Comprehensive Emphysema Subtype Diagnosis Using Structured Expert Knowledge

Jaron Schaeffer[1], Mamatha Rudrapatna[1], Arcot Sowmya[1], Peter Wilson[2]

[1] University of New South Wales, Sydney, NSW 2052, Australia
[2] I-Med Network, Sydney, NSW 2000, Australia

Abstract. Computerised emphysema quantification has received a lot of research attention due to the mass availability of CT. Yet, to our knowledge, no existing method is able to recognise all common subtypes of the disease, a diagnosis routinely given by radiologists. In this paper, we present a HRCT-based Computer Assisted Diagnosis system for emphysema subtype diagnosis. The system first detects low-attenuation regions using adaptive density mask, a novel refinement to the classic density mask method. Detected regions are then classified individually and results combined in a bottom-up manner to achieve per-patient diagnosis and quantification. Expert knowledge necessary for classification decisions was acquired incrementally using a multi-level Ripple Down Rules system. Evaluation shows that the multi-level approach well reflects the pathological characteristics of the subtypes, and RDR knowledge management provided robust diagnosis using very little training data.

1 Introduction

Emphysema is a chronic obstructive pulmonary disease (COPD) typically caused by exposure to tobacco smoke. Alveolar walls break down due to inflammatory responses to the particles inhaled and the affected regions show low attenuation on high-resolution computed tomography (HRCT) scans. Patients suffer from limited respiratory capabilities as the disease progresses.

Automated emphysema detection has been researched for almost as long as CT has existed; Sluimer *et al.* recently provided an overview and performance comparison in their survey on computerised CT analysis of the lung [1]. Despite its age, density mask (DM), a method introduced by Muller *et al.* in 1988 [2], is still considered the de-facto standard in computerised quantification [3]. DM calculates the percentage of the lung showing below-normal attenuation and has been shown to correlate well with pulmonary function tests, the standard for emphysema diagnosis in pre-CT times [4, 5]. However, DM and other emphysema quantification methods are unable to detect disease subtypes, a diagnosis made routinely by radiologists when diagnosing scans manually.

To build a computer assisted diagnosis (CAD) system capable of comprehensive subtype recognition, extensive expert knowledge must be transferred into the system systematically and in a time-efficient way. Ripple Down Rules

(RDR), a knowledge engineering technology for expert systems introduced by Compton and Jansen [6], is suitable for the task. RDR are ordered lists of rules with exceptions. If a misclassification is detected for an input case while reviewing classification results, an exception is appended to the firing rule to correct the conclusion; a new rule is added to the knowledge base if no rules fire at all; and no change is made for correctly classified cases (*review/approve-or-create-rule cycle*). This makes RDR an *incremental* or *per-case* knowledge acquisition method, where knowledge base testing and rule creation merges into a single working step to make use of every sample reviewed. New rules are validated against the existing knowledge base to ensure knowledge consistency. For any new rule or exception created, the human expert is responsible for formulating a rule condition based on some features of the case investigated. The choice of these conditions is a trade-off between rule specificity and the desired generalisation to similar but unseen future cases.

Feature extraction and feature design are crucial for the success of any CAD system, and often, new features are added as the expert finds that the features currently available are insufficient. If the diagnosis model was represented using a machine learning approach, this would require re-training; with RDR being an incremental learner, the expert would simply incorporate new features into the newly built rules. Another important property of RDR is that the system is built while being already in use. RDR can be seamlessly integrated into the day-to-day workflow of the experts as demonstrated by the commercially successful version of RDR for pathology domain [7, 8]. It is known that RDR produce knowledge bases (KB) similar in size to those developed by machine learning [9] and the time taken to add a rule remains roughly constant regardless of the size of the KB [8]. An enhanced version of RDR known as Multiple Classification RDR (MCRDR) [10] can provide multiple conclusions to a given case.

In this paper, we present a CAD system for classification and quantification of *centrilobular* (*centriacinar*), *panlobular* (*panacinar*) and *paraseptal* (*bullous*) emphysema (Fig. 1). The system design is inspired by the way we observed the radiologist in our group handle the task manually: raw emphysema regions are detected (Sect. 2) and classified individually based on a set of region features. Results are combined to provide input to a higher-level classification step (Sect. 3). Both steps use RDR rule bases for expert knowledge management and reasoning, forming a novel multi-level RDR classification system. We discuss evaluation methods and some results in Sect. 4 and summarise remaining challenges and ideas for future work in Sect. 5.

2 Emphysema Detection and Feature Extraction

Raw emphysema region detection is based on density mask, as this method is widely accepted by radiologists and studies have been carried out to select good threshold values [4, 5]. However, a pure DM approach fails in our case for later stage emphysema cases (*confluent emphysema*), where large regions of very serrated appearance are detected that provide no meaningful shape information

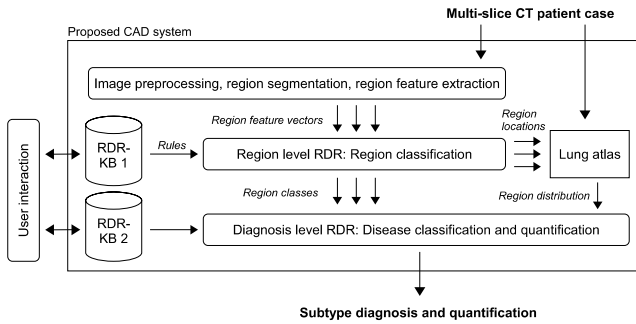


Fig. 1. CAD System overview

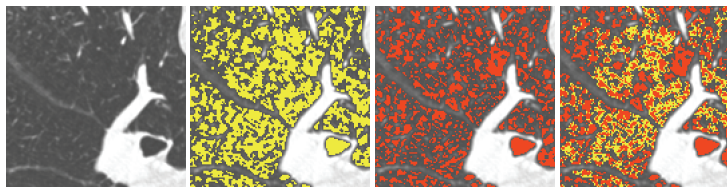
(see Fig. 2). To compensate, we have developed a DM variant called Adaptive Density Mask (ADM), somewhat similar to but computationally more efficient than the standard Watershed transform [11]. ADM was designed to give results similar to standard DM except for confluent cases, where it creates more and smaller unconnected regions more suitable for subsequent region analysis.

2.1 Adaptive Density Mask

The idea behind ADM is to select a binarisation threshold for each emphysema region automatically to maximise the number of distinct regions while simultaneously maximising their area, both within the limits imposed by minimum/maximum density masks. Traditional minima detection methods capable of achieving this [11, 12] are expensive, over-segment the image and require advanced de-noising beforehand. ADM provides a simple, fast and robust tradeoff between sensitivity and simplicity for the drawback of having to select a number of additional parameters.

It is worth pointing out that, by definition, DM (and any method based on it such as ADM) does not produce false positives in the traditional sense; every detected image pixel/voxel is considered as indication of the presence of emphysema and thus clinically relevant. Fig. 3 illustrates ADM schematically for a real-valued signal; in particular, the method works as follows:

Given a set of thresholds $t_1, \ldots, t_n$, a DM is calculated on the input image for each threshold (for e.g. -965 HU to -945 HU using 5 HU steps in Fig. 3). For each resulting distinct region (solid horizontal black lines), the minimum attenuation is determined and the corresponding x/y position is associated as an unique index to the region (squares and vertical lines). We can think of the different density mask outputs as an ordered stack of layers as indicated by the solid horizontal lines (binary one) and the dotted horizontal lines (binary zero) in Fig. 3. Each region is recursively visited by the following procedure, starting on the topmost layer:



(a) (b) (c) (d)

Fig. 2. Standard density mask (DM) compared to the adaptive density mask (ADM) method for raw emphysema segmentation. (a) Original image region, a late case of centrilobular emphysema. (b) DM output using a threshold of -950 HU (yellow). (c) ADM output using -975 HU, -970 HU, ..., -945 HU thresholds (red). (d) Overlay of the DM (yellow) and ADM (red) segmentations from (b) and (c). For large segments of late panlobular and centrilobular emphysema, ADM creates more and smaller unconnected regions as compared to DM that results in large, serrated regions.

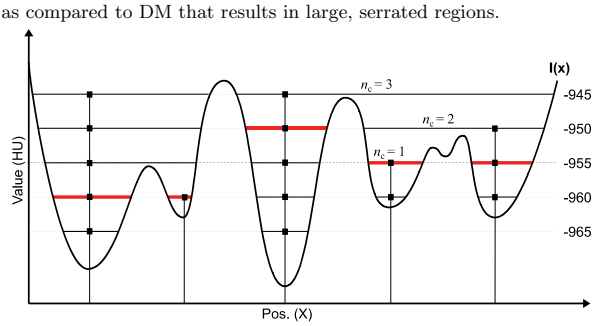


Fig. 3. Adaptive Density Mask: schema for a real-valued signal. The density mask outputs for the different thresholds are depicted in solid (binary one) and dotted (binary zero) lines, resulting emphysema regions are highlighted in bold red. For some regions, the number of regions below the current layer n_c is given.

- Find the number n_c of regions below the current layer covered by the current region. For the rightmost region in Fig. 3 on the -945 HU density mask, $n_c = 3$.
- If $n_c \leq 1$, the current region on the current layer will cover a maximum area while simultaneously maximising the number of regions: recursing to a lower layer will not increase the number of regions but will decrease the region area. Mark the current region in the result mask as binary one and stop recursion.
- If $n_c > 1$, visit the covered region in the layer below.

We observe the following two properties of ADM:

1. *Local minima detection:* The image is re-quantised using $t_1, \ldots, t_n$, and every local minimum detectable in the re-quantised image creates a distinct region

in the binary result image. This ensures that even for late confluent cases, shape information is retained.
2. *Maximum region area:* While the number of regions in the binary result image is determined by the number of detected local minima, each such region has maximum area. This makes ADM behave similarly to DM where the emphysema regions are generally well formed and isolated (non-confluent).

2.2 Region Feature Extraction

We extract multiple features for each low-attenuation region: Shape (compactness and elongation using image moments), first-order statistics (mean HU, HU standard deviation and uniformity using min/max normalised standard deviation), connectivity to the lung boundary (*pleura*), edge-radius-symmetry transform [13] for wall detection and a morphological wall detection approach in which the region is dilated and first order statistics are calculated for the new region pixels, resulting in a total of 12 features per region.

3 RDR Based Multi-level Classification

Our multi-level RDR system design is motivated by the finding that presence and distribution of the different emphysema region classes in a patient are commonly used by radiologists for subtype diagnosis [3]. Emphysema region identification and classification is a task on its own and must not be confused with the higher-level per-patient subtype classification.

Each disease subtype is dominated by a special type of low attenuation region [3] (see Table 1 for a brief characterisation). Accordingly, the extracted region feature vectors are used to classify regions into one of these region classes using a *region level* RDR layer. In a second step, a *diagnosis level* RDR layer builds on the region level results and produces the desired subtype diagnosis as outlined in the schema in Figure 1.

Table 1. Region class characterisation and relation to emphysema subtypes (characterisation extracted from [3])

Region class	Appearance	Predominant in subtype
Bullous	Round, compact, any size. Very low density, typically touching the pleura, visible walls	Paraseptal
Centrilobular	Small/medium size; typically rounded shape	Centrilobular
Diffuse	Large, often smoothly outlined by interlobular fissures and pleura; uniform density with smaller vessels within	Panlobular

Expertise for the region classification decision was transferred into the system using the common RDR cycle: new rules or exceptions to existing rules are added

by the expert either to correct a misclassified sample or to remove false positives. The radiologist in our group, a specialist on lung CT with more than 30 years of experience, created a total of 22 rules for the region classification rule base. Effectively, 22 regions from a number of typical cases were thus used as the gold standard (see Figure 4 for a sample rule created).

– TouchingPleura = 1, Area > 6.95 mm^2, Compactness > 40%, MeanDensity < −974 HU → **Bullous**
except
• WallsMorphMean < −890 HU → **Centrilobular**
except
∗ WallsMorphStddev < 100 HU, Uniformity > 85% → **Bullous**

– EmphysemaDistribution={Diffuse,Bilateral}, EmphysemaTotalLungPct > 30% → **Moderate panlobular emphysema**

Fig. 4. Two sample RDR rules created during system development: a region level rule with an exception that in turn has an exception (top) and a diagnosis level rule (bottom).

Percentage involvement of all the three region types detected in the previous step are calculated for various lung regions. These percentages as well as distribution and predominance attributes (such as e.g. diffuse/focal, unilateral/bilateral distributions or apical/middle/basal predominance, calculated for each region class separately using [14]) are presented as features at the diagnosis level RDR. We use multi-classification RDR [10] to allow for co-existence of different emphysema subtypes, and each subtype is individually quantified as either *absent*, *mild*, *moderate* or *severe*. Figure 4 shows a sample diagnosis level rule based on these features.

4 Classifier Performance Evaluation

Evaluations were carried out separately for the region level and the diagnosis level layers. In addition, in order to examine whether our region features are powerful enough for a clear discrimination between the three region classes, we compared region classification results obtained using standard machine learning methods to expert opinion, where high agreement would indicate sufficiently powerful region features.

A dataset consisting of 4,514 manually labelled regions (176 bullous, 4,193 centrilobular and 145 diffuse) from 9 scans of different patients was created using a designated labelling tool to train and evaluate the following classifiers: a decision tree (C4.5), a naive Bayesian, decision tables using the Inducer of Decision Table Majority (IDTM) induction algorithm [15] and a fully connected single hidden layer perceptron classifier, all through the WEKA data mining

suite for Java [16]. Sample sizes for each class were chosen to roughly reflect the frequency distribution of the different region classes. Classifiers were trained using 10-fold stratified cross validation using 10 repetitions. Table 2 shows high agreement for all classifiers tested and proves that the proposed region features are sufficiently powerful for region discrimination.

For comparison, Table 3 displays the confusion matrix for the region-level RDR classifier after the creation of the 22 rules mentioned above for the same dataset. Very good performance is observed for the centrilobular and the diffuse region classes, while almost 33% of the bullous class are missed. Apparently, an insufficient number of samples were reviewed to trigger rule creation and knowledge transfer necessary for reliable classification of this region type. Also, since the same expert hand-labelled the region dataset and created the RDR rules, we cannot eliminate the possibility of some bias and thus over-estimation of the RDR classifier performance.

Table 2. Class-specific and overall classification results for standard machine learning region level classifiers using 4514 training regions. *F-Measure* is defined as the harmonic mean between classifier precision and recall; the *Total F-Measure* column contains a weighted sum of the class-specific F-Measures (weights proportional to class sample size).

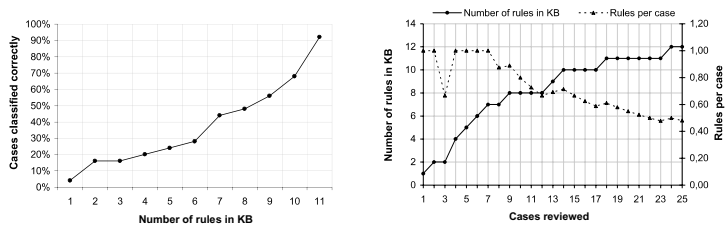
	Class-specific F-Measure (%)			
Classifier	Bullous (176 inst.)	Centrilobular (4,193 inst.)	Diffuse (145 inst)	Total F-Measure
C4.5	67.50	98.50	89.80	97.01
Naive Bayes	50.40	94.80	64.20	92.09
Decision Tables	63.70	98.40	89.10	96.75
Perceptron	65.20	98.20	85.00	96.49

Table 3. Confusion matrix for RDR region level classifier using 22 training regions. The value at the bottom right shows the combined F-Measure for the RDR classifier as a result of a weighted average of the class-specific F-Measures (class weight proportional to class sample size).

↓ Label/classified →	Bullous	Centrilobular	Diffuse	F-Measure (%)
Bullous	119	55	2	66.48
Centrilobular	57	4,133	3	98.49
Diffuse	6	12	127	91.70
Combined				97.02

For the diagnosis level layer, the radiologist in our team reviewed an additional 25 scans (different from the ones used to create the region dataset) manually. Since raw emphysema detection is out of scope for this work, we only used scans that appeared to contain emphysema through their relatively high DM coverage ($31.3\% \pm 8.4\%$ for the selected scans using a standard -950 HU threshold). Emphysema subtypes present were identified and a severity diagnosis was given for each case. 5 scans were diagnosed not to show any emphysema. Among the others, 1 showed mild paraseptal (bullous) emphysema, 7 centrilobular (mild: 1, moderate: 2, severe: 4) and 14 panlobular (mild: 10, moderate: 2, severe: 2) emphysema. The diagnosis level RDR knowledge base was initially empty, and during the course of the review, the radiologist added new rules whenever the system came to a wrong or no conclusion.

Altogether, 12 rules were created this way; we recorded the order in which scans were reviewed and whether one or more rules were created to adjust the system behaviour. Two different visualisations of the process can be seen in Fig. 5. Figure 5 (a) shows that 12 rules were needed on the whole to classify the 25 scans correctly. Figure 5 (b) clearly highlights a plateau effect (solid line) and a decay (dotted line) as the number of scans reviewed increases: 10 out of the 12 rules (83.3%) were added for the first 17 scans (68%), confirming the well-known RDR benefit of robust classification results after using only a small number of training samples. Rather than testing on unseen cases, every new case is used for training in the RDR tradition.



(a) (b)

Fig. 5. Results of diagnosis level RDR evaluation. (a) Number of rules in the knowledge base and overall classifier performance on the 25 scan test set. (b) Scans reviewed and number of rules in diagnosis level RDR KB for the 25 scan test set used for evaluation (left axis)/average number of rules per scan (right axis).

5 Conclusion and Outlook

We have presented the first CAD system capable of a comprehensive emphysema diagnosis including disease subtypes. The RDR subtype diagnosis is parameterless and only influenced by the knowledge transferred into the system by the

radiologist(s), suggesting a scenario for a routine clinical use: a fraction of the scans passing through the CAD system in daily use can be reviewed manually and inappropriate classification behaviour can be adjusted. Evaluation of the RDR region classification showed results comparable to state-of-the-art machine learning techniques using only a fraction of training samples.

Longer-term clinical application is one of our goals and still the only way to demonstrate the system's practicability. However, for an incremental system such as the proposed one, results can be expected to improve as more knowledge is gradually added to the adapting system. The presented region detection and multi-level RDR classification may also be used as a generalised framework for subtype classification of other lung diseases (e.g. asbestos-related diseases using pleural plaque, diffuse thickening and pleural rind regions).

For the raw emphysema detection step, we have presented a refined version of density mask called Adaptive Density Mask (ADM) as an extension to the standard density mask method. Only ADM makes it possible to analyse the shape of low-attenuation regions even with late cases of emphysema.

Data for this work was recorded using scan intervals of 10-15 mm, preventing accurate 3D processing. However, extending the proposed system to 3D should be a straight forward task and can be expected to increase robustness and accuracy.

Throughout this paper, we assume that the CT scans under investigation contain emphysema and no other lung disease characterised by low-attenuation regions. This assumption might not hold in practice, for e.g. fibrosis often co-occurs with emphysema and might influence ADM detection results. In addition to that, non-emphysema low attenuation regions (such as e.g. the bronchial tree) should be segmented beforehand and excluded to improve the reliability of the computed volumes.

References

1. Sluimer, I., Schilham, A., Prokop, M., van Ginneken, B.: Computer analysis of computed tomography scans of the lung: a survey. IEEE Transactions on Medical Imaging **25**(4) (April 2006) 385–405
2. Muller, N., Staples, C., Miller, R., Abboud, R.: Density mask. An objective method to quantitate emphysema using computed tomography. Chest **94**(194) (1988) 782–787
3. Webb, W.R., Mueller, N.L., Naidich, D.P.: High-Resolution CT of the Lung. 3rd edition edn. Lippincott Williams & Wilkins (2001)
4. Stern, E., Frank, M.: CT of the lung in patients with pulmonary emphysema: diagnosis, quantification, and correlation with pathologic and physiologic findings. American Journal of Roentgenology **162**(4) (1994) 791–798
5. Kinsella, M., Muller, N., Abboud, R., Morrison, N., DyBuncio, A.: Quantitation of emphysema by computed tomography using a" density mask" program and correlation with pulmonary function tests. Chest **97**(2) (1990) 315
6. Compton, P., Jansen, R.: Knowledge in context: a strategy for expert system maintenance. In: AI '88: Proceedings of the second Australian joint conference on Artificial intelligence, New York, NY, USA, Springer-Verlag New York, Inc. (1990) 292–306

7. Edwards, G., Compton, P., Malor, R., Srinivasan, A., Lazarus, L.: Peirs: A pathologist-maintained expert system for the interpretation of chemical pathology reports. Pathology **25**(1) (1993) 27–34
8. Compton, P., Peters, L., Edwards, G., Lavers, T.: Experience with Ripple-Down Rules. Applications And Innovations in Intelligent Systems XIII: Proceedings of AI-2005, the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, December 2005 (2006)
9. Compton, P., Preston, P., Kang, B., Yip, T.: Local patching produces compact knowledge bases. A Future for Knowledge Acquisition/EKAW '94 **94**
10. Kang, B., Compton, P., Preston, P.: Multiple Classification Ripple Down Rules: Evaluation and Possibilities. Proceedings 9th Banff Knowledge Acquisition for Knowledge Based Systems Workshop (1995) 17–1
11. Roerdink, J., Meijster, A.: The Watershed Transform: Definitions, Algorithms and Parallelization Strategies. Mathematical Morphology **41** (2000) 187–S28
12. Fairfield, J.: Toboggan contrast enhancement for contrast segmentation. Pattern Recognition, 1990. Proceedings., 10th International Conference on **1** (Jun 1990) 712–716 vol.1
13. Chabat, F., Hu, X.P., Hansell, D.M., Yang, G.Z.: ERS transform for the automated detection of bronchialabnormalities on CT of the lungs. IEEE Transactions on Medical Imaging **20**(9) (September 2001) 942–952
14. Zrimec, T., Busayarat, S., Wilson, P.: A 3D Model of the Human Lung. Medical Image Computing and Computer-Assisted Intervention-MICCAI (2004) 1074–1075
15. Kohavi, R.: The power of decision tables. Proceedings of the Eighth European Conference on Machine Learning (1995) 174–189
16. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd edition edn. Morgan Kaufmann (June 2005)