined in this review included suggestions to improve the quality of data produced through crowdsourcing. These suggestions focused on refining the task design, crowdsourcing platform and post-processing of annotations. We summarize these recommendations here.

1) Task design: As discussed, crowdsourcing has been applied effectively to many medical imaging applications. However, careful study design remains necessary to ensure generation of data of sufficient quality.

The selection and design of an appropriate crowdsourcing task is central to project success. Effort should be made to make the task simple and unambiguous [Rajchl et al., 2016, Gurari et al., 2016], and to present study data appropriately [McKenna et al., 2012]. For unavoidably challenging tasks, crowdsourcing may still provide useful data, for instance, through enabling a rapid first-pass evaluation of large scale data sets [Della Mea et al., 2014, Park et al., 2017]. Particularly challenging tasks may be made tractable through gamification [Albarqouni et al., 2016b] or careful reframing of the task, e.g. crowdsourcing of emphysema assessment was made possible through reframing the task as a question of image similarity [Ørting et al., 2017]. Alternatively, it may be possible to achieve the desired data quality simply through asking a larger cohort of crowd workers to perform each task per data point. An interesting example of task design is given in [Gurari et al., 2016] where quality and speed of crowdsourced segmentations in natural images are increased by flipping images, suggesting that familiarity with an image can be detrimental.

2) Crowdsourcing platform: The choice of crowdsourcing platform can influence study cost and completion time, as well as the size and demographics of the crowd. Furthermore, different platforms offer distinct features which may influence the quality of data produced. For example, Heller et al. noted that user interface features, such as zoom and intuitive controls, can increase data quality. Contingent on the complexity of the task and interface design, training materials should be provided, as this can improve results [McKenna et al., 2012]. However, this is not always necessary - in some cases minimal [Brady et al., 2014] or no training [Ganz et al., 2017] was required.

3) Post-processing: Post-processing of annotations is recommended to improve annotation quality by removing annotations from poorly performing workers. Alternatively, if multiple workers annotate the same data it is possible to improve annotation quality by aggregating annotations.

The surveyed papers propose a variety of criteria for filtering individual annotations. For example, time spend on task [O'Neil et al., 2017], expected shape of segmentation [Cheplygina et al., 2016, Chávez-Aragón et al., 2013], correlation with other workers' results [Sharma et al., 2017, Chávez-Aragón et al., 2013] and correlation to experts annotations or ground truth [Sameki et al., 2016, Keshavan et al., 2018, Irshad et al., 2017, 2015, Foncubierta Rodríguez and Müller, 2012]. However, due to the lack of comparisons between different filtering approaches, the only clear recommendation from these works is to use some form of filtering.

Contrary to this recommendation, Nguyen et al. found that filtering unreliable workers did not have a significant influence when annotations from multiple workers are aggregated. However, aggregating without taking individual performance into account might not be the best approach. Malpani et al. compared different aggregation methods, and found that weighted voting, with weights based on self-reported confidence scores, improved results compared to simple majority voting. Simi-

larly, Irshad et al. found that aggregating segmentations from 3-5 workers, using weights based on consensus and worker trust scores, improved performance over using single worker annotations. Further, Cheplygina and Pluim [2018] found that disagreement between workers was predictive of melanoma diagnosis in skin lesions, suggesting that simple aggregation, such as majority voting or mean statistics, might not be the best approach.

VII. Discussion

A. Trends

As discussed in Section II, crowdsourcing is applied to a variety of medical images, however, it is most commonly applied to histology or microscopy images. The trend for crowdsourcing of this image type may be due to the ease of which these (typically 2D) images can be incorporated into a crowdsourcing or citizen science project. Alternatively, the microscopy images examined in these papers may have not been derived from a patient, and would therefore not require the consent of an individual to use for crowdsourcing purposes.

The most common crowd task is rating entire images. This is somewhat surprising, given that we would expect such tasks to rely more on prior knowledge than other crowdsourced tasks, such as drawing outlines of objects. Again, this trend might be facilitated due to the ease with which rating images can be incorporated in existing platforms.

Most crowdsourcing studies are set up on commercial platforms, followed by custom platforms. Each image is annotated by multiple crowd workers, who typically receive less than $0.10 per annotation. On the one hand, this low reimbursement might be a product of researchers trying to optimize the total number of annotations given a particular budget. On the other hand, it could be a lack of awareness of what appropriate compensation should be [Hara et al., 2017].

A surprising finding is that, often, important details about the crowd and their compensation are missing. Besides missing details in terms of crowd compensation, we find missing details regarding the number of requested annotations per unit. While for some of the surveyed papers, we could infer an approximation of the number of annotations gathered per unit by checking the scale of the experiment and the total amount of annotations gathered, for at least a third of the surveyed papers (34%) this was not possible due to a lack of detail when describing the crowdsourcing experimental methodology.

Crowdsourced annotations are generally processed prior to evaluating how well the annotations solved the intended purpose. Simply excluding workers based on platform scores or a single test task is not as popular as continuously monitoring worker performance. 29% of the surveyed papers aggregate annotations from multiple crowd workers. This is most commonly done by simple majority voting, but some papers use estimates of task difficulty and/or worker performance to obtain a weighted aggregation.

The most common approach to evaluating the quality of preprocessed annotations is by comparing to an expert defined gold standard. A smaller set of papers use the annotations to train an ML method and evaluate the performance of the trained method.

The studies we reviewed almost unanimously conclude that crowdsourcing is a a viable solution for medical image annotation, which may seem unexpected given the complexity of medical imaging as a field in general. There might be several possible reasons for the lack of negative results. One is researchers selecting tasks which they already expect to be suitable for crowdsourcing. Another reason is publication bias, with papers demonstrating negative results having less chance of being published, which is also an issue in computer vision [Borji, 2018].

B. Limitations

There are a number of limitations in the way that the current studies are being conducted. There is generally a lack of clarity in the reporting of experimental design and evaluation protocols. Additionally, ethical questions regarding worker compensation, image content and patient privacy are rarely discussed, but seem crucial to address. In several papers the study design appears to be ad-hoc. Characteristics such as the platform, number of annotators, how the task is explained and so forth, are not always motivated, or even described. This creates difficulties in understanding what leads to a successful crowdsourcing study and increases the barrier for researchers who have not used crowdsourcing before. The studies which do examine such factors are often conducted on a single application, making it difficult to generalize lessons learned to other applications. Detailed documentation of experiments is a crucial factor for ensuring reproducible science and essential for replication studies.

Another problem is the evaluation of results. The quality of crowdsourced annotations is generally estimated by comparing directly to expert annotations. However, variation in both expert defined gold standard and crowd annotations are not systematically accounted for, making it difficult to assess if crowd annotations are actually good enough. When using annotations to train ML methods, noisy crowd annotations might not be a problem if handled by the ML algorithm. However, variation in annotations should still be investigated in this case. A related problem is using expert annotations to filter crowd annotations, which would not be possible for real unlabeled data, thus leading to overly optimistic results.

Overall, surveyed papers reported successful results. However, from our personal experience and discussions with other researchers, it is non-trivial to setup a crowdsourcing project for medical images. Due to the lack of negative results, the current literature does not inform researchers inexperienced with crowdsourcing about the main considerations of such a project. Furthermore, very few articles report on pilot experiments which aim to calibrate and identify the optimal crowdsourcing parameter settings such as the number of annotators per image.

There are important ethical issues which are largely not mentioned in the papers we surveyed. First of all, details about compensation are often missing, whereas this can have an important effect on the crowd [Hara et al., 2017]. Furthermore, what is reasonable compensation in one country, may be too low for another country due to different cost of living. How to set the compensation fairly is an open issue that researchers should consider in their work.

Another ethical concern is whether it is possible and/or appropriate to share images with the crowd. Some images (for example of surgery) may be traumatic to view or unsuitable for children, which is more unique to the medical domain than other areas where crowdsourcing is applied e.g. astronomy or ecology projects. Another issue is sharing images from the perspective of patient consent, which is an issue that must be considered case by case.

C. Opportunities

Several papers discuss directions they want to take in further research. One of the popular directions is increasing the role of machine learning. Several papers not using machine learning plan to do so in future [Brady et al., 2017, Cheplygina and Pluim, 2018, Sullivan et al., 2018]. Papers that already use machine learning discuss improvements to their algorithms or crowd-algorithm combinations [Sharma et al., 2017, Sameki et al., 2016].

Related to the above, tailoring the tasks to individual workers is another possibility. The rating score given to workers by platforms only reflects an overall completion rate, and might be artificially high because employers tend to rate the majority of the tasks positive and apply a filtering afterwards. Considering worker scores on different task types could help to make a better selection of workers beforehand.

Another strategy discussed as future work is the use of gamification. Several papers have explored this idea [Luengo-Oroz et al., 2012, Mavandadi et al., 2012, Albarqouni et al., 2016b, Sullivan et al., 2018] citing increased motivation of annotators. While the earlier papers [Luengo-Oroz et al., 2012, Mavandadi et al., 2012, Albarqouni et al., 2016b] have task-specific games, Sullivan et al. [2018] takes a more task-independent approach of a mini-game within an existing, larger game. This could be an opportunity for many other researchers, without the need to design a game from scratch. Finally, annotating images at a festival as in [Timmermans et al., 2016] could be an interesting direction.

Beyond the opportunities that the papers discuss as future research, we see a number of other future directions for the community as a whole. Perhaps the most important future direction is openly sharing our experiences with crowdsourcing, including failures. Due to publication bias, current papers may not reflect the performance and difficulties encountered in a typical crowdsourcing project.

More generally, there is an opportunity to create a set of guidelines for crowdsourcing medical imaging studies. Rather than relying on ad-hoc choices, researchers could then make informed decisions about the platform, reward of the annotators and other variables. For example, the European Citizen Science Initiative has a selection of guides for performing citizen science studies[6]. A further opportunity is to interact more with other fields where crowdsourcing has been in use longer, and to see which of their best practices are also applicable to medical imaging.

[6]https://ecsa.citizen-science.net/blog/collection-citizen-science-guidelines-and-publications

Interacting with workers could both improve projects and help establish guidelines. Workers have created communities (e.g. Reddit, Facebook) and discussion boards (https://www.mturkforum.com, https://www.turkernation.com) for some platforms. Chandler et al. found that 28% $\pm$ 5% of the workers on Mechanical Turk read discussion boards and blogs related to Mechanical Turk. The topics of conversations, in order of frequency, are: pay, gratification, completion time, difficulty, how to successfully complete, purpose and the requesters' reputation of the HIT. These forums are a valuable source for researchers for gathering information, measuring opinions and getting feedback on improving their project. This is particularly important because high throughput workers are more likely to discuss HITs [Chandler et al., 2014]. This subgroup (less than 10 % of the workers do more than 75% of the work [Hara et al., 2017]) is likely to have experience with similar tasks [Chandler et al., 2014], and interaction with these workers may result in various improvements such as improvements of the user interface as in [Bruggemann et al., 2018].

Next to image analysis, crowdsourcing could also be a way to collect, rather than curate, data to improve medical knowledge. This could vary from donating your own medical images (such as http://www.medicaldatadonors.org) to contributing experiences about rare diseases. Since such initiatives do not focus on image analysis we did not include them in this survey, however [Ranard et al., 2014, Wazny, 2017] may be good starting points for readers interested in these topics.

ACKNOWLEDGMENTS

The authors would like to thank eScience-Lorentz grant 2018 and Ms Gerda Filippo (Lorentz center) for their support in organizing the workshop where this paper was conceived.

REFERENCES

S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5): 1313–1321, May 2016a.

S. Albarqouni, S. Matl, M. Baust, N. Navab, and S. Demirci. Playsourcing: a novel concept for knowledge creation in biomedical research. In *Deep Learning and Data Labeling for Medical Applications*, pages 269–277. Springer, 2016b.

R. Alialy, S. Tavakkol, E. Tavakkol, A. Ghorbani-Aghbologhi, A. Ghaffarieh, S. H. Kim, and C. Shahabi. A review on the applications of crowdsourcing in human pathology. *Journal of pathology informatics*, 9, 2018.

S. Boorboor, S. Nadeem, J. H. Park, K. Baker, and A. Kaufman. Crowdsourcing lung nodules detection and annotation. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, page 105791D. International Society for Optics and Photonics, 2018.

A. Borji. Negative results in computer vision: A perspective. *Image and Vision Computing*, 69:1–8, 2018.

C. J. Brady, A. C. Villanti, J. L. Pearson, T. R. Kirchner, O. Gup, and C. Shah. Rapid grading of fundus photos for diabetic retinopathy using crowdsourcing. *Investigative Ophthalmology & Visual Science*, 55(13):4826–4826, 2014.

C. J. Brady, L. I. Mudie, X. Wang, E. Guallar, and D. S. Friedman. Improving consensus scoring of crowdsourced data using the rasch model: development and refinement of a diagnostic instrument. *Journal of medical Internet research*, 19(6), 2017.

J. Bruggemann, G. C. Lander, and A. I. Su. Exploring applications of crowdsourcing to cryo-EM. *Journal of structural biology*, 203(1):37–45, 2018.

M. Cabrera-Bean, A. Pages-Zamora, C. Diaz-Vilor, M. Postigo-Camps, D. Cuadrado-Sánchez, and M. A. Luengo-Oroz. Counting malaria parasites with a two-stage EM based algorithm using crowsourced data. In *Engineering in Medicine and Biology Society (EMBC)*, pages 2283–2287. IEEE, 2017.

J. Chandler, P. Mueller, and G. Paolacci. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, 46(1):112–130, 2014.

A. Chávez-Aragón, W.-S. Lee, and A. Vyas. A crowdsourcing web platform-hip joint segmentation by non-expert contributors. In *Medical Measurements and Applications Proceedings (MeMeA), 2013 IEEE International Symposium on*, pages 350–354. IEEE, 2013.

V. Cheplygina and J. P. W. Pluim. Crowd disagreement about medical images is informative. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*, pages 105–111. Springer, 2018.

V. Cheplygina, A. Perez-Rovira, W. Kuo, H. Tiddens, and M. de Bruijne. Early experiences with crowdsourcing airway annotations in chest CT. In *Large-scale Annotation of Biomedical data and Expert Label Synthesis (MICCAI LABELS)*, pages 209–218, 2016.

V. Cheplygina, M. de Bruijne, and J. P. Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *arXiv preprint arXiv:1804.06353*, 2018.

A. G. S. de Herrera, A. Foncubierta-Rodríguez, D. Markonis, R. Schaer, and H. Müller. Crowdsourcing for medical image classification. *Swiss Medical Informatics*, 30, 2014.

V. Della Mea, E. Maddalena, S. Mizzaro, P. Machin, and C. A. Beltrami. Preliminary results from a crowdsourcing experiment in immunohistochemistry. In *Diagnostic pathology*, volume 9, page S6. BioMed Central, 2014.

F. J. C. dos Reis, S. Lynn, H. R. Ali, D. Eccles, A. Hanby, E. Provenzano, C. Caldas, W. J. Howat, L.-A. McDuffus, B. Liu, et al. Crowdsourcing the general public for large scale molecular pathology studies in cancer. *EBioMedicine*, 2(7):681–689, 2015.

C. Eickhoff. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Gamification for Information Retrieval (GamifIR)*, pages 53–56. ACM, 2014.

A. Foncubierta Rodríguez and H. Müller. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In *ACM Multimedia workshop on Crowdsourcing for Multimedia*, pages 9–14. ACM, 2012.

M. Ganz, D. Kondermann, J. Andrulis, G. M. Knudsen, and L. Maier-Hein. Crowdsourcing for error detection in cortical surface delineations. *International journal of computer assisted radiology and surgery*, 12(1):161–166, 2017.

H. Greenspan, B. Van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

Y. Gur, M. Moradi, H. Bulu, Y. Guo, C. Compas, and T. Syeda-Mahmood. Towards an efficient way of building annotated medical image collections for big data studies. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 87–95. Springer, 2017.

D. Gurari, D. Theriault, M. Sameki, B. Isenberg, T. A. Pham, A. Purwada, P. Solski, M. Walker, C. Zhang, J. Y. Wong, and M. Betke. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In *Winter Conference on Applications of Computer Vision, (WACV)*, pages 1169–1176, 2015.

D. Gurari, M. Sameki, and M. Betke. Investigating the influence of data familiarity to improve the design of a crowdsourcing image annotation system. In *Human Computation (HCOMP)*, 2016.

K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. Bigham. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. *arXiv preprint arXiv:1712.05796*, 2017.

E. Heim. *Large-scale medical image annotation with quality-controlled crowdsourcing*. PhD thesis, German Cancer Research Center (DKFZ), 2018.

N. Heller, P. Stanitsas, V. Morellas, and N. Papanikolopoulos. A web-based platform for distributed annotation of computerized tomography scans. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*, pages 136–145. Springer, 2017.

D. Holst, T. M. Kowalewski, L. W. White, T. C. Brand, J. D. Harper, M. D. Sorensen, M. Truong, K. Simpson, A. Tanaka, R. Smith, et al. Crowd-sourced assessment of technical skills: differentiating animate surgical skill through the wisdom of crowds. *Journal of endourology*, 29(10):1183–1188, 2015.

J. Howe. The rise of crowdsourcing. *Wired magazine*, 14(6): 1–4, 2006.

M. Huang and G. Hamarneh. SwifTree: Interactive extraction of 3D trees supporting gaming and crowdsourcing. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*, pages 116–125. Springer, 2017.

H. Irshad, L. Montaser-Kouhsari, G. Waltz, O. Bucur, J. Nowak, F. Dong, N. W. Knoblauch, and A. H. Beck.

Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. In *Pacific Symposium on Biocomputing*, pages 294–305. World Scientific, 2015.

H. Irshad, E.-Y. Oh, D. Schmolze, L. M. Quintana, L. Collins, R. M. Tamimi, and A. H. Beck. Crowdsourcing scoring of immunohistochemistry images: Evaluating performance of the crowd and an automated computational method. *Scientific Reports*, 7:43286, 2017.

A. Keshavan, J. Yeatman, and A. Rokem. Combining citizen science and deep learning to amplify expertise in neuroimaging. *bioRxiv*, page 363382, 2018.

A. Kovashka, O. Russakovsky, L. Fei-Fei, and K. Grauman. Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 10(3):177–243, 2016.

J. Lawson, R. J. Robinson-Vyas, J. P. McQuillan, A. Paterson, S. Christie, M. Kidza-Griffiths, L.-A. McDuffus, K. A. Moutasim, E. C. Shaw, A. E. Kiltie, et al. Crowdsourcing for translational research: analysis of biomarker expression using cancer microarrays. *British journal of cancer*, 116(2): 237, 2017.

A. Y. Lee and A. Tufail. Mechanical turk based system for macular OCT segmentation. *Investigative Ophthalmology & Visual Science*, 55(13):4787–4787, 2014.

A. Y. Lee, C. S. Lee, P. A. Keane, and A. Tufail. Use of mechanical turk as a mapreduce framework for macular OCT segmentation. *Journal of ophthalmology*, 2016, 2016.

G. Leifman, T. Swedish, K. Roesch, and R. Raskar. Leveraging the crowd for annotation of retinal images. In *International Conference of the Engineering in Medicine and Biology Society (EMBC)*, pages 7736–7739. IEEE, 2015.

L. Lejeune, M. Christoudias, and R. Sznitman. Expected exponential loss for gaze-based video and volume ground truth annotation. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*, pages 106–115. Springer, 2017.

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

M. A. Luengo-Oroz, A. Arranz, and J. Frean. Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. *Journal of medical Internet research*, 14(6), 2012.

L. Maier-Hein, S. Mersmann, D. Kondermann, S. Bodenstedt, A. Sanchez, C. Stock, H. G. Kenngott, M. Eisenmann, and S. Speidel. Can masses of non-experts train highly accurate image classifiers? In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 438–445. Springer, 2014a.

L. Maier-Hein, S. Mersmann, D. Kondermann, et al. Crowdsourcing for reference correspondence generation in endoscopic images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 349–356. Springer, 2014b.

L. Maier-Hein, D. Kondermann, T. Roß, S. Mersmann, E. Heim, S. Bodenstedt, H. G. Kenngott, A. Sanchez, M. Wagner, A. Preukschas, et al. Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences. *International Journal of Computer Assisted Radiology and Surgery*, 10(8):1201–1212, 2015.

L. Maier-Hein, T. Ross, J. Gröhl, B. Glocker, S. Bodenstedt, C. Stock, E. Heim, M. Götz, S. Wirkert, H. Kenngott, et al. Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 616–623. Springer, 2016.

A. Malpani, S. S. Vedula, C. C. G. Chen, and G. D. Hager. A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *International journal of computer assisted radiology and surgery*, 10(9):1435–47, Sept. 2015.

S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere, S. Padmanabhan, K. Nielsen, and A. Ozcan. Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study. *PLoS ONE*, 7 (5), 2012.

M. T. McKenna, S. Wang, T. B. Nguyen, J. E. Burns, N. Petrick, and R. M. Summers. Strategies for improved interpretation of computer-aided detections for ct colonography utilizing distributed human intelligence. *Medical image analysis*, 16(6):1280–1292, 2012.

D. Mitry, T. Peto, S. Hayat, J. E. Morgan, K.-T. Khaw, and P. J. Foster. Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis of images in the epic norfolk cohort on behalf of the ukbiobank eye and vision consortium. *PLoS ONE*, 8(8):e71154, 2013.

D. Mitry, T. Peto, S. Hayat, P. Blows, J. Morgan, K.-T. Khaw, and P. J. Foster. Crowdsourcing as a screening tool to detect clinical features of glaucomatous optic neuropathy from digital photography. *PLoS ONE*, 10(2):1–8, 2015.

D. Mitry, K. Zutis, B. Dhillon, T. Peto, S. Hayat, K.-T. Khaw, J. E. Morgan, W. Moncur, E. Trucco, and P. J. Foster. The accuracy and reliability of crowdsource annotations of digital retinal images. *Translational vision science & technology*, 5(5):6–6, 2016.

T. B. Nguyen, S. Wang, V. Anugu, N. Rose, M. McKenna, N. Petrick, J. E. Burns, and R. M. Summers. Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology*, 262(3):824–833, 2012.

A. Q. O'Neil, J. T. Murchison, E. J. van Beek, and K. A. Goatman. Crowdsourcing labels for pathological patterns in ct lung scans: Can non-experts contribute expert-quality ground truth? In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*, pages 96–105. Springer, 2017.

S. N. Ørting, V. Cheplygina, J. Petersen, L. H. Thomsen, M. M. W. Wille, and M. de Bruijne. Crowdsourced emphysema assessment. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*, pages 126–135. Springer, 2017.

J. H. Park, S. Mirhosseini, S. Nadeem, J. Marino, A. Kaufman, K. Baker, and M. Barish. Crowdsourcing for identification of polyp-free segments in virtual colonoscopy videos. In *Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications*, volume 10138, page 101380V. International Society for Optics and Photonics, 2017.

J. H. Park, S. Nadeem, J. Marino, K. Baker, M. Barish, and A. Kaufman. Crowd-assisted polyp annotation of virtual colonoscopy videos. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, page 105790M. International Society for Optics and Photonics, 2018.

M. Rajchl, M. C. Lee, F. Schrans, A. Davidson, J. Passerat-Palmbach, G. Tarroni, A. Alansary, O. Oktay, B. Kainz, and D. Rueckert. Learning under distributed weak supervision. *arXiv preprint arXiv:1606.01100*, 2016.

M. Rajchl, L. M. Koch, C. Ledig, J. Passerat-Palmbach, K. Misawa, K. Mori, and D. Rueckert. Employing weak annotations for medical image analysis problems. *arXiv preprint arXiv:1708.06297*, 2017.

B. L. Ranard, Y. P. Ha, Z. F. Meisel, D. A. Asch, S. S. Hill, L. B. Becker, A. K. Seymour, and R. M. Merchant. Crowdsourcing: harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, 29(1):187–203, Jan. 2014.

V. Roethlingshoefer, S. Bittel, H. Kenngott, M. Wagner, S. Bodenstedt, T. Ross, S. Speidel, and M.-H. L. How to create the largest in-vivo endoscopic dataset. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*, 2017.

M. Sameki, D. Gurari, and M. Betke. ICORD: Intelligent Collection of Redundant Data ? A Dynamic System for Crowdsourcing Cell Segmentations Accurately and Efficiently. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1380–1389, 2016.

M. Sharma, O. Saha, A. Sriraman, R. Hebbalaguppe, L. Vig, and S. Karande. Crowdsourcing for chromosome segmentation and deep classification. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 786–793. IEEE, 2017.

P. Smittenaar, A. K. Walker, S. McGill, C. Kartsonaki, R. J. Robinson-Vyas, J. P. McQuillan, S. Christie, L. Harris, J. Lawson, E. Henderson, et al. Harnessing citizen science through mobile phone technology to screen for immunohistochemical biomarkers in bladder cancer. *British journal of cancer*, 119(2):220, 2018.

A. M. Sonabend, B. E. Zacharia, M. B. Cloney, A. Sonabend, C. Showers, V. Ebiana, M. Nazarian, K. R. Swanson, A. Baldock, H. Brem, et al. Defining glioblastoma resectability through the wisdom of the crowd: a proof-of-principle study. *Neurosurgery*, 80(4):590–601, 2017.

D. P. Sullivan, C. F. Winsnes, L. Åkesson, M. Hjelmare, M. Wiking, R. Schutten, L. Campbell, H. Leifsson, S. Rhodes, A. Nordgren, et al. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature biotechnology*, 36(9):820, 2018.

B. Timmermans, Z. Szlávik, and R.-J. Sips. Crowdsourcing ground truth data for analysing brainstem tumors in children. In *Belgium Netherlands Artificial Intelligence Conference (BNAIC)*, 2016.

K. Wazny. Crowdsourcing ten years in: A review. *Journal of global health*, 7(2), 2017.

5 Learning from visual similarity assessment

This section is based on two manuscripts. The first [33] investigates training convolutional neural networks (CNNs) using similarity measurements derived from visual scoring. The second [36] investigates how visual similarity assessments can be obtained and used to learn expressive representations using CNNs.

[33] Silas Nyboe Ørting, Jens Petersen, Veronika Cheplygina, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. *Feature learning based on visual similarity triplets in medical image analysis: A case study of emphysema in chest CT scans*. In Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS 2018). 2018.

[36] Silas Nyboe Ørting, Jens Petersen, Veronika Cheplygina, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. *Visual similarity comparisons for medical image analysis*. In preparation, 2019.

Feature Learning Based on Visual Similarity Triplets in Medical Image Analysis: A Case Study of Emphysema in Chest CT Scans

Silas Nyboe Ørting[1(✉)], Jens Petersen[1], Veronika Cheplygina[2], Laura H. Thomsen[3], Mathilde M. W. Wille[4], and Marleen de Bruijne[1,5]

[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
silas@di.ku.dk
[2] Medical Image Analysis (IMAG/e), Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[3] Department of Internal Medicine, Hvidovre Hospital, Copenhagen, Denmark
[4] Department of Diagnostic Imaging, Bispebjerg Hospital, Copenhagen, Denmark
[5] Biomedical Imaging Group Rotterdam, Departments of Radiology and Medical Informatics, Erasmus MC - University Medical Center Rotterdam, Rotterdam, The Netherlands

Abstract. Supervised feature learning using convolutional neural networks (CNNs) can provide concise and disease relevant representations of medical images. However, training CNNs requires annotated image data. Annotating medical images can be a time-consuming task and even expert annotations are subject to substantial inter- and intra-rater variability. Assessing visual similarity of images instead of indicating specific pathologies or estimating disease severity could allow non-experts to participate, help uncover new patterns, and possibly reduce rater variability. We consider the task of assessing emphysema extent in chest CT scans. We derive visual similarity triplets from visually assessed emphysema extent and learn a low dimensional embedding using CNNs. We evaluate the networks on 973 images, and show that the CNNs can learn disease relevant feature representations from derived similarity triplets. To our knowledge this is the first medical image application where similarity triplets has been used to learn a feature representation that can be used for embedding unseen test images.

Keywords: Feature learning · Similarity triplets
Emphysema assessment

1 Introduction

Recent years have demonstrated the enormous potential of applying convolutional neural networks (CNNs) for medical image analysis. One of the big challenges when training CNNs is the need for annotated image data. Annotating

D. Stoyanov et al. (Eds.): CVII-STENT 2018/LABELS 2018, LNCS 11043, pp. 140–149, 2018.
https://doi.org/10.1007/978-3-030-01364-6_16

medical images can be a time-consuming and difficult task requiring a high level of expertise. A common issue with annotations is substantial inter- and intra-rater variability. There are many sources of rater variability in annotations, for example, level of expertise, time-constraints and task definition. A common approach to defining annotation tasks is to ask raters for an absolute judgment, "segment the tumor", "count number of nodules", "assess extent of emphysema". Evidence from social psychology suggests humans in some cases are better at making comparative ratings than absolute ratings [3,5,11]. Redefining annotation tasks in terms of relative comparisons could improve rater agreement.

An annotation task that is especially prone to rater variations and may be better suited for comparative ratings is visual assessment of emphysema extent in chest CT scans. Emphysema is a pathology in chronic obstructive pulmonary disease (COPD), a leading cause of death worldwide [4]. Emphysema is characterized by destruction of lung tissue and entrapment of air. The appearance of emphysema in CT scans can be quite varied and in many cases it is difficult to precisely define where healthy tissue starts and emphysema stops. Current visual scoring systems for assessing emphysema extent are coarse yet still subject to considerable inter-rater variability [2,15]. Emphysema assessment based on visual similarity of lung tissue could improve rater agreement while also improving the granularity of ratings and because it is not limited by current radiological definitions, it could be used to uncover new patterns.

Current practice for visual assessment of emphysema is to consider the full lung volume and decide how much is affected by emphysema [2,15]. Comparing visual similarity of several 3D lung volumes simultaneously could be a difficult and time-consuming task, leading to worse rater agreement compared to assessing each volume by itself. Comparing visual similarity of 2D slices is a much easier task that could even be performed by non-experts with a little instruction. Simplifying the task to this degree opens the possibility of substituting medical experts with crowdworkers, leading to dramatic reductions in time consumption and costs. Crowdsourced image similarities have successfully been used for fine-grained bird classification [12], clustering of food images [14] and more recently as a possibility for assessment of emphysema patterns [6].

There is a growing body of recent work on learning from similarities derived from absolute labels [8,13] illustrating that learning from similarities can be better than learning directly from labels. The triplet learning setting used in these works is for learning from visual image similarity where ratings for a triplet of images (x_i, x_j, x_k) are available in the form of "x_i is more similar to x_j than to x_k".

In this work we also consider similarity triplets derived from absolute labels in the form of expert assessment of emphysema extent. However, our focus is on investigating the feasibility of learning in this setting, with the future goal of learning from actual visual similarity assessment of lung images. We aim to learn descriptive image features, relevant for emphysema severity assessment, directly on the basis of visual similarity triplets. We investigate if CNNs can extract enough relevant information from a single CT slice to learn a disease relevant

representation from similarity triplets. In our previous work on crowdsourcing emphysema similarity triplets [6] we did not learn a feature representation that could be used for unseen images. We believe this work is the first medical image application where similarity triplets has been used to learn a feature representation for embedding unseen images.

2 Materials and Method

In this section we define the triplet learning problem and present a CNN based approach for learning a mapping from input images to a low dimensional representation that reflects the characteristics of the visual similarity measurements.

2.1 The Triplet Learning Problem

Let $\mathbf{X}$ be an image space and $x_i \in \mathbf{X}$ an image. We define a similarity triplet as an ordered triplet of images (x_i, x_j, x_k) such that the ordering satisfies the triplet constraint, given by

$$\delta(x_i, x_j) \leq \delta(x_i, x_k) \tag{1}$$

where δ is some, potentially unknown, measure of dissimilarity. Let $\mathbf{T} \subseteq \mathbf{X}^3$ be a set of ordered triplets that satisfies (1). We want to find a mapping from image space to a low dimensional embedding space, $h^* : \mathbf{X} \to \mathbb{R}^d$, that minimizes the expected number of violated triplets

$$h^* = \arg\min_h \mathbb{E}_{(i,j,k)\in\mathbf{T}} \left[\mathbb{1}\{\tilde{\delta}(h(x_i), h(x_j)) \leq \tilde{\delta}(h(x_i), h(x_k))\} \right]. \tag{2}$$

where $\mathbb{1}$ is the indicator function and $\tilde{\delta} : \mathbb{R}^d \to \mathbb{R}$ is a known dissimilarity.

2.2 Learning a Mapping

End-to-end learning using CNNs is a convenient and powerful method for learning concise representations of images. Optimization of CNNs is based on gradient descent and we cannot optimize (2) directly, because the subgradient is not defined. A commonly used approach is to define a loss function based on how much a triplet is satisfied or violated

$$L((x_i, x_j, x_k)) = \max\{0, \tilde{\delta}(h(x_i), h(x_j)) - \tilde{\delta}(h(x_i), h(x_k) + C)\} \tag{3}$$

where C is a fixed offset used to avoid trivial solutions and encourage oversatisfying triplet constraints. Large violations can dominate the loss (3) and force the optimization to focus on outliers. Since we expect some inconsistencies in the similarity triplets, we consider a variant of (3) that bounds the loss on both sides

$$L((x_i, x_j, x_k)) = \text{clip}_{l,u}(\tilde{\delta}(h(x_i), h(x_j)) - \tilde{\delta}(h(x_i), h(x_k))) \tag{4}$$

where

$$\text{clip}_{l,u}(x) = \begin{cases} 0 & \text{if } x < l \\ 1 & \text{if } x > u \\ \frac{x-l}{u-l} & \text{otherwise} \end{cases} \tag{5}$$

We consider two CNN architecture setups loosely based on VGGnet [9], one with increasing and one with a fixed number of filters in each layer. In both cases a layer is comprised of zeropadding, 3×3 convolution and maxpooling. After the final layer we add a global average pooling layer, and d fully connected units to obtain a d-dimensional embedding of the input. We use squared Euclidean distance as dissimilarity, i.e. $\tilde{\delta} = ||\cdot||_2^2$.

2.3 Data

We use CT scans of 1947 subjects from a national lung cancer screening study [7] with visual assessment of emphysema extent [15] and segmented lung masks. Emphysema is assessed on a six-point extent scale for six regions of the lung: the upper, middle and lower regions of the left and right lung. Here we restrict our attention to the upper right region, defined as the part of the right lung lying above the carina. The six-point extent scale is defined by the intervals {0, 1–5%, 6–25%, 26–50%, 51–75%, 76–100%}. Distribution of emphysema scores is skewed towards 0% with about 73% having 0% and only about 13% having more than 1–5%. Example slices with varying emphysema extent are shown in Fig. 1.

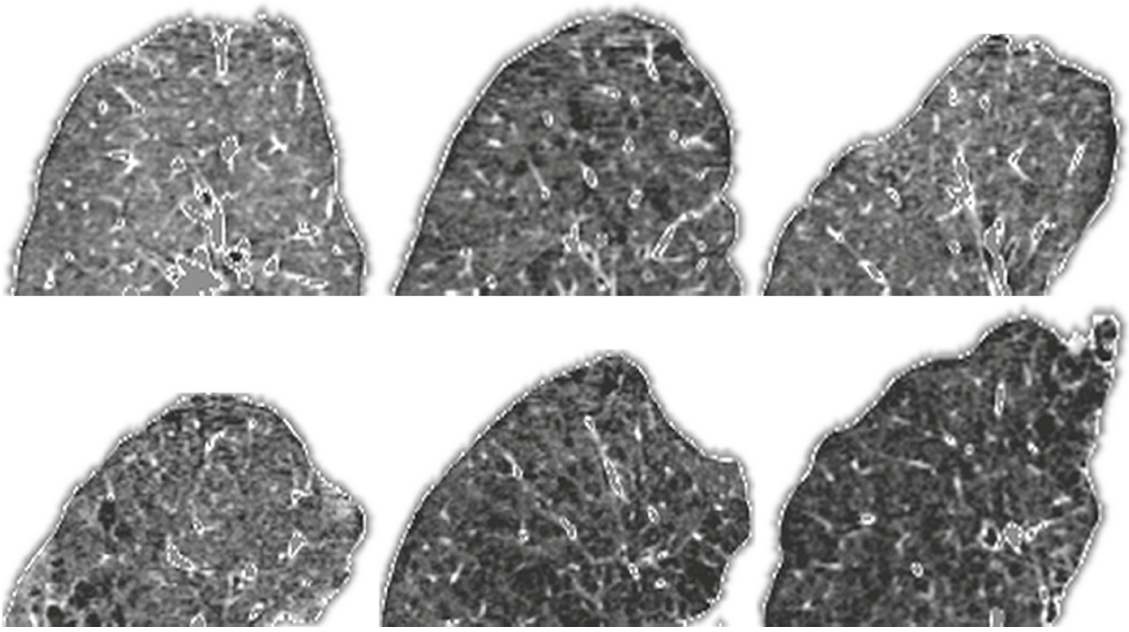

Fig. 1. Example slices. From top left, visually assessed emphysema extent is 0%, 1–5%, 6–25%, 26–50%, 51–75% and 76–100%. Window level −780HU, window width 560HU.

3 Experiments and Results

We split subjects randomly into a training group of 974 subjects and a test group of 973 subjects. For each experiment we then split the training group randomly in half and use one half for training and the other half for validation. Each experiment was run 10 times and we report median statistics calculated over these 10 runs. We use the same clip function for all experiments, with $[l, u] = [-0.01, 0.1]$.

3.1 Preprocessing

A single coronal slice was extracted from the center of the upper right region. Bounding boxes were calculated for the lung mask of each extracted slice in the training data and all images were cropped to the size of the intersection of these bounding boxes (57×125 pixels). Pixels outside the lung mask were set to -800HU to match healthy lung tissue. This aggressive cropping was introduced to avoid background pixels dominating the input data. Finally all pixel intensities were scaled by $\frac{1}{1000}$ resulting in values roughly in the range $[-1, 0]$.

3.2 Selecting Training Triplets

For 974 images there are close to 10^6 possible triplets. Many of these triplets will contain very little information, and choosing the right strategy for selecting which triplets to learn from could result in faster convergence and reduce the required number of triplets needed. When class labels are available they can be used to select triplets as suggested in [8]. However, we are primarily interested in the setting where we do not have class labels. To understand the importance of triplet selection we compare uniform sampling of all possible triplets to sampling based on emphysema extent labels.

When selecting triplets based on emphysema extent we pick the first image uniformly at random from all images, the second uniformly at random from all images with the same emphysema extent as the first, and the third from from all images with different emphysema extent. For the third image we sample images with probability proportional to the absolute difference between the labels.

3.3 Simulating Similarity Assessment

We use visually assessed emphysema extent to simulate similarity assessment of image triplets. For a triplet of images (x_i, x_j, x_k) with emphysema extent labels (y_i, y_j, y_k) the ordering of the triplets satisfies

$$|y_{\sigma(1)} - y_{\sigma(2)}| \leq |y_{\sigma(1)} - y_{\sigma(3)}| \tag{6}$$

$$|y_{\sigma(1)} - y_{\sigma(2)}| \leq |y_{\sigma(2)} - y_{\sigma(3)}| \tag{7}$$

$$|y_{\sigma(1)} - y_{\sigma(3)}| \leq |y_{\sigma(2)} - y_{\sigma(3)}| \tag{8}$$

This corresponds to asking a rater to order images based on similarity.

3.4 CNN Selection

We implemented all CNNs in Keras [1] and used the default Adam optimizer. We searched over networks with $\{3, 4, 5\}$ convolution layers. We used 16 filters for the setup with a fixed number of filters, and 8,16,32,64,128 for the setup with an increasing number of filters. We used a batch size of 15 images and trained the models for 100 epochs or until 10 epochs passed without decrease in triplet violations on the validation set. We then selected the weights with the lowest triplet violations on the validation set. We expect an untrained network with randomly initialized weights will show some degree of class separation and include it as a baseline. Table 1 summarizes median validation triplet violations of the selected models and the median number of epochs used for training. Triplet selection based on emphysema results in somewhat faster convergence and slightly fewer violations compared to uniform triplet selection. The difference in median epochs between uniform triplet selection and extent based triplet selection corresponds to 7500 extra training triplets for uniform selection.

Table 1. Validation set performance. The letter in model type indicates **F**ixed or **I**ncreasing number of filters and the digit indicates number of convolution layers.

Sampling scheme	Model type	Median epochs	Median violations
Untrained	F3	–	46.80 ± 0.94
Uniform	I4	23.0 ± 7.0	40.84 ± 0.71
Extent	F4	18.0 ± 5.0	39.30 ± 0.58

3.5 Triplet Prediction Performance

Selecting Test Triplets. Because we simulate similarity assessments from class labels, the selection of test triplets will have a large influence on the interpretation of performance metrics. In our case about 71% of subjects in the test set do not have emphysema. This implies that selecting triplets uniformly at random results in about 36% of the triplets having no emphysema images. We choose to ignore these same-class triplets when measuring test performance.

In addition to the issue of same-class triplets, we are also faced with a dataset where more than 50% of those subjects that have emphysema only have 1–5% extent. Ignoring this issue will lead to performance metrics dominated by the ability of the network to distinguish subjects with very little emphysema from those without emphysema. This is a difficult task even when given access to the full volume. To more fully understand how well the network embeds images with varying levels of emphysema extent, we calculate test metrics under five different test triplet selection schemes. (1) two images with same extent and one image with different extent, (2) two images without emphysema and one with

emphysema, (3) two images with 0–5% and one with >5%, (4) two images with 0–25% and one with >25%, (5) two images with 0–50% and one with >50%.

Table 2 summarizes the results. As expected we see that the networks are much better at distinguishing between subjects with moderate to severe emphysema versus mild and no emphysema (0–5%), than subjects with emphysema versus subjects with no emphysema (0%). We also see that the untrained network provides decent separation of images with severe emphysema versus moderate to no emphysema (0–50%). In all cases we see that using information about emphysema extent for generating training triplets leads to better performance compared with uniform sampling of triplets.

Table 2. Median triplet violations on test set for the selected models from Table 1 using different schemes for selecting test triplets. See text for explanation of column names.

Sampling scheme	Test triplet selection method				
	All	0%	0–5%	0–25%	0–50%
Uniform	41.0	40.2	30.0	19.0	11.6
Extent	39.3	39.0	26.4	14.6	9.4
Untrained	48.5	48.9	44.3	37.2	29.2

An example embedding of the test set is shown in Fig. 2. We used the models with best performance on the validation set to generate the embedding. Although we see significant overlap between subjects with and without emphysema, both of the trained embeddings have a reasonably pure cluster of subjects with emphysema. There is a clear tendency towards learning a one dimensional embedding. We hypothesize that several factors contribute to this tendency, (1) clipping at $[-0.01, 0.1]$ encourages small distances, (2) pairwise distances for uniformly distributed points increase as the dimensionality is increased, (3) the underlying

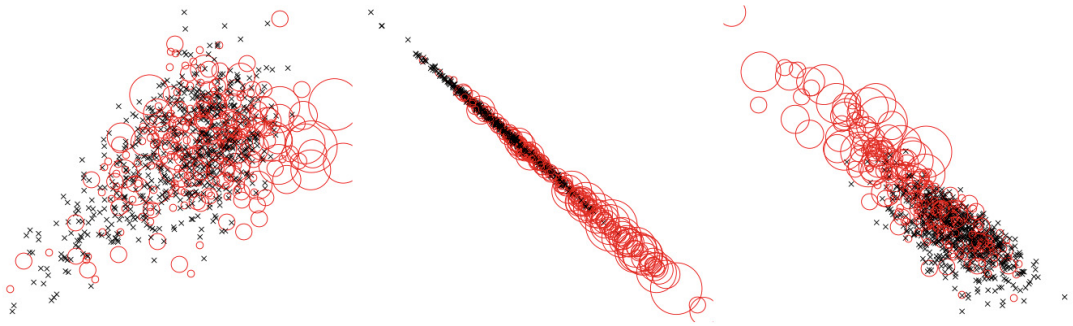

Fig. 2. Example embedding of test data. Black crosses are subjects without emphysema, red circles are subjects with emphysema. Size of circle correspond to emphysema extent. From left: Untrained (48.3% testset violations), uniform (39.5% testset violations), visual (38.8% testset violations). (Color figure online)

dissimilarity space, emphysema extent, is one dimensional and all triplets can in principle be satisfied by embedding unto the real line.

4 Discussion and Conclusion

We formulated assessment of emphysema extent as a visual similarity task and presented an approach for learning an emphysema relevant feature representation from similarity triplets using CNNs. We derived similarity triplets from visual assessment and investigated the importance of selecting informative triplets.

It is slightly surprising that a single cropped 2D slice contains enough information for the level of separation illustrated by the embeddings in Fig. 2. This shows that learning can be accomplished from simple annotation tasks. However, there are likely instances where the particular slice is not representative for the image as a whole, which may explain why there is a large overlap between subjects with and without emphysema in Fig. 2. We suspect that with triplet similarities based on individual slice comparisons, class overlap would be less.

As a proof of concept, in this work we simulated slice similarity assessment from experts' emphysema extent scores. Potentially such triplets could be gathered online via crowdsourcing platforms such as Amazon Mechanical Turk. Our previous results [6] showed that crowdsourced triplets could be used to classify the emphysema type (rather than extent) with a better than random performance. Preliminary results indicate that the crowdsourced triplets are too few or too noisy for training the proposed CNNs. However, we expect that improving the quality and increasing the quantity of crowdsourced triplets will allow CNNs to learn an emphysema sensitive embedding without needing expert assessed emphysema extent for training.

We investigated the importance of triplet selection and found that performance improved slightly when selecting triplets based on emphysema extent, in particularly for subjects with moderate emphysema extent (columns 0–5% and 0–25% in Table 2). While using disease class labels to select triplets is not a viable solution, for medical images we often have access to relevant clinical information that could be used to select triplets. In the context of emphysema, measures of pulmonary function are potential candidates for triplet selection. However, our preliminary results indicate that using pulmonary function measures for triplet selection is not straightforward and can harm performance compared to uniform triplet selection.

We assumed that there is a single definition of visual similarity between the slices. However, this does not have to hold in general. For emphysema it is relevant to consider both pattern and extent as measures of similarity. The idea of having multiple notions of similarity is explored in [10], where different subspaces of the learned embedding corresponds to different notions of similarity. Simultaneously modeling multiple notions of similarity could lead to more expressive feature representations. Additionally, it be useful when learning from crowdsourced triplets, where some raters might focus on irrelevant aspects, such as size and shape of the lung.

In conclusion, we have shown that CNNs can learn an informative representation of emphysema based on similarity triplets. We believe this to be a promising direction for learning from relative ratings, which may be more reliable and intuitive to do, and therefore could allow the collection of large data sets that CNNs benefit from. The next step is to explore embeddings resulting from directly annotated similarity triplets. We expect such embeddings to show different notions of similarity and it will be interesting to see how these notions compare to current radiological definitions.

References

1. Chollet, F., et al.: Keras (2015). https://keras.io
2. COPDGene CT Workshop Group, Graham Barr, R., et al.: A combined pulmonary-radiology workshop for visual evaluation of COPD: study design, chest CT findings and concordance with quantitative evaluation. COPD J. Chronic Obstr. Pulm. Dis. **9**(2), 151–159 (2012)
3. Goffin, R.D., Olson, J.M.: Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. Perspect. Psychol. Sci. **6**(1), 48–60 (2011)
4. Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) (2017)
5. Jones, I., Wheadon, C.: Peer assessment using comparative and absolute judgement. Stud. Educ. Eval. **47**, 93–101 (2015)
6. Ørting, S.N., Cheplygina, V., Petersen, J., Thomsen, L.H., Wille, M.M.W., de Bruijne, M.: Crowdsourced emphysema assessment. In: Cardoso, M.J., et al. (eds.) LABELS/CVII/STENT -2017. LNCS, vol. 10552, pp. 126–135. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67534-3_14
7. Pedersen, J.H., et al.: The Danish randomized lung cancer CT screening trial-overall design and results of the prevalence round. J. Thorac. Oncol. **4**(5), 608–614 (2009)
8. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint: arXiv:1409.1556
10. Veit, A., Belongie, S., Karaletsos, T.: Conditional similarity networks. In: Computer Vision and Pattern Recognition (CVPR 2017) (2017)
11. Wagner, S.H., Goffin, R.D.: Differences in accuracy of absolute and comparative performance appraisal methods. Organ. Behav. Hum. Decis. Process. **70**(2), 95–103 (1997)
12. Wah, C., Van Horn, G., Branson, S., Maji, S., Perona, P., Belongie, S.: Similarity comparisons for interactive fine-grained categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 859–866 (2014)
13. Wang, J., et al.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386–1393 (2014)

14. Wilber, M.J., Kwak, I.S., Belongie, S.J.: Cost-effective hits for relative similarity comparisons. In: Second AAAI Conference on Human Computation and Crowdsourcing (2014)
15. Wille, M.M., Thomsen, L.H., Dirksen, A., Petersen, J., Pedersen, J.H., Shaker, S.B.: Emphysema progression is visually detectable in low-dose CT in continuous but not in former smokers. Eur. Radiol. **24**(11), 2692–2699 (2014)

Feature learning using visual similarity triplets for lung texture analysis

Silas Nyboe Ørting, Jens Petersen, Veronika Cheplygina, Laura H. Thomsen, Mathilde M. W. Wille and Marleen de Bruijne

***Abstract*—For many medical image analysis tasks state-of-the-art methods are based on supervised machine learning. Obtaining expert labels required for this is costly and generally under-utilizes experts knowledge by reducing expert knowledge about anatomy, imaging, and disease patterns to a single label. Relative comparisons of images, focusing on visual similarity of image content, could increase the information extracted from experts and allow more general representations to be learned. Additionally, since assessing visual similarity does not require classification of patterns, it is likely that non-experts can perform the task allowing crowdsourcing to be used as a means of reducing labeling costs. We present a study of how visual similarity can be assessed and used to learn representations useful for detecting lung abnormalities. We focus on visual similarity in the form of similarity triplets, "image a is more similar to image b than to image c", and obtain 60,000 visual similarity triplets for 300 images. Inter and intra-rater agreement for these assessments are 58% and 65%. We use the similarity triplets to learn a low-dimensional embedding using a convolutional neural network. We propose a model to handle multiple notions of similarity and show that representations learned with this model results in less correlated features. Decorrelating features could help learn a representation where feature dimensions can be mapped to semantic concepts. Using a single slice from each scan we use the learned representations to fit logistic regression models for detecting visually assessed emphysema and ILD patterns. We obtain AUC of 0.74 (0.54 using random features) for emphysema detection and 0.58 (0.53 using random features) for ILD detection.**

I. Introduction

SUPERVISED machine learning, and in particular deep learning, has greatly advanced medical image analysis [1], [2], [3]. One of the main challenges when applying supervised machine learning in medical image analysis is that labels are costly to obtain due to the need for expert knowledge.

Crowdsourcing, where groups of non-experts perform the labeling tasks, is an increasingly popular solution to reducing the cost of labeling [4]. Although many labeling tasks are amenable to crowdsourcing, many medical image analysis tasks require a certain level of expertise. For example, assessing severity of emphysema requires knowledge, not only about the appearance of emphysema in CT, but also about the appearance variations of healthy lung tissue and other pathologies. It can therefore be necessary to adapt the labeling task to ensure non-experts can perform it quickly and accurately.

There is a discrepancy between how experts label images, and the kind of labels used by machine learning algorithms. Machine learning solutions generally focus on a very specific problem, e.g. segment a tumor, count micro-bleeds, or estimate severity of emphysema. Labels are then obtained to match the specific problem. However, when making an assessment a radiologist will not just segment a tumor, count micro-bleeds or estimate severity of emphysema. Rather, the radiologist will also consider the entire image and how it compares to what is to be expected. Some of this information will make its way into a radiology report, for example as incidental findings, some will stay in the head of the radiologist. Reducing all this information to a single label under-utilizes the radiologist's expertise, thereby increasing the cost of labeling images.

In this work we focus on learning from visual similarity comparisons as an approach to improved utilization of experts knowledge and reduced labeling costs. By focusing on visual similarity, we hope to learn representations that yield a more complete characterization of image content compared to representations optimized for a single task.

Additionally, learning from visual similarity could reduce the need for expert knowledge. Distinguishing pathologies with similar appearance is a challenging task requiring expertise and, in many cases, context not directly available in the image. However, when characterizing the visual appearance of an image, it is not necessary to make this distinction. Indeed, if two pathologies are visually similar it is reasonable that the learned representation of them is also similar. Accepting this limitation, it is reasonable to believe non-experts can assess visual similarity about as well as experts.

The idea of learning from visual similarity comparisons has been explored in computer vision applications for tasks such as clustering [5], learning image similarity metrics [6] and object categorization [7]. Closely related to learning from visual similarity comparisons is learning from similarity measurements derived from image labels, e.g. two images are similar if they share a label. This idea was explored in [8] where the Siamese network architecture was proposed for face verification. More recently, [9] explored how learning from

This study was financially supported by the Danish Council for Independent Research (DFF) and the Netherlands Organization for Scientific Research (NWO). The sponsors had no involvement in the work.

S. Ørting and J. Petersen are with Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

V. Cheplygina is with Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

L. Thomsen is with Department of Internal Medicine, Hvidovre Hospital, Copenhagen Denmark

M. M. W. Wille is with Department of Diagnostic Imaging, Bispebjerg Hospital, Copenhagen, Denmark

M. de Bruijne is with Department of Computer Science, University of Copenhagen, Copenhagen, Denmark and Biomedical Imaging Group Rotterdam, Departments of Radiology and Medical Informatics, Erasmus MC - University Medical Center Rotterdam, The Netherlands

Manuscript received NN

derived similarity triplets, "a is more similar to b than to c", improved facial recognition over learning directly from image labels. Learning from similarity triplets was further explored in [10] where a conditional triplet network was proposed to learn from similarity triplets when different notions of similarity are used, e.g. color for some triplets, size for others.

The closest to our work is probably [11] where visual similarity assessments of endomicroscopy videos are used to improve a video retrieval system based on SIFT features. However, learning from visual similarity remains under-explored in image analysis, especially for end-to-end learning.

An alternative approach to reducing the cost of labeling is weakly supervised learning where models are trained on image level labels. For example, learning from lung function measurements [12] or presence of emphysema and interstitial lung disease [13]. Convolutional neural networks (CNNs) has proved useful in this setting. For example, learning from emphysema presence [14]. However, learning from weak labels does not directly address the challenge of extracting more information than what can be summarized in a single label.

The aim of this study is to explore how visual similarity can be used to learn a low-dimensional representation of medical images. We believe the idea of learning from visual similarity is generally applicable where texture and other visual characteristics are of primary concern, and we here provide a study of chest CT scans as an illustrative example.

This work extends our previous work in [15] where we investigated crowdsourcing of visual similarity comparisons for emphysema assessment and in [16] where we investigated how well different convolutional neural networks (CNNs) could learn from similarity measurements that were derived from expert visual assessment of individual scans.

In this work we show how visual similarity triplets can be gathered and analyze how well raters agree on visual similarity assessments. In contrast to [15], where we used a small stratified set of images with maximally different emphysema patterns, we here consider a more realistic setting by sampling a larger set of images uniformly at random from all CT scans acquired in a lung cancer screening trial. We use the visual similarity triplets to learn a low-dimensional representation using a CNN and show that the learned representation is useful for detecting both emphysema and interstitial lung disease abnormalities. We propose an iterative approach for learning when multiple notions of similarity are used by the annotators, and investigate how this model of multiple similarity notions influences the learned embeddings.

II. Materials and methods

A. Learning from similarity triplets

Consider three images a, b, c. Assume we know that image a is more similar to image b than to image c. Using a suitable distance function δ we can express this as

$$\Delta(a, b, c) = \delta(a, b) - \delta(a, c) < 0, \tag{1}$$

which we refer to as a triplet constraint. We represent such a triplet constraint as the ordered triplet (a, b, c). We wish to find a low-dimensional embedding of the images a, b, c that satisfies the triplet constraint.

Let $\mathcal{F}$ be a class of functions that maps from image space to some embedding space E and δ a distance function on E. We want to find a function $f^* \in \mathcal{F}$ that satisfies the triplet constraint from Equation 1

$$f^* \in \{f \,|\, \Delta_f(a, b, c) < 0\} \tag{2}$$

where $\Delta_f(a, b, c) = \Delta(f(a), f(b), f(c))$.

For a set of images I and a set of triplets $T \subseteq I^3$ we are interested in finding a function that satisfies all triplet constraints in T

$$f^* \in \{f \,|\, \Delta_f(a, b, c) < 0 \,.\, \forall (a, b, c) \in T\}. \tag{3}$$

In general we cannot guarantee that there exists such an f. Instead we consider the problem

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{(a,b,c) \in T} \max\{0, h - \Delta_f(a, b, c)\} \tag{4}$$

that minimizes the number of violated triplet constraints. The $\max$ operation ensures that easily satisfied triplets will not dominate the optimization, and $h > 0$ ensures that the trivial solution of mapping everything to the same point, $\Delta_f(a, b, c) = 0 \,.\, \forall (a, b, c)$, is sub-optimal.

B. Similarity triplet networks

In this work we restrict the class of functions $\mathcal{F}$ to a small set of convolutional neural networks (CNNs). We consider relatively shallow and simple CNN models, because our previous work [16] indicates that these models have enough capacity to learn lung texture representations. We consider CNNs comprised of three convolution blocks, each comprised of: 1px zero padding, one layer c convolution kernels of size 3×3, a rectified linear activation function and 2×2 max pooling. The number of convolution kernels c in the three convolution blocks is respectively set to 8, 16 and 32. After the final convolution block we use k 1×1 convolution kernels with hyperbolic tangent activations, followed by masked global average pooling to obtain average feature responses within the lung mask. The last two steps results in an embedding lying in the k-dimensional unit cube.

1) Masked global average pooling: Let x be an image and m a binary mask indicating which part of x should be pooled. Masked global average pooling of x is then

$$\frac{\sum_i x_i m_i}{\sum_i m_i} \tag{5}$$

C. Multiple notions of similarity

When assessing relative visual similarity of images, a rater uses some notion of similarity. For a given set of similarity triplets, it is very likely that the similarity notion varies between triplets. Consider the example in Fig. 1, where the similarity notion on the left could be "amount of bullae" and the similarity notion on the right could be "overall intensity".

Using the same embedding to represent both cases could be problematic. Which notion of similarity is used depends on

the context, so the same image could occur in triplets using different notions of similarity, resulting in triplet constraints that cannot be satisfied simultaneously in the same embedding. This problem is investigated in [10] where it is shown that it is beneficial to force different parts of the embedding space to correspond to different notions of similarity. In [10] triplets are derived from image labels and it is thus known which similarity notion is used for each triplet. We do not have that knowledge. Instead we propose to learn which similarity notion is used for each triplet.

We propose an EM like approach where we interchange between optimizing the embedding given an assignment of similarity notions and optimizing the assignment of similarity notions given an embedding. A similarity notion is assigned to each triplet and not constrained to be the same for triplets with the same anchor image.

We represent a similarity notion as a weight vector w that encodes which dimensions of the embedding are relevant for a specific triplet. The triplet constraint from Equation 2 then becomes

$$\Delta_f(a, b, c; w) = \Delta_f(w \circ f(a), w \circ f(b), w \circ f(c)) \quad (6)$$

where $\circ$ is the Hadamard, or component-wise, product.

For a given triplet we will have that some features are relevant, whereas others should be ignored. In order to achieve this, we restrict w to be a binary vector that selects which dimensions are relevant. For a triplet (a, b, c) we find the w that minimizes Equation 6 by considering each component of w separately

$$w_i = \begin{cases} 0 & \text{if } \Delta_f(a, b, c; e_i) \geq 0 \\ 1 & \text{if } \Delta_f(a, b, c; e_i) < 0 \end{cases} \quad (7)$$

where e_i is the i'th unit vector. This corresponds to masking out dimensions where the triplet is not satisfied. In the case where the triplet cannot be satisfied in any dimension ($\forall i \,.\, w_i = 0$), we set all $w_i = 1$ for that triplet. By allowing more than one $w_i = 1$, we ensure that complex similarities requiring multiple dimensions can still be modeled.

The suggested approach for assigning similarity notions is only possible when the triplet constraint is known, and can thus only be used during training. Regardless of this limitation, we hypothesize that using multiple similarity notions can help the network learn embeddings where dimensions correspond to different features of interest.

D. Data

We used low-dose chest CT scans from the Danish Lung Cancer Screening Study (DLCST) [17]. We randomly selected 300 baseline scans from DLCST and split these 300 scans into three non-overlapping datasets (D_1, D_2, D_3) with 100 scans in each.

In Section III-B3 we conduct experiments to estimate how much information about lung abnormalities is contained in the learned representations. For these experiments we use assessments of lung abnormalities from [18] where all DLCST scans where visually assessed by medical experts. We used assessment from a single rater and included assessment of emphysema extent on a 6-point scale in the upper right region; and assessment of three interstitial lung disease (ILD) patterns: honeycombing, ground glass opacities and reticulation. Due to low prevalence, we considered all three ILD patterns jointly and refer to this as ILD.

TABLE I
DISTRIBUTION OF VISUALLY ASSESSED ILD PATTERNS AND EMPHYSEMA EXTENT IN THE UPPER RIGHT REGION.. COLUMNS 0-6 CORRESPOND TO 0% IN SCAN, 0% IN REGION, 1-5%, 6-25%, 26-50%, 51-75%, 76-100%

Dataset	0	1	2	3	4	5	6	ILD present
D_1	70	4	12	9	4	1	0	17
D_2	66	3	18	7	5	0	1	12
D_3	69	5	11	12	3	0	0	6

The distribution of emphysema extent and prevalence of ILDs is summarized per dataset in Table I. To asses similarity between emphysema extent distributions of the three datasets we computed the Earth Mover's Distance between distributions. We find $EMD(D_1, D_3) = 5$, $EMD(D_1, D_2) = 12$ and $EMD(D_2, D_3) = 15$. In the sense of emphysema extent, D_1 and D_3 are more similar to each other than to D_2, and both are about equally similar to D_2. For ILDs we see a difference in 5% for D_1, D_2, 11% for D_1, D_3 and 6% for D_2, D_3. In the sense of ILD presence, D_1 and D_3 are each more similar to D_2 than to each other.

1) Preprocessing: We used previously segmented lung masks from [19] to extract the region above the carina from the right lung (right upper region). We then extracted three coronal slices from the upper right region of each scan. We chose the center slice, and the two neighboring slices spaced 1cm away from the center. We masked all slices to only include the lung fields and set areas outside the lung mask to -800HU ($\approx$ density of healthy lung tissue). Finally, to avoid scaling issues when fitting models, we scaled intensity values by $1/1000$ to obtain pixel values approximately in the range $[-1, 0]$. Additionally, we extracted a single slice covering the full lung field of both lungs and processed it in the same manner. This full lung field slice was used for testing how well the learned representation can predict ILD presence, for which no regional visual assessments were available.

2) Obtaining visual similarity triplets: We used only the center slice from the upper right region when assessing visual similarity.

Visual similarity was assessed by looking at a 3×3 grid of images and choosing the two images that are most similar to the center images, see Fig. 2 for an example. We refer to such an image grid as query and to the center image as the anchor image. We obtained 12 similarity triplets from each query (six similarity triplets for each of the two selected images). This approach of choosing multiple images from a grid was shown to be effective in [20] where they found that the increased annotation quantity outweighed the decrease in annotation quality. In out study, we have a skewed dataset with mostly no pathology and we are primarily interested in images with pathology. We believe that a large set of annotations with more pathology images (in absolute numbers), is more beneficial than a smaller set of higher quality annotations.

For each image in a dataset we generated 50 queries with

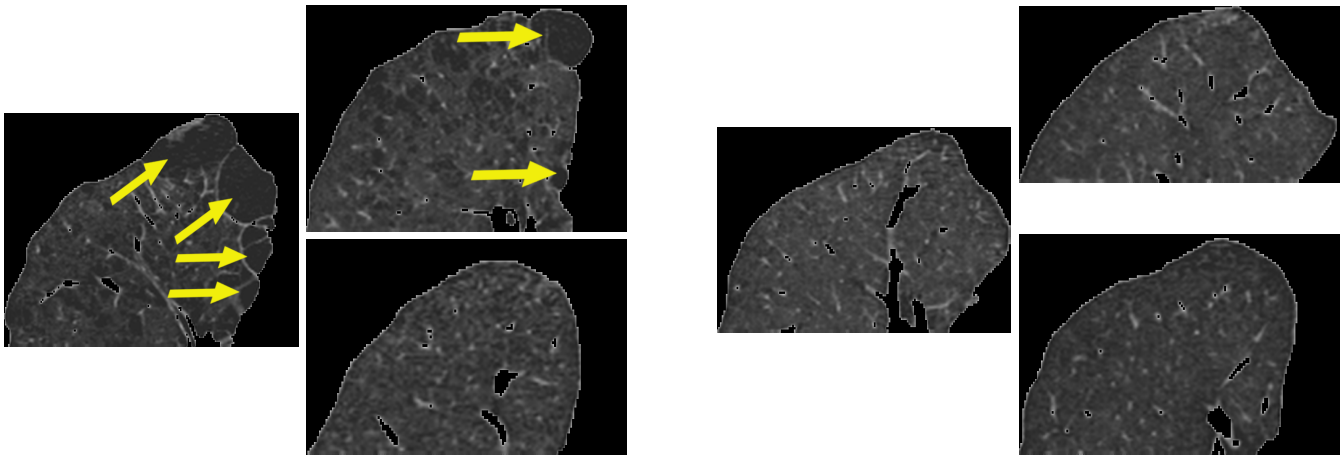

Fig. 1. Multiple notions of similarity. **Left:** Amount of bullae (indicated with yellow arrows) as similarity notion. **Right:** Overall intensity as similarity notion

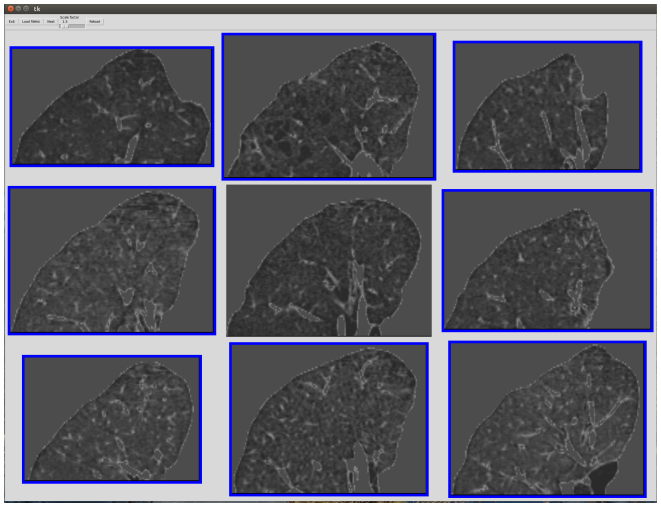

Fig. 2. Annotation UI

that image as anchor. The eight other images in each query where chosen uniformly at random among the remaining 99 images in the dataset. This resulted in 5.000 queries per dataset, with no query containing images from multiple datasets. All 15.000 queries where then randomly shuffled and split into batches of 100 queries, such that each batch contained queries from all datasets.

A single rater (SØ) annotated all batches. Additionally, two raters (JP and MdB), annotated the same four batches to assess inter-rater variability. The same four batches where also annotated twice by SØ, with at least 2 weeks between, to asses intra-rater variability. All three raters have 3+ years experience analyzing emphysema in chest CT images from a computer science perspective, but have no clinical training. Instructions for annotation where informal and general, along the lines of "focus on differences in lung tissue texture".

E. Measuring rater agreement

We measure how well raters agree on the two images selected as most similar to the anchor image. In this paper there are three possibilities when measuring pairwise agreement (1) raters agree on two images (2) raters agree on one image and (3) raters agree on no images. For a set of queries Q we define agreement between two raters, r_1 and r_2, that select k images as

$$\cap(r_1, r_2|Q) = \frac{1}{|Q|} \sum_{q \in Q} \frac{|r_1(q) \cap r_2(q)|}{k}. \tag{8}$$

There will be some agreement due to chance. The size of chance agreement depends on the size of the query. There are $\binom{8}{2} = 28$ ways of picking two images among eight candidates. Let $\{a, b\}$ and $\{c, d\}$ be two sets of two images picked uniformly at random from eight candidates. The expected agreement is then

$$\begin{aligned} & E\left[\frac{1}{2}\left|\{a,b\} \cap \{c,d\}\right|\right] \\ &= E\left[\frac{1}{2}\left(\left|\{a\} \cap \{c,d\}\right| + \left|\{b\} \cap \{c,d\}\right|\right)\right] \\ &= E\left[\left|\{a\} \cap \{c,d\}\right|\right] \\ &= E\left[\mathbb{1}\{a \in \{c,d\}\}\right] \\ &= P\left(a \in \{c,d\}\right) \\ &= 1 - P\left(a \notin \{c,d\}\right) \\ &= 1 - \frac{\binom{7}{2}}{\binom{8}{2}} \\ &= 0.25 \end{aligned} \tag{9}$$

III. Experiments and Results

We conducted two sets of experiments. First we analyzed the triplet annotations by measuring rater agreement and estimating how consistent the triplets are. then we investigated how and how well a network can learn a low-dimensional representation of images from the triplets.

A. Analysis of annotations

1) Annotation variability: Four batches of 100 randomly selected queries where annotated by all three raters to assess inter-rater variability. Annotating a single batch took between

TABLE II
RATER AGREEMENT IN PERCENT. FIRST ROW IS INTRA-RATER AGREEMENT. COLUMNS SHOW OVERALL AGREEMENT AND STRATIFIED BY EMPHYSEMA EXTENT AND ILD PRESENCE IN ANCHOR IMAGE.

Raters	Overall	1	2	3	4	5	6	ILD
SØ1/SØ2	65	62	66	75	83	69	100	71
SØ/JP	61	59	68	60	68	74	100	61
SØ/MdB	55	54	49	59	66	65	71	57
JP/MdB	55	54	56	61	56	50	71	55

10 and 30 minutes. The same four batches where annotated twice by the primary rater, with at least two weeks in between, to assess intra-rater agreement. Agreement was measured as described in Section II-E and is summarized in Table II. Columns 1-6 show agreement stratified by visually assessed emphysema extent in the anchor image and column ILD show agreement for anchor images with ILD presence. Agreement between the primary rater (P) and the secondary raters (S1,S2) is averaged across the two sets of annotations from the primary rater. Overall agreement is substantially better than random (25%). Overall intra-rater agreement is better than inter-rater agreement. Agreement between P and S1 is close to intra-rater agreement and better than P/S2 an S1/S2. There appears to be a tendency of higher agreement for larger emphysema extent in anchor image. Intra-rater agreement is higher for ILD than overall, whereas inter-rater agreement on ILD is the same as overall inter-rater agreement.

2) Annotation consistency: We have no guarantee that all triplet annotations can be satisfied. Triplet constraints induce a partial ordering and can be represented as a directed graph. If this graph contains any cycles the constraints are inconsistent. Although detecting cycles is simple, deciding which edges to prune is not. McFee and Lanckriet [21] view inconsistent triplets as label noise and aims at retaining as much information as possible, however this is equivalent to the NP-Complete maximum acyclic subgraph problem. Although good approximate solutions exist, we take a different view. Inconsistent triplets arise because images are very similar or because different notions of similarity are used. In both cases, the inconsistencies are valuable and should not be discarded.

We measure the scale of inconsistencies by estimating how many triplets can be satisfied when the embedding is not constrained by image data. This also provides an approximate upper bound on the proportion of triplets that can be satisfied by a triplet network using a single notion of similarity.

Based on our experience annotating the data we believe that 5-10 dimensions are sufficient for capturing the different similarity notions. We used an 8-dimensional embedding space for the consistency experiments.

We used t-distributed stochastic triplet embedding (t-STE) [22] to embed each of the three datasets. Consistency was measured as the percentage of violated triplet constraints. The experiment was run 1000 times with different random initializations of t-STE. The median percentage of violated triplets was $10.1\%, 10.9\%, 9.6\%$ for D_1, D_2, D_3, with median absolute deviation less than 0.07%. Although there is some difference in violated triplets (1.3% beween D_2, D_3), it is not clear if this difference is large enough too be of importance. For 60000 triplets, a difference of 1.3% corresponds to 780 triplets.

TABLE III
CROSS VALIDATION FOLDS USING THE THREE DISJOINT DATASETS D_1, D_2, D_3.

Fold	Train	Validation	Test
123	D_1	D_2	D_3
213	D_2	D_1	D_3
132	D_1	D_3	D_2
312	D_3	D_1	D_2
231	D_2	D_3	D_1
321	D_3	D_2	D_1

B. Triplet network performance

We compare two approaches for training networks. In the first approach we train the networks using multiple similarity notions as described in Section II-C. We stop training after 10 iterations. In the second approach we follow the same procedure with the exception that all triplets use the same similarity weight ($w = \mathbf{1}$) in all iterations. All experiments are carried out with both approaches.

We use a 3×2-fold cross validation scheme for all experiments where we train a network. The scheme is summarized in Table III. We run all experiments 10 times for each fold, we refer to each run as a "replication". Parameters for the network are kept fixed throughout all experiments.

The networks are implemented in Keras [23] using the Tensorflow backend. We use the ADAM optimizer with default parameters. We use a batch size of 30 with 2000 steps per epoch for at most 100 epochs. We use early stopping by stopping optimization after 10 epochs without reduction in violations on the validation set.

Although similarity triplets where only collected for the center slice, we assume the same triplet holds for neighboring slices. Based on this assumption we use the two neighboring slices for data augmentation during training, by randomly selecting one of the three slices in each training epoch. We do not use augmentation for the validation and test sets. We used a variety of GPUs for the experiments with training times around 1-2 hours for the first iteration and 6-8 hours for all ten iterations.

1) Stability: The purpose is to investigate how stable model fitting is with respect to initialization and training data.

For each fold we measure proportion of violations on the test set using the networks from the first training iteration. Note that the training procedure is exactly the same for both training approaches in the first iteration, since all initial weights are set to 1 for all triplets. We therefor report results jointly.

The results are summarized in Fig. 3. With a few exceptions, test set violations is below 35% with the best cases being close to 25%. This clearly indicates that the networks learn a useful representation for predicting visual similarity.

There is clearly large variation across folds. The folds trained on D_2 (213,231) have some replications that fail to learn anything useful. The best, and very similar, performance

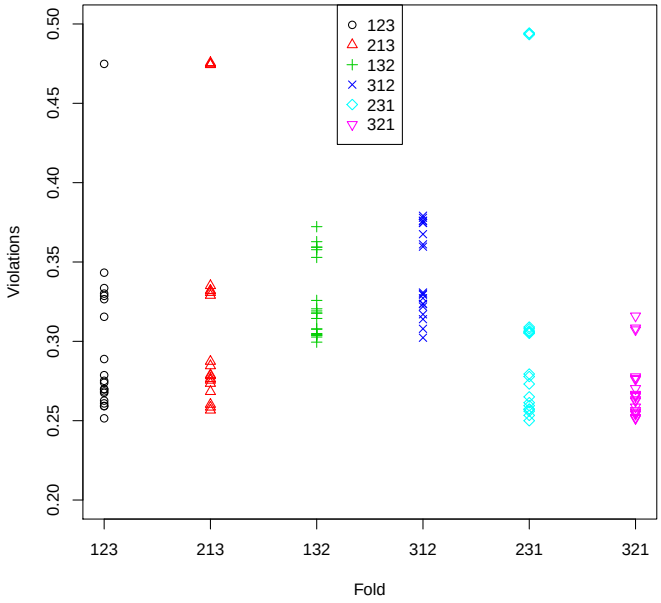

Fig. 3. Proportion of test set violations across folds.

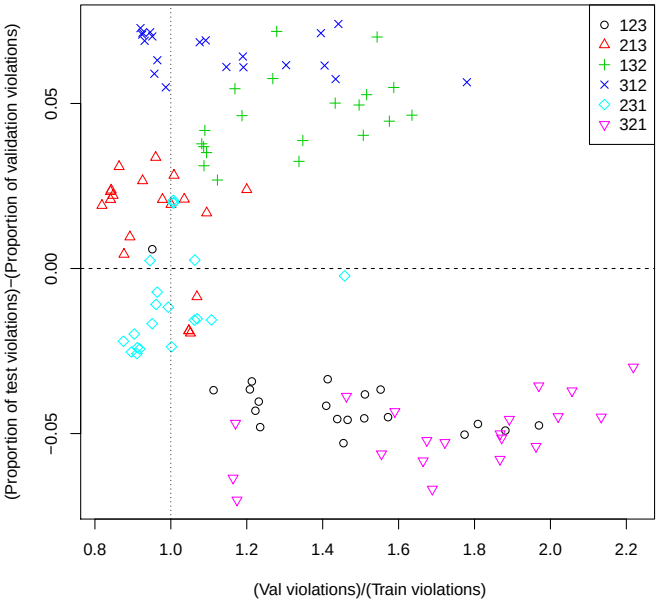

Fig. 4. Difference in violations on test and validations versus validation generalization error. Points below the dashed line have lower test violations than validation violations. Points to the left of the dotted line have lower validation violations than training violations.

is on folds 123 and 321. If we take the difference in emphysema distribution, reported in Table II, into account, this performance difference appears to correlate with the similarity of emphysema extent distribution between train and test data.

The difference in test and validation violations is plotted against the validation set generalization error in Fig. 4. The maximum difference in proportion of test and validation violations is never larger than 7.4% indicating that poorly performing networks, e.g. test violations $> 45\%$, can be avoided by discarding networks with large validation violations.

There appears to be three clusters mostly containing two folds each: (123,321) where validation violations overestimate test violations, (132,312) where validation violations underestimate test violations, (213,231) where generalization error is less than one and validation violations are close to test violations. This again indicates that D_2 is dissimilar from D_1 and D_3.

2) Training with multiple similarity notions: The purpose is to investigate if using multiple similarity notions during training decreases the proportion of violated test triplets.

For each fold we measure the proportion of test set violations using the networks from each iteration. The distribution of test set violations for all folds and replications is summarized in Fig. 5. Using multiple iterations is generally beneficial. When training with multiple similarity notions there is a clear reduction in variation and the median proportion of violations decreases steadily from around 31% to 28%.. When training without multiple similarity notions there is a decrease in variation and median proportion of violations, from around 30% to 27%, the first four iterations. After that it appears to have no effect.

When training with multiple similarity notions we encourage the network to find embeddings where triplets are satisfied in at least one dimension. To investigate the effect of this we measure how well the learned embeddings satisfy triplets in at

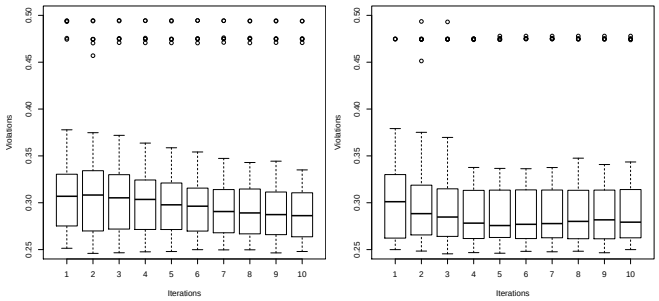

Fig. 5. Test set violations. Left: Training with multiple similarity notions. Right: Training without multiple similarity notions.

least one dimensions, by assigning weights to test triplets as described in Section II-C. The distribution of test set violations for all folds and replications is summarized in Fig. 6. Training with multiple similarity notions clearly reduces the amount of test triplets that cannot be satisfied in any dimensions. Whereas the results is about the same across iterations when training without.

To further investigate the effect of training with multiple similarity notions we measure how the correlation between samples changes between iterations. By masking out dimensions i for some triplets, we reduce the number of constraints that are active for dimensions i. In the extreme case, where dimension i is masked for all triplets containing a specific image a, the representation of a is unconstrained in dimension i. The representation of a in i will not be random, because constraints from other triplets force different dimensions to be correlated, we do however expect the correlation to decrease. We measure change in correlation between two iterations as the average change in absolute correlation between all pairs of dimensions. Let ρ^n, ρ^m be the correlation matrices at iteration

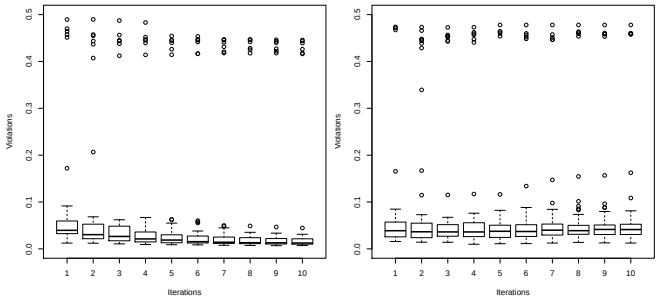

Fig. 6. Test set violations when triplets need only be satisfied in a single dimension. Left: Training with multiple similarity notions. Right: Training without multiple similarity notions.

TABLE IV
CHANGE IN CORRELATION FROM ITERATION 1 TO ITERATION 10. AGGREGATED ACROSS REPLICATIONS AND FOLDS. COLUMNS WITH $\Delta < -x$ SHOW THE PERCENTAGE OF REPLICATIONS WHERE DECREASE IN CORRELATION IS MORE THAN x.

Experiment	Median Δ	$\Delta < 0$	$\Delta < -0.1$	$\Delta < -0.2$
Without	-0.009	61.7%	15.0%	5.0%
With	-0.108	85.0%	53.3%	13.3%

n and m. Using a k-dimensional embedding, the change in correlation from iteration n to iteration m is then

$$\frac{2}{k(k-1)} \sum_i \sum_{j>i} |\rho^m_{i,j}| - |\rho^n_{i,j}| \tag{10}$$

The change in correlation from iteration 1 to iteration 10 is summarized in Table IV. Although correlation is reduced both when training with and without multiple similarity notions, the decrease is largest and most consistent when training with multiple similarity notions, indicating that the similarity weights pushes the embedding to decorrelate dimensions.

3) Predictive power of embeddings: The purpose is to investigate how well the learned embeddings can predict visually assessed lung pathologies.

We use the networks trained in Section III-B2 to embed center slices from the upper right region and from the full lung field. For each fold, we then fit a regularized logistic regression model on the training splits to predict 1) visually assessed emphysema presence in the upper right region and 2) visually assessed ILD presence in the scan.

We use a weighted cost function to handle class imbalance. The cost for predicting absence is set to 1 and the cost of predicting presence is set to $2 \times (1 - \text{prevalence})$. The regularization parameter of the logistic regression model is optimized over ($C \in \{2^i | i = 0, 1 \ldots, 9)$ on the validation splits.

The results are summarized in Table V with the last row showing performance when using random features sampled from an 8-dimensional normal distribution. Although random features are better than random guessing, performance of the trained networks is clearly better. Overall, training without multiple similarity notions is slightly better than training with.

Predicting presence of emphysema at iteration 1 yields AUC of around .70 (.70 - .74 at iteration 10). Predicting emphysema extent larger than 5% is substantially easier with AUC around .83 (.84 - .86 at iteration 10). AUCs only increase slightly when predicting more than 25% extent, indicating that the main challenge is in distinguishing 0% and 1-5% emphysema extent. Presence of the three ILD pathologies is overall lower than emphysema (11.6% versus 31.6%) with more variation between datasets (6%, 12%, 17% presence). Average AUC for predicting ILD presence is .58 at both iteration 1 and 10. A possible explanation for the relatively poor detection of ILD patterns could be that the patterns are rarely visible in the slices extracted from the upper right region, resulting in very little ILD signal during optimization. However, these results indicate that the networks can learn features related to both emphysema and ILDs from visual similarity triplets.

TABLE V
AVERAGE ROC AUC ($\times 100$) FOR PREDICTING LUNG PATHOLOGIES.

		Emphysema			
Experiment	Iteration	Presence	$> 5\%$	$> 25\%$	ILD
With	1	69	82	86	58
Without	1	70	83	87	59
With	10	70	84	88	58
Without	10	74	86	89	58
Random	–	54	58	59	53

IV. DISCUSSION & CONCLUSION

We have presented and analyzed an approach for obtaining and learning from visual similarity triplets. We have shown that the approach for obtaining triplets yield reasonably high agreement between annotators and that CNNs can learn to predict unseen triplets. We have shown that the learned representations capture features that are relevant for both emphysema and ILD prediction. Emphysema and ILD predictions are based on a single slice in the upper right lung region (emphysema) and from the full lung fields (ILD). Detecting pathologies from a single slice is not always possible, because they may not be present in the slice, and performance should be viewed in this light.

We have proposed a method for modeling multiple notions of similarity during training and showed that it pushes the learned representations to decorrelate dimensions. Although this decorrelation did not improve performance, it could prove helpful for a better understanding of the representation. The decorrelated representations make it easier to satisfy triplets in at least one dimension. If we can learn a good predictor of which similarity notion to use for a specific triplet, it is likely that we can learn a better embedding by including this in the optimization.

A possible weakness of using multiple similarity notions as suggested here is that it provides few constraints. For eight dimensions we have 255 ways of picking a non-zero binary vector. Restricting the weight vectors to a smaller set would provide a stronger constraint that could improve learning. One way of restricting the weights would be to have a set of partially overlapping weight vectors (110 and 011 in the 3D case), which would allow each similarity notion to adapt independently, while still constraining all notions to partially respect each other.

Any data-driven analysis is at the mercy of the data. In this work we have randomly selected three sets of 100 scans from the same screening study population. We have deliberately not stratified subjects by ILD presence or emphysema extent scores. Our aim in this study was to consider visual similarity comparisons as an alternative to assessing individual image labels. Stratifying by image labels would defeat this aim and hide issues arising from variation in datasets. This has led to three datasets with different distributions of pathological findings. Although we have randomized the order in which queries where answered, we found that the triplet constraints in the three datasets where not equally consistent. We also found that performance was best (and similar) when training/testing on D_1/D_3 and vice versa. These results indicate that the observed variation in data and label distribution could be the cause of variation in performance across datasets. Future work would need to consider the scale of experiments to ensure the variation in images presented to raters is large enough to characterize the full dataset.

In this work, we have used a single rater to answer all queries. One of the aims of this study is learning representations that provide a more complete characterization of images, than representations learned for a specific task. Having multiple raters should provide a more diverse characterization supporting this aim. Additionally, having multiple raters allow us to estimate how difficult a query is by measuring inter-rater agreement for each query. High agreement could indicate easy triplets that must be respected, whereas low agreement could indicate queries with very similar images or different notions of similarity.

We used a 3×3 grid query to obtain visual similarity triplets. This introduces some dependence between the triplets from each query, which we ignore in this work by treating each triplet independently. Instead of learning from triplets, it could be useful to learn directly from 3 queries. The input to the network would then be one anchor image and eight query images. The output would be an embedding of the nine images, and the loss should ensure that the partial ordering from the visual similarity assessment was respected. This could be helpful in cases where query images are similar. If images are embedded close together, then it is probably not very important to get the ordering exactly right. Additionally, similarity notions could be assigned per query, which probably matches reality better than assigning a similarity notion to each triplet independently.

In summary, we have shown that assessing visual similarity of lung texture in a population with relatively low pathology prevalence is possible, and that visual similarity triplets can be used to learn low-dimensional embeddings that capture features of visually assessed pathologies. This positions learning from visual similarity as an approach that could reduce annotation cost by learning rich representations of image content from possibly crowdsourced annotations.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *arXiv preprint arXiv:1804.06353*, 2018.

[3] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, pp. 1–19, 2018.

[4] A. Kovashka, O. Russakovsky, L. Fei-Fei, K. Grauman *et al.*, "Crowdsourcing in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 10, no. 3, pp. 177–243, 2016.

[5] R. G. Gomes, P. Welinder, A. Krause, and P. Perona, "Crowdclustering," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 558–566. [Online]. Available: http://papers.nips.cc/paper/4187-crowdclustering.pdf

[6] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[7] C. Wah, S. Maji, and S. Belongie, "Learning localized perceptual similarity metrics for interactive categorization," in *2015 IEEE Winter Conference on Applications of Computer Vision*, Jan 2015, pp. 502–509.

[8] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 539–546 vol. 1.

[9] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[10] A. Veit, S. Belongie, and T. Karaletsos, "Conditional similarity networks," *Computer Vision and Pattern Recognition (CVPR 2017)*, 2017.

[11] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "Learning semantic and visual similarity for endomicroscopy video retrieval," *IEEE Transactions on Medical Imaging*, vol. 31, no. 6, pp. 1276–1288, June 2012.

[12] L. Sørensen, M. Nielsen, P. Lo, H. Ashraf, J. 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, Jan 2012.

[13] J. Hofmanninger and G. Langs, "Mapping visual features to semantic profiles for retrieval in medical imaging," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[14] G. Bortsova, F. Dubost, S. Ørting, I. Katramados, L. Hogeweg, L. Thomsen, M. Wille, and M. de Bruijne, "Deep learning from label proportions for emphysema quantification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 768–776.

[15] S. N. Ørting, V. Cheplygina, J. Petersen, L. H. Thomsen, M. M. Wille, and M. de Bruijne, "Crowdsourced emphysema assessment," in *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS 2017)*, 2017.

[16] S. N. Ørting, J. Petersen, V. Cheplygina, L. H. Thomsen, M. M. Wille, and M. de Bruijne, "Feature learning based on visual similarity triplets in medical image analysis: A case study of emphysema in chest ct scans," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS 2018)*, 2018.

[17] J. H. Pedersen, H. Ashraf, A. Dirksen, K. Bach, H. Hansen, P. Toennesen, H. Thorsen, J. Brodersen, B. G. Skov, M. Døssing, J. Mortensen, K. Richter, P. Clementsen, and N. Seersholm, "The Danish randomized lung cancer CT screening trial–overall design and results of the prevalence round," *Journal of Thoracic Oncology*, vol. 4, no. 5, 2009.

[18] M. M. Wille, L. H. Thomsen, A. Dirksen, J. Petersen, J. H. Pedersen, and S. B. Shaker, "Emphysema progression is visually detectable in low-dose CT in continuous but not in former smokers," *Eur Radiol*, vol. 24, no. 11, pp. 2692–2699, Nov 2014.

[19] 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.

[20] M. J. Wilber, I. S. Kwak, and S. J. Belongie, "Cost-effective hits for relative similarity comparisons," in *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[21] B. McFee and G. Lanckriet, "Learning multi-modal similarity," *J. Mach. Learn. Res.*, vol. 12, pp. 491–523, Feb. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.1953063

[22] L. Van Der Maaten and K. Weinberger, "Stochastic triplet embedding," in *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*. IEEE, 2012, pp. 1–6.
[23] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

6 Discussion and directions

The importance of machine learning in medical image analysis is growing. For tasks, such as segmentation, that lend themselves well to machine learning solutions, fully automated solutions seem more and more realistic. A major challenge when developing machine learning solutions for medical imaging tasks is that the methods are trained on data from a specific distribution and generalize poorly to new distributions. The problems can be subtle. Imagine a diagnostic tool that has been validated and found to have a false positive rate of 5% when applied on a population with 50% prevalence. Now apply the same tool in a population with 25% prevalence. Assuming the same false positive rate yields a 50% increase in number of false positives for a fixed population size, which could tip the scale from useful tool to unacceptable over-diagnosis.

Why is this interesting in the context of this thesis? Because it illustrates that training a machine learning method on one population and applying it on another population is problematic. Instead of hoping performance generalizes despite changes in data distribution, it might be better to focus on developing models that are cheap and easy to train. Weak supervision, crowdsourcing and learning from visual similarity are all aimed at reducing the labeling burden when developing machine learning methods.

Other approaches exists. Unsupervised methods aimed at characterizing common patterns in the images is a very interesting approach, because it eliminates the need for labels. Semi-supervised learning, where only a small part of the data is labeled, could be a better approach because it allows experts to guide the learning process such that it centers on the relevant features. This work has aimed at reducing the need for labels, without losing the benefits of supervised learning.

The following three sections discuss the three tracks of contributions presented in Sections 3, 4 and 5, with an emphasis on directions for future work.

6.1 Weakly supervised learning

The weakly supervised methods presented in this thesis have all used a representation based on classic scale-space features. There is little doubt now that CNNs can yield superior performance by learning a representation more closely tied to the task. Multiple layers enable CNNs to learn complex combinations of features that are hard or impossible to learn with traditional methods. The Simple MIL approach with a logistic classifier can in principle learn complex interactions between features, but the combinatorical explosion of possibilities limits this in practice, unless strong regularization, such as the spatial regularization enforced by convolutions, is used.

The fact that fine-tuning CNNs trained on ImageNet is effective for many tasks, suggests that as long as the representation is rich enough, it is interactions between features that is important for specific tasks. Although pre-training on ImageNet can be beneficial for medical imaging tasks, it is likely better to pre-train CNNs on a large and diverse set of medical imaging tasks, a medical ImageNet. This could provide representations that are generally useful for medical image analysis, and enable machine learning methods to be trained for specific tasks using much less annotated data.

It is quite interesting that the Simple MIL approach presented in [34] works as well as it does. The CNN analog is to add a global pooling layer after a pixel prediction layer. This approach also works quite well and does not require more data or better annotations [3]. An obvious question to ask is "what benefits are there to not using a CNN?". One answer is that we avoid optimizing features for a specific dataset. Compared to machine learning systems, humans are exceptionally good at ignoring intensity shifts, noise and changes in resolution. Ideally we want our computer systems to have the same qualities.

Emphysema can be detected quite well with a simple blob detector [50]. If we know that we need a blob detector, and we know that interesting blobs will always have intensities in [-1000HU, -850HU] then it is simple to create a set of blob detectors that covers this intensity range. On the other hand, if we have a dataset where interesting blobs never have intensities above -920HU, learned filters will not work well for data where blobs have intensities close to -850HU. Avoiding feature learning can improve generalization because we avoid over-fitting to the available data. However, the rise of CNNs have clearly illustrated that knowing which features are important is very difficult and learning filters is often the better option.

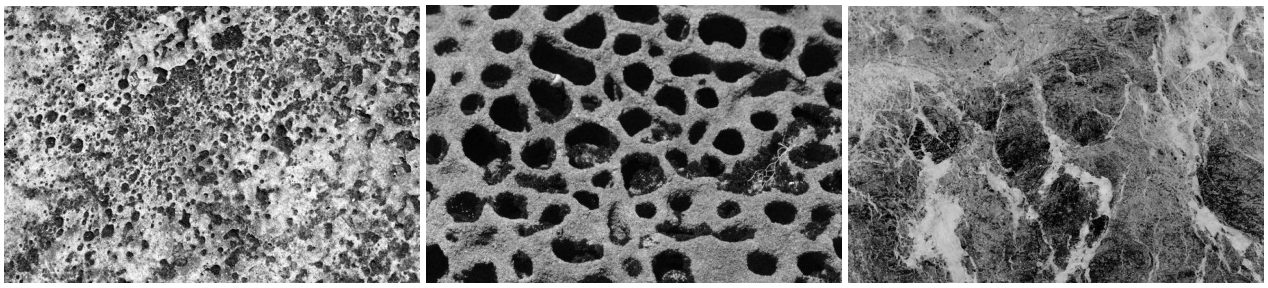

Figure 3: Naturally occurring textures that vaguely resembles lung tissue textures.

6.2 Crowdsourcing

The thesis argues that crowdsourcing is a path worth exploring for medical image analysis. However, bringing the power of crowdsourcing to medical imaging is a challenge. In many cases we are dealing with data that is private and cannot be shared with the public. In recent years, growing use and abuse of personal data has brought the issue to the attention of the public and to researchers working in medical image analysis.

Ideally we would not need medical data for training. For instance, we could try to train on naturally occurring texture patterns, such as those shown in Figure 3. However, applying this model directly on chest CT images will yield nonsensical results. If we instead of applying the model directly, first find a mapping from chest CT images to natural texture and then apply the model, it might work. Generative Adversarial Networks (GANs) have been shown to be very powerful for this kind of domain adaptation tasks, for example in robotics where simulated training environments are mapped to real world [27].

A likely easier task is mapping between closely related domains, such as chest CT scans from different studies. One of the challenges when using GANs is to constrain the mapping, such that new information is not introduced and important information is not lost. GANs learn a mapping by matching the data distributions. However, different studies will generally have different label distributions. Ignoring differences in label distributions is highly problematic, since the mapping will then force one study population to appear more or less diseased than it is. This problem was nicely illustrated in [6], where GANs learned to add and remove tumors when translating between brain MRI modalities.

When working with medical images we often have access to information about major sources of variation: scanners, scan protocols and populations. This information can be used to constrain the mappings. For example, if we know that one study used a hard reconstruction kernel and another used a soft reconstruction, we can probably make them more similar by blurring the images in the first study. Another example, imagine we want to apply a model for assessing emphysema trained on one domain in a new domain. Even though we do not have access to the distribution of emphysema in the new domain, we generally have knowledge about prevalence and severity of COPD, which could be incorporated to reduce issues arising from differences in label distribution.

Although crowdsourcing raises privacy concerns, it is not impossible to respect privacy while crowdsourcing labels. And although the privacy concern is most evident when crowdsourcing, the same issue faces anyone working within medical image analysis. If people do not trust researchers to ensure their data is not abused, they will increasingly block access to their data. Solving this challenge requires first and foremost that informed consent is obtained from the people whose data is to be used. Clearly communicating what data is needed, why it is needed and how it will be used is crucial to enable informed consent. Additionally, placing stricter demands on obtaining informed consent encourages researchers to define clear research goals and analysis protocols that, in general, leads to more robust findings.

A major challenge is allowing a breadth of exploratory work while still ensuring informed consent. One solution would be to establish a large public database of images, using crowdsourcing to extensively characterize the images. Combined with domain adaptation, this could allow models to be trained on the public database and applied on private databases.

In addition to proper handling of personal information, crowdsourcing also pose technical challenges . The primary technical concern when crowdsourcing labels is ensuring acceptable label quality. As illustrated in [32], lack of proper quality control leads to large variation in label quality, negatively impacting performance. Expertise from research areas such as quality control and user interface design, could help realize crowdsourcing as a valuable tool for developing of machine learning solutions to medical image analysis tasks.

6.3 Learning from visual similarity

A hypothesis in this thesis is that learning from visual similarity annotations can yield a more general characterization of images than learning from image labels. [33] and [36] investigate how to collect and learn from visual similarity triplets, but do not provide a comparison between learning from image labels and learning from visual similarity. Key elements in such a comparison are the cost of labeling and the information content of learned representations. If non-experts can perform visual assessment of emphysema, the cost can be reduced significantly. It is unclear how much expertise is needed for visual assessment of emphysema. No significant difference between assessments by radiologists and pulmonologists was found in [7], indicating that daily experience with assessing CT scans are not paramount. However, pulmonologists may lack experience with routine assessment of CT scans, but they have a deep knowledge of the lungs and abnormalities. It is unlikely that untrained crowds can achieve a similar quality of assessment.

In [28] a diverse group of annotators are tasked with segmenting six patterns in CT scans. Although individual annotators perform worse than an expert, combining the annotators yields performance similar to expert segmentations. However, annotators are informed about which pattern to segment and the study only includes slices from twenty subjects, so it is unlikely that the results transfer to assessment of emphysema in CT scans "in the wild".

Measuring which representation is better is tricky. If the task is to detect honeycombing, then CNNs trained to detect emphysema will not perform as well as CNNs trained to detect honeycombing. There is no technical need to restrict CNNs to a single class setting, and a CNN trained to predict both emphysema and honeycombing is likely as good as a CNN trained to detect one of the patterns. A proper comparison should compare learning from multi-class labels to learning from visual similarity.

One area where visual similarity has an advantage is when we do not know which patterns to look for. There is some controversy over the definition of emphysema sub-types. Annotating centrilobular, paraseptal and panlobular emphysema assumes that these patterns can be distinguished and that they cover all possibilities. Annotating visual similarity allows the CNN to learn without being forced to find three distinct classes, which could improve characterization and understanding of emphysema appearance in CT scans. On the other hand, learning from visual similarity could force the CNN to learn similarities that are irrelevant.

6.4 Closing remark

While writing this thesis an important question has presented itself. Is machine learning based on visual assessment of emphysema the right path? Densitometry is useful and easy to interpret. Many of the issues can be overcome by careful standardization of scan protocols and analysis software. Densitometry can also be complemented by relatively simple bullae detectors that provides some characterization of patterns and have been shown to be useful for lung cancer risk prediction [50]. Methods that learn from clinical outcomes, such as COPD stage and mortality, could also complement densitometry without requiring visual assessment.

The central question to answer is, which part of visual assessment is it that provides beneficial information. Is it robustness to noise, pattern recognition, holistic analysis or something else? The powerful feature learning aspect of CNNs could be used to analyze which low-level features are important when radiologists assess emphysema. Comparing this to densitometry and bullae detectors could improve understanding of the factors that are important in terms of disease progression and risk prediction.

The role of machine learning in deepening our understanding of emphysema and COPD is as, if not more, important as its role in enabling automated emphysema assessment.

7 Publications

[31] Silas Nyboe Ørting, Jens Petersen, Mathilde M W Wille, Laura H Thomsen, and Marleen de Bruijne. *Quantifying emphysema extent from weakly labeled CT scans of the lungs using label proportions learning*, The Sixth International Workshop on Pulmonary Image Analysis. 2016.

[34] Silas Nyboe Ørting, Jens Petersen, Laura H Thomsen, Mathilde M W Wille, and Marleen de Bruijne. *Detecting emphysema using multiple instance learning*. In 2018 IEEE 15th International Symposium on Biomedical Imaging. 2018.

[32] Silas Nyboe Ørting, Veronika Cheplygina, Jens Petersen, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. *Crowdsourced emphysema assessment*. In Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. 2017.

[33] Silas Nyboe Ørting, Jens Petersen, Veronika Cheplygina, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. *Feature learning based on visual similarity triplets in medical image analysis: A case study of emphysema in chest CT scans*. In Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS 2018). 2018.

7.1 In preparation

[35] Silas Nyboe Ørting, Jens Petersen, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. *Learning to quantify emphysema extent: What labels do we need?* arXiv preprint arXiv:1810.07433. 2018.

[29] Silas Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R. Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina. *A survey of crowdsourcing in medical image analysis*. arXiv e-prints, art. arXiv:1902.09159. Feb 2019.

[36] Silas Nyboe Ørting, Jens Petersen, Veronika Cheplygina, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. *Visual similarity comparisons for medical image analysis*. In preparation, 2019.

References

[1] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Transactions on Medical Imaging*, 35(5):1207–1216, May 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2535865.

[2] Polina Binder, Nematollah K Batmanghelich, Raúl San José Estépar, and Polina Golland. Unsupervised discovery of emphysema subtypes in a large clinical cohort. In *International Workshop on Machine Learning in Medical Imaging*, pages 180–187. Springer, 2016.

[3] Gerda Bortsova, Florian Dubost, Silas Ørting, Ioannis Katramados, Laurens Hogeweg, Laura Thomsen, Mathilde Wille, and Marleen de Bruijne. Deep learning from label proportions for emphysema quantification. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 768–776, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00934-2.

[4] Peter J. Castaldi, Raúl San José Estépar, Carlos S. Mendoza, Craig P. Hersh, Nan Laird, James D. Crapo, David A. Lynch, Edwin K. Silverman, and George R. Washko. Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers. *American Journal of Respiratory and Critical Care Medicine*, 188(9):1083–1090, Aug 2013. ISSN 1073-449X. doi: 10.1164/rccm.201305-0873OC. URL http://dx.doi.org/10.1164/rccm.201305-0873OC.

[5] Veronika Cheplygina, Lauge Sørensen, David MJ Tax, Jesper Holst Pedersen, Marco Loog, and Marleen de Bruijne. Classification of copd with multiple instance learning. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1508–1513. IEEE, 2014.

[6] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 529–536, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00928-1.

[7] COPDGene CT Workshop Group. A combined pulmonary-radiology workshop for visual evaluation of COPD: Study design, chest CT findings and concordance with quantitative evaluation. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 9(2):151–159, 2012.

[8] Harvey O Coxson, Asger Dirksen, Lisa D Edwards, Julie C Yates, Alvar Agusti, Per Bakke, Peter MA Calverley, Bartolome Celli, Courtney Crim, Annelyse Duvoix, et al. The presence and progression of emphysema in copd as determined by ct scanning and biomarker expression: a prospective analysis from the eclipse study. *The lancet Respiratory medicine*, 1(2):129–136, 2013.

[9] Adrien Depeursinge, Alejandro Vargas, Alexandra Platon, Antoine Geissbuhler, Pierre-Alexandre Poletti, and Henning Müller. Building a reference multimedia database for interstitial lung diseases. *Computerized medical imaging and graphics*, 36(3):227–238, 2012.

[10] Maggie A Flower. *Webb's physics of medical imaging*. CRC Press, 2012.

[11] Leticia Gallardo-Estrella, David A. Lynch, Mathias Prokop, Douglas Stinson, Jordan Zach, Philip F. Judy, Bram van Ginneken, and Eva M. van Rikxoort. Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification. *European Radiology*, pages 1–9, 2015. ISSN 0938-7994. doi: 10.1007/s00330-015-3824-y. URL http://dx.doi.org/10.1007/s00330-015-3824-y.

[12] Mingchen Gao, Ulas Bagci, Le Lu, Aaron Wu, Mario Buty, Hoo-Chang Shin, Holger Roth, Georgios Z Papadakis, Adrien Depeursinge, Ronald M Summers, et al. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(1):1–6, 2018.

[13] Pierre Alain Gevenois, Viviane De Maertelaer, Paul De Vuyst, Jacqueline Zanen, and Jean-Claude Yernault. Comparison of computed density and macroscopic morphometry in pulmonary emphysema. *American journal of respiratory and critical care medicine*, 152(2):653–657, 1995.

[14] GOLD. Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Pulmonary Disease (2019 Report), 2019. URL `http://www.goldcopd.org/`.

[15] Germán González, Samuel Y Ash, Gonzalo Vegas-Sánchez-Ferrero, Jorge Onieva Onieva, Farbod N Rahaghi, James C Ross, Alejandro Díaz, Raúl San José Estépar, and George R Washko. Disease staging and prognosis in smokers using deep learning in chest computed tomography. *American journal of respiratory and critical care medicine*, 197(2):193–203, 2018.

[16] Nadia N. Hansel, George R. Washko, Marilyn G. Foreman, MeiLan K. Han, Eric A. Hoffman, Dawn L. DeMeo, R. Graham Barr, Edwin JR Van Beek, Ella A. Kazerooni, Robert A. Wise, Robert H. Brown, Jennifer Black-Shinn, John E. Hokanson, Nicola A. Hanania, Barry Make, Edwin K. Silverman, James D. Crapo, and Mark T. Dransfield. Racial differences in ct phenotypes in copd. *COPD*, 10(1):20–27, Feb 2013. ISSN 1541-2555. doi: 10.3109/15412555.2012.727921. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321889/`. 23413893[pmid].

[17] Charles Hatt, Craig Galban, Wassim Labaki, Ella Kazerooni, David Lynch, and Meilan Han. Convolutional neural network based copd and emphysema classifications are predictive of lung cancer diagnosis. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 302–309. Springer, 2018.

[18] Johannes Hofmanninger and Georg Langs. Mapping visual features to semantic profiles for retrieval in medical imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 457–465, 2015.

[19] Stephen M. Humphries, Aleena M. Notary, Juan Pablo Centeno, and David A. Lynch. Automatic classification of centrilobular emphysema on ct using deep learning: Comparison with visual scoring. In Danail Stoyanov, Zeike Taylor, Bernhard Kainz, Gabriel Maicas, Reinhard R. Beichel, Anne Martel, Lena Maier-Hein, Kanwal Bhatia, Tom Vercauteren, Ozan Oktay, Gustavo Carneiro, Andrew P. Bradley, Jacinto Nascimento, Hang Min, Matthew S. Brown, Colin Jacobs, Bianca Lassen-Schmidt, Kensaku Mori, Jens Petersen, Raúl San José Estépar, Alexander Schmidt-Richberg, and Catarina Veiga, editors, *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 319–325, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00946-5.

[20] Y. Häme, 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)*, pages 109–113, April 2015. doi: 10.1109/ISBI.2015.7163828.

[21] Song Soo Kim, Joon Beom Seo, Namkug Kim, Eun Jin Chae, Young Kyung Lee, Yeon Mok Oh, and Sang Do Lee. Improved correlation between ct emphysema quantification and pulmonary function test by density correction of volumetric ct data based on air and aortic density. *European Journal of Radiology*, 83(1):57–63, 09 2012. doi: 10.1016/j.ejrad.2012.02.021. URL `http://dx.doi.org/10.1016/j.ejrad.2012.02.021`.

[22] David A Lynch, John HM Austin, James C Hogg, Philippe A Grenier, Hans-Ulrich Kauczor, Alexander A Bankier, R Graham Barr, Thomas V Colby, Jeffrey R Galvin, Pierre Alain Gevenois, et al. CT-definable

subtypes of chronic obstructive pulmonary disease: a statement of the Fleischner Society. *Radiology*, 277 (1):192–205, 2015.

[23] David A Lynch, Camille M Moore, Carla Wilson, Dipti Nevrekar, Theodore Jennermann, Stephen M Humphries, John HM Austin, Philippe A Grenier, Hans-Ulrich Kauczor, MeiLan K Han, et al. CT-based visual classification of emphysema: Association with mortality in the COPDGene study. *Radiology*, page 172294, 2018.

[24] Lena Maier-Hein, Sven Mersmann, Daniel Kondermann, et al. Crowdsourcing for reference correspondence generation in endoscopic images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 349–356. Springer, 2014.

[25] Annette McWilliams, Martin C Tammemagi, John R Mayo, Heidi Roberts, Geoffrey Liu, Kam Soghrati, Kazuhiro Yasufuku, Simon Martel, Francis Laberge, Michel Gingras, et al. Probability of cancer in pulmonary nodules detected on first screening ct. *New England Journal of Medicine*, 369(10):910–919, 2013.

[26] N L Müller, C A Staples, R R Miller, and R T Abboud. "density mask". an objective method to quantitate emphysema using computed tomography. *Chest*, 94(4):782–787, 1988. doi: 10.1378/chest.94.4.782. URL +http://dx.doi.org/10.1378/chest.94.4.782.

[27] Phuong DH Nguyen, Tobias Fischer, Hyung Jin Chang, Ugo Pattacini, Giorgio Metta, and Yiannis Demiris. Transferring visuomotor learning from simulation to the real world for robotics manipulation tasks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6667–6674. IEEE, 2018.

[28] Alison Q O'Neil, John T Murchison, Edwin JR van Beek, and Keith A Goatman. Crowdsourcing labels for pathological patterns in ct lung scans: Can non-experts contribute expert-quality ground truth? In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (MICCAI LABELS)*, pages 96–105. Springer, 2017.

[29] Silas Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R. Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina. A survey of crowdsourcing in medical image analysis. *arXiv e-prints*, art. arXiv:1902.09159, Feb 2019.

[30] Silas Nyboe Ørting. Automatic estimation of emphysema extent in low-dose CT scans of the lungs. Master's thesis, University of Copenhagen, Denmark, 2016.

[31] Silas Nyboe Ørting, Jens Petersen, Mathilde MW Wille, Laura H Thomsen, and Marleen de Bruijne. Quantifying emphysema extent from weakly labeled CT scans of the lungs using label proportions learning. *The Sixth International Workshop on Pulmonary Image Analysis*, pages 31–42, 2016.

[32] Silas Nyboe Ørting, Veronika Cheplygina, Jens Petersen, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. Crowdsourced emphysema assessment. In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS 2017)*, 2017.

[33] Silas Nyboe Ørting, Jens Petersen, Veronika Cheplygina, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. Feature learning based on visual similarity triplets in medical image analysis: A case study of emphysema in chest ct scans. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS 2018)*, 2018.

[34] Silas Nyboe Ørting, Jens Petersen, Laura H Thomsen, Mathilde M W Wille, and Marleen de Bruijne. Detecting emphysema using multiple instance learning. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018.

[35] Silas Nyboe Ørting, Jens Petersen, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. Learning to quantify emphysema extent: What labels do we need? *arXiv preprint arXiv:1810.07433*, 2018.

[36] Silas Nyboe Ørting, Jens Petersen, Veronika Cheplygina, Laura H Thomsen, Mathilde MW Wille, and Marleen de Bruijne. Visual similarity comparisons for medical image analyis. *In preparation*, 2019.

[37] David G. Parr, Asger Dirksen, Eeva Piitulainen, Chunqin Deng, Marion Wencker, and Robert A. Stockley. Exploring the optimum approach to the use of ct densitometry in a randomised placebo-controlled study of augmentation therapy in alpha 1-antitrypsin deficiency. *Respir Res*, 10(1):75–75, Aug 2009. ISSN 1465-9921. doi: 10.1186/1465-9921-10-75. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2740846/`. 1465-9921-10-75[PII].

[38] Jesper H. Pedersen, Haseem Ashraf, Asger Dirksen, Karen Bach, Hanne Hansen, Phillip Toennesen, Hanne Thorsen, John Brodersen, Birgit Guldhammer Skov, Martin Døssing, Jann Mortensen, Klaus Richter, Paul Clementsen, and Niels Seersholm. The Danish randomized lung cancer CT screening trial–overall design and results of the prevalence round. *Journal of Thoracic Oncology*, 4(5), 2009. ISSN 1556-0864.

[39] Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, 2010.

[40] J. P. Sieren, J. D. Newell, P. F. Judy, D. A. Lynch, K. S. Chan, J. Guo, and E. A. Hoffman. Reference standard and statistical model for intersite and temporal comparisons of ct attenuation in a multicenter quantitative lung study. *Med Phys*, 39(9):5757–5767, Sep 2012. ISSN 0094-2405. doi: 10.1118/1.4747342. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3448623/`. 049209MPH[PII].

[41] Lauge Sørensen, Saher B Shaker, and Marleen De Bruijne. Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE transactions on medical imaging*, 29(2):559–569, 2010.

[42] Lauge Sørensen, Mads Nielsen, Pechin Lo, Haseem Ashraf, Jesper H Pedersen, and Marleen De Bruijne. Texture-based analysis of copd: a data-driven approach. *IEEE transactions on medical imaging*, 31(1): 70–78, 2012.

[43] Marco Stolpe and Katharina Morik. Learning from label proportions by optimizing cluster model selection. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6913 of *Lecture Notes in Computer Science*, pages 349–364. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-23807-9.

[44] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[45] Ruwan Tennakoon, Gerda Bortsova, Silas Ørting, Amirali K. Gostar, Mathilde M. W. Wille, Zaigham Saghir, Reza Hoseinnezhad, Marleen de Bruijne, and Alireza Bab-Hadiashar. Deep multi-instance volumetric imageclassification with extreme value distributions. *Under review*, 2018.

[46] Gerard J Tortora and Bryan Derrickson. *Principles of Anatomy and Physiology, 2009.* Hoboken, N.J. : Wiley, 12 edition, 2009. ISBN 978-0-470-08471-7.

[47] Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

[48] Q. Wang, Y. Zheng, G. Yang, W. Jin, X. Chen, and Y. Yin. Multiscale rotation-invariant convolutional neural networks for lung texture classification. *IEEE Journal of Biomedical and Health Informatics*, 22 (1):184–195, Jan 2018. ISSN 2168-2194. doi: 10.1109/JBHI.2017.2685586.

[49] Mark O Wielpütz, Diana Bardarova, Oliver Weinheimer, Hans-Ulrich Kauczor, Monika Eichinger, Bertram J Jobst, Ralf Eberhardt, Marcel Koenigkam-Santos, Michael Puderbach, and Claus P Heussel. Variation of densitometry on computed tomography in copd–influence of different software tools. *PloS one*, 9(11):e112898, 2014.

[50] Rafael Wiemker, Merlijn Sevenster, Heber MacMahon, Feng Li, Sandeep Dalal, Amir Tahmasebi, and Tobias Klinder. Automated assessment of imaging biomarkers for the PanCan lung cancer risk prediction model with validation on NLST data. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 1013421. International Society for Optics and Photonics, 2017.

[51] Mathilde Marie Winkler Wille, Laura H. Thomsen, Jens Petersen, Marleen de Bruijne, Asger Dirksen, Jesper H. Pedersen, and Saher B. Shaker. Visual assessment of early emphysema and interstitial abnormalities on CT is useful in lung cancer risk analysis. *European Radiology*, pages 1–8, 2015. ISSN 0938-7994.

[52] Jie Yang, Elsa D. Angelini, Pallavi P. Balte, Eric A. Hoffman, John H. M. Austin, Benjamin M. Smith, Jingkuan Song, R. Graham Barr, and Andrew F. Laine. Unsupervised discovery of spatially-informed lung texture patterns for pulmonary emphysema: The mesa copd study. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pages 116–124, Cham, 2017. Springer International Publishing. ISBN 978-3-319-66182-7.