Crowdsourced emphysema assessment

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, 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 [13]. Visual scores can be used for training classifiers to automatically assess presence and extent of emphysema [14, 9]. However, visual scoring of emphysema by experts is both expensive and prone to high rater disagreement [13]. Instead of performing a full visual scoring, which requires expert knowledge 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 [12] 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 [6, 2, 8, 1]. 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 & 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 [13]. 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 (10mm) and located such that the center slice coincided with the center slice of the region. In this way we covered a depth of 80mm and avoided slices at the very boundary of the lungs. An example of an extracted set of slices is given in Figure 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.

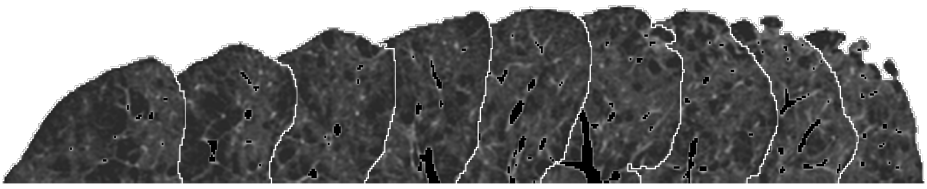

Fig. 1. Nine slices extracted from a single volume. There is a large extent of centrilobular emphysema. We can see that neighboring slices tend to have more similar texture patterns than slices that are far away from each other. White border added for clarity.

2.2 Crowdsourced triplets

We used Amazon MTurk[6] 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 Figure 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 seconds. The most productive worker submitted 131 HITs and the lowest work time for a HIT was 4 seconds. More than 92% of the HITs were finished within 30 minutes of the first HIT being available. The total cost was $388.80.

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

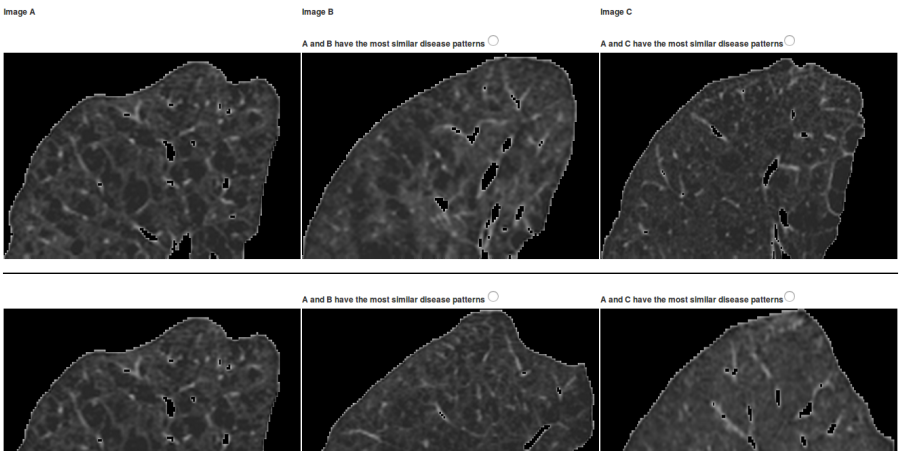

Fig. 2. Amazon MTurk user interface for collecting the similarity triplets

2.3 Similarity embedding

We used t-distributed stochastic triplet embedding (t-STE) [11] 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. 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[7].

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

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

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

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

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[8].

3 Experiments & 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

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

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

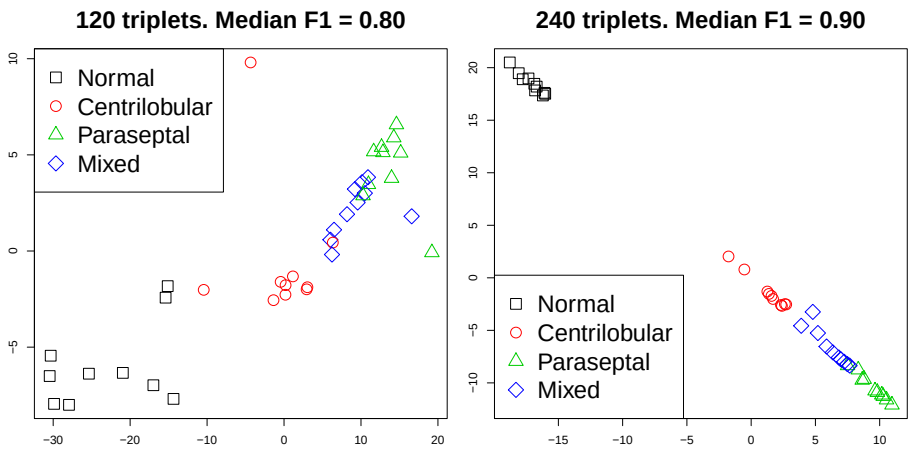

Fig. 3. Example embeddings from simulated triplets. Left: 120 triplets. Right: 240 triplets. While there is no overlap between emphysema and normal classes in both cases, there is some overlap between emphysema classes for 120 triplets.

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 figure 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.

3.2 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.

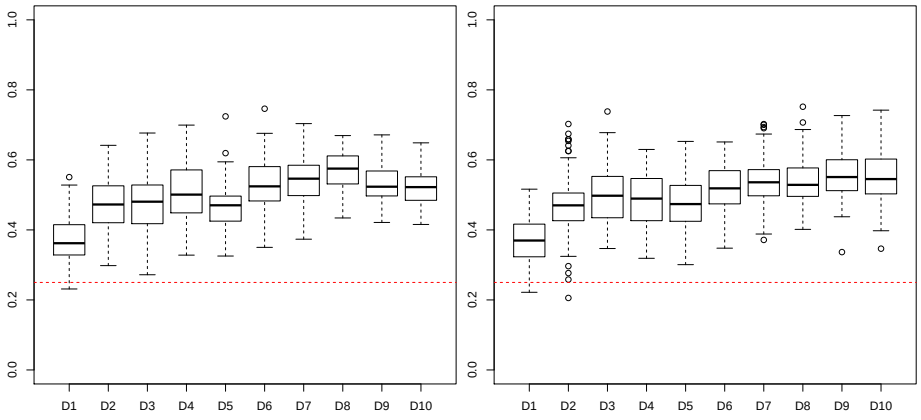

Fig. 4. Distribution of mean F_1 scores for classification of emphysema type. Left stratified, right combined. The dashed red line indicate random performance ($F_1 = 0.25$).

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.

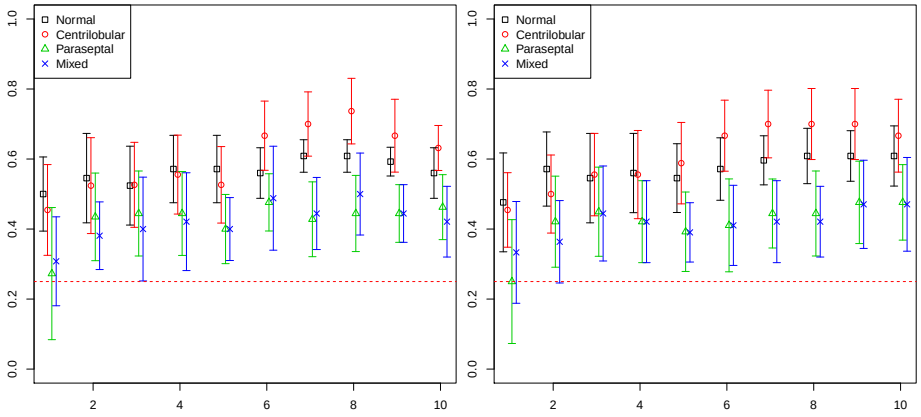

Fig. 5. F_1 scores for classification of emphysema type. Left for stratified, right for combined. The dashed red line indicate random performance ($F_1 = 0.25$). Symbols indicate median values and bars indicate ± 1 median absolute deviation.

4 Discussion & 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. Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE TMI*, 35(5):1313–1321, May 2016.
2. V. Cheplygina, A. Perez-Rovira, W. Kuo, H. Tiddens, and M. de Bruijne. Early experiences with crowdsourcing airway annotations in chest CT. In *MICCAI LABELS*, pages 209–218, 2016.
3. From the Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2015.
4. Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. Crowdsourcing in computer vision. *Found. Trends. Comp. Graphics and Vision*, 10(3):177–243, 2016.
5. Pechin Lo, Jon Sporring, Haseem Ashraf, Jesper JH Pedersen, and Marleen de Bruijne. Vessel-guided airway tree segmentation: A voxel classification approach. *Med Image Anal*, 14(4):527–538, 2010.
6. Lena Maier-Hein, Sven Mersmann, Daniel Kondermann, Sebastian Bodenstedt, Alexandro Sanchez, Christian Stock, Hannes Gotz Kenngott, Mathias Eisenmann, and Stefanie Speidel. Can masses of non-experts train highly accurate image classifiers? In *MICCAI*, pages 438–445. Springer, 2014.
7. Brian McFee and Gert Lanckriet. Learning multi-modal similarity. *J. Mach. Learn. Res.*, 12:491–523, February 2011.
8. Danny Mitry, Kris Zutis, Baljean Dhillon, Tunde Peto, Shabina Hayat, Kay-Tee Khaw, James E Morgan, Wendy Moncur, Emanuele Trucco, and Paul J Foster. The accuracy and reliability of crowdsource annotations of digital retinal images. *Transl Vis Sci Technol*, 5(5):6–6, 2016.
9. Mizuho Nishio, Kazuaki Nakane, Takeshi Kubo, Masahiro Yakami, Yutaka Emoto, Mari Nishio, and Kaori Togashi. Automated prediction of emphysema visual score using homology-based quantification of low-attenuation lung region. *PLOS ONE*, 12(5):1–12, 05 2017.
10. 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. *J Thorac Oncol*, 4(5), 2009.
11. Laurens van der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *IEEE MLSP*, pages 1–6, 2012.
12. Catherine Wah, Grant Van Horn, Steve Branson, Subhransu Maji, Pietro Perona, and Serge Belongie. Similarity comparisons for interactive fine-grained categorization. In *IEEE CVPR*, pages 859–866, 2014.
13. 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*, 24(11):2692–2699, Nov 2014.
14. Silas Nyboe Ørting, Jens Petersen, Mathilde Wille, Laura Thomsen, and Marleen de Bruijne. Quantifying emphysema extent from weakly labeled CT scans of the lungs using label proportions learning. In *MICCAI PIA*, pages 31–42. CreateSpace Independent Publishing Platform, 2016.