Applying Rough Sets Algorithm for Radiography Diagnosis

Oliviu MATEI
Technical University of Cluj-Napoca
str.G. Baritiu nr. 26-28, 400027 Cluj-Napoca
oliviu.matei@holisun.com

***Abstract*— The researchers and practitioners of today create models, algorithms, functions, and other constructs defined in abstract spaces. The research of the future will likely be data driven. Symbolic and numeric data that are becoming available in large volumes will define the need for new data analysis techniques and tools. Machine learning is an emerging area of computational intelligence that offers new theories, techniques, and tools for analysis of large data sets. In this paper, a novel approach for autonomous decision-making is developed based on the rough set theory of data mining (and machine learning). The approach has been tested on a medical data set for patients with lung abnormalities. The two independent algorithms developed in this paper either generate an accurate diagnosis or make no decision. The methodology discussed in the paper depart from the developments in data mining as well as current medical literature, thus creating a variable approach for autonomous decision-making.**

***Index Terms*—rough sets, radiographs**

I. INTRODUCTION

The interest in medical decision-making has been gaining momentum in recent years. In this paper, new algorithms for decision-making based on prior data are proposed. The algorithms are built on the concepts from rough set theory, cluster analysis, and measure theory. Computational analysis indicates that the proposed algorithms offer numerous advantages over other approaches such as neural networks and regression analysis, namely:

- simplicity;
- high accuracy;
- low computational complexity. Regression analysis and neural networks share the following characteristics.
- Each involves a learning phase and a decision-making phase.
- Both make decisions essentially for all objects with unknown outcomes, however, with an error.
- Both require specialized software or even hardware (some neural networks).
- The models associated with neural networks and regression models are "population based," which means that one model is developed for all cases in a training data set. Such a model uses a fixed number of features.

One of the two algorithms proposed in this paper uses decision rules extracted from a training set. The feature extraction approach follows an "individual (data object) based" paradigm. A feature extraction algorithm identifies unique features (test results, symptoms, etc.) of an object (e.g., a patient) and checks whether these unique features are shared with other objects. It is obvious that the "population based" and "individual based" paradigms differ and, in general, the set of features derived by each of the two paradigms is different. In the feature extraction approach, a set of features applies to a group of objects. These features are expressed as a decision rule. The properly derived decision rules accurately assign outcomes for a large percentage of cases with unknown decisions (i.e., make predictions for new cases). The drawback of the feature extraction approach is high computational complexity of the learning phase; however, it offers a greater promise for applications in decision-making than any of the "population-based" based approaches. The approach presented in this paper follows the emerging concepts from the rough set theory [15] of data mining. The reasoning behind the rough set theory is that a group of objects (patients) with a unique subset of features shares the same decision outcome. The feature extraction algorithm dynamically analyzes a large database and identifies unique features of each group of objects. Importantly, the subset of features is not specified in advance. In expert systems, rules guiding diagnostic decisions are fixed, while the rules generated with the rough set theory approach are dynamic and unique to each group of objects [3]. An important aspect of the approach proposed in this paper is that the decisions (diagnoses) are accurate for of objects with unknown outcomes, where and possibly could approach zero. To accomplish such high decision-making accuracy, the diagnostic decisions are made by two independent algorithms:

- primary decision-making algorithm;
- confirmation algorithm.

Both algorithms utilize features, however, in an orthogonal way. The computer-generated decision is accepted only if the solutions generated by the primary and confirmation algorithms agree. The proposed approach is illustrated with a medical case study involving diagnosis of patients with various pulmonary diseases using information from noninvasive tests.

Before that, we present the pulmonary abnormalities with their characteristics. Based on them, we determine those feature which are worth to be taken into account in applying the two algorithms.

II. RELATED WORK

Machine learning is an emerging area of computational intelligence that offers new theories, techniques, and tools for processing large data sets. It has gained considerable attention among practitioners and researchers especially for the practical applications, including in medical imaging. The growing volume of data that is available in a digital form spurs this accelerated interest. One of the few theories developed specifically for machine learning is the rough set theory [16]. It has found applications in industry, service organizations, healthcare [12], software engineering [20], edge detection [26], data filtration [22], and clinical decision making [23], [25].

A comprehensive comparative analysis of prediction methods included in [11] indicates that automatically generated diagnostic rules outperform the diagnostic accuracy of physicians. The authors' claim is supported by a comprehensive review of the literature on four diagnostic topics: localization of a primary tumor, prediction of reoccurrence of a breast cancer, thyroid diagnosis, and rheumatoid prediction. In this paper, the concept of feature extraction, cluster analysis, and measure theory are used to develop low computational complexity and accurate algorithms for the diagnosis of lung diseases.

Bram van Ginneken [5] classified the tumors as benign or malignant using the k-nearest neighbor (kNN) algorithm.

However, a great amount of work was invested in analysis of the radiological images. The thoracic applications of greatest interest include the detection and volume measurement of lung nodules [1], [6]. The article by Lee et al. [14] in this issue of the Korean Journal of Radiology is one of the few studies to examine the influence of a commercially available CAD system on the detection of lung nodules.

Many studies have revealed that CAD systems are effective at detecting small pulmonary nodules on radiographs [7], [19], and the ultimate goal of CAD systems is the detection of malignant lung nodules. Although Armato et al. [1] reported that a large fraction of missed lung cancers were detected using a CAD system, no observer-based study has assessed CAD schemes for lung cancer detection.

III. ABNORMALITIES FEATURES

The features taken into account when discussing the lung abnormalities are:

Structure: which may be homogeneous, inhomogeneous or reticular. According to [4], the homogeneity is defined as:

$$H = \sum_i \sum_j \frac{P_d(i,j)}{1+|i-j|}$$

In the above formula, P_d is the gray level co-occurrence matrix [10] for a displacement vector $d = (d_x, d_y)$. The entry (i, j) of P_d is the number of occurrences of the pair of gray levels i and j which are a distance d apart. Formally, it is given as

$$P_d(i,j) = \left|\{((r,s),(t,v)) : I(r,s) = i, I(t,v) = j\}\right|$$

Where:

$$((r,s),(t,v)) \in N \times N)$$

$$(t,v) = (r+dx, s+dy)$$

and |.| is the cardinality of a set.

$$G = [G_x G_y]$$

$$G_m = \sqrt{G_x^2 + G_y^2}$$

$$q = \tan^{-1}\left(\frac{G_x}{G_y}\right)$$

There are several well known gradient filters [4], such as Roberts, Sobel or Prewitt operators. In our experiments, the isotropic operators were used:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -\sqrt{2} & 0 & \sqrt{2} \\ -1 & - & 1 \end{bmatrix}, G_y = \begin{bmatrix} -1 & -\sqrt{2} & -1 \\ 0 & 0 & 0 \\ 1 & \sqrt{2} & 1 \end{bmatrix}$$

Intensity: is another important characteristic of an opacity. In radiographs, lung cancer may appear as a solid nodule, a partly solid nodule, or as a non-solid nodule. Many studies have suggested that these non-solid or partly solid nodules represent precursors to an early adenocarcinoma. Despite their potential clinical significance, nodules in this category may not be detected by radiographs [14], and most CAD schemes for detecting lung nodules are designed and optimized for the detection of solid nodules. Much research is currently targeted at resolving this problem.

Each pixel of a gray scale image has a pixel value which describes how bright that pixel is. The most common pixel format is the byte image, where this number is stored as an 8-bit integer giving a range of possible values from 0 to 255. Typically zero is taken to be black, and 255 is taken to be white. Values in between make up the different shades of gray.

Surrounding tissues: Some diseases may compact the surrounding tissues, creating the so-called "atelectasis".

Evolution: Some opacities may evolve in time and this is an important aspect that makes the difference between diseases. For instance, benign tumors do not change in time, whereas the malignant tumors grow in years or even months.

Lung: usually any lung may be affected by pulmonary diseases, but some of them are preponderant in the right lung. Others, such as pulmonary edema appear in both lungs in the same time.

Position:depending on the origin of each disease, specific parts of the lung may be affected. However, most pulmonary diseases may be located anywhere in the lung field. On the other hand, the pulmonary edema starts from the lower lobe of the lung and grows up in time. Pneumococcal pneumonia and secondary tuberculosis are located in the upper part of the lung, and the primary tuberculosis generates opacities in the lower lobe as well as around the hilum.

To define the positions of the opacities, the lung is divided into smaller regions. As a first step, each lung is divided into 4 parts: supra-clavicular, infraclavicular, median and basal. This is done for a hypothetical chest

image with the lung fields of the training images at their mean location, by computing horizontal lines that divide the lung fields in four parts of approximately area.

Except the first one, the other regions are again subdivided vertically. This corresponds to the medical representation of the lung field regions [18] (see figure below).

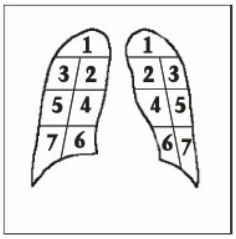

Figure 1. Lung fields divisions.

After segmentation, the region maps are warped to the segmentation result, using interpolation with radial basis functions [21]. After warping, the borders between the regions need no longer be horizontal or vertical.

Unlike us, Bram van Ginneken [5] divided the lung fields into 42 subregions. However, we consider his representation artificial. We preferred a division used by physicians in topographic anatomy [17].

Age: Some ages are bent for specific pulmonary diseases; other illnesses may appear at any age. Pneumococcal pneumonia appears after 1 year of age; bronchopneumonia is more frequent to children and old men; malignant tumors appear at older ages.

IV. CASE STUDY

Computational results will be illustrated with the data set for 120 patients with known diagnoses, confirmed by pathology tests. Twelve features for each patient were used in the computational study. The 120 patients records were checked for completeness and reduced to 50. Each of the 70 records rejected was missing at least one of the 12 features used in our study.

The selected features are listed next.

- **F1**: patients age
- F2: maximum radius (mm)
- **F3** shape: 1 = nodular, 2 = smoothly lobulated, 3 = segmentary, 4 = many small spiculations, 5 = large irregular spiculation
- **F4** calcification type: 0 = none, 1 = central calcification, 2 = laminated, 3 = dense
- **F5** lung: L = left lung, R = right lung, B = both lungs
- **F6:** location in thorax based on the segmentation shown in figure 1
- **F7** gender: M = male, F = female
- **F8** homogeneity: H = homogeneous, I = inhomogeneous
- **F9** oneness: U = unique, M = multiple
- **F10** edge: T = strong, S = soft
- **F11** surrounding tissues: A = atelectasis, NA = no affection
- **F12** evolution in time: 0 = no variations, 1 = radius/opacities decreasing, 2 = location variable in hours, 3 = location variable in days, 4 = radius constant in years, 5 = radius increasing
- **D**: diagnosis

The 50-patient data set is shown in the Appendix A .

A. Primary Algorithm Results

The primary decision-making algorithm uses decision rules extracted from the training data set. Numerous alternative rules (included in perspectives) have been generated with the rule extraction algorithm. Table I includes 9 decision rules generated with the rule extraction algorithm partially based on the concepts presented in [7].

TABLE I. DECISION RULES (PERSPECTIVE 1)

1.	IF F3 = 3 AND F6 = 1 THEN D = pneumococcal pneumonia (Patients 1, 43)
2.	IF F3 = 3 AND F4 = 0 THEN D = pneumococcal pneumonia (Patients 1, 4, 20, 32, 43, 44)
3.	IF F3 = 1 AND F4 = 0 AND F9 = M AND F11 = NA THEN D = bronchopneumonia (Patients 7, 14, 22, 39, 41, 46)
4.	IF (F6 = 2 OR F6 = 3) AND F9 = M AND F11 = NA THEN D = bronchopneumonia (Patients 7, 14, 22, 39, 41, 46)
5.	IF F3 = 1 AND F4 = 0 AND F8 = H AND F10 = T THEN D = undrained abscess (Patients 5, 33, 34, 47)
6.	IF F3 = 1 AND F11 = A THEN D = undrained abscess (Patients 5, 33)
7.	IF F1 ≤ 31 AND F3 = 1 AND F8 = H AND F9 = U AND F10 = T AND F11 = NA THEN D = benign tumor (Patients 18, 26, 36, 50)
8.	IF F1 ≤ 53 AND F2≥51 THEN D = malignant tumor (Patients 27, 31)
9.	IF F3 = 5 AND F5 = B THEN D = pulmonary edema (Patients 10, 16, 24, 28, 35, 49)

The rules in table I accurately describe the 50 patients. Each decision rule indicates the patients that it represents. Some patients are described by more that one decision rule, such as patients 1, 7, 14, 43 etc. To increase the value of the decision redundancy factor (DRF[1]), it is desirable that each object in the training set be represented by multiple rules. As the decision rules in table I ensure DRF = 0 for most rules, we call these decision rules Perspective 1 (the basic decision making perspective). Alternative decision-making perspectives, e.g. Perspective 2, will increase the value of DRF for the objects (patients) in the training set and the objects in the test data set.

TABLE II. DECISION RULES (PERSPECTIVE 2)

1.	IF F3 = 3 AND F8 = H AND F9=U AND F11 = NA THEN D = pneumococcal pneumonia (Patients 1, 4, 6, 20, 32, 43, 44)
2.	IF F3 = 5 AND F4 = 0 AND (F5 *6*= B) AND F8 = I AND F9 = U AND F10 = S AND F11 = NA THEN D = interstitial pneumonia (Patients 2, 11, 17, 21, 29, 38, 42, 45)
3.	IF F1 · 31 AND F2 · 42 AND F11 = NA THEN D = benign tumor (Patients 18, 26, 36, 50)
4.	IF F1 ¸ 53 AND F3 = 2 AND F7 = M AND F8 = I AND F11 = A THEN D = malignant tumor (Patients 27, 31)
5.	IF F3 = 5 AND F4 = 0 AND F5 = B AND F8 = I AND F10 = S THEN D = pulmonary edema (Patients 10, 16, 24, 28, 35, 49)

The rules in table II use features that partially overlap with the features used in Perspective 1 of table I. Mutually

[1] The DRF is the number of times an object can be independently represented with the reduced number of features minus one. For an object with single-feature reducts, *DRF* = *k* ¡1, where is the number of features included in the reduced objects. This measure will also reflect the user's confidence in predictions.

exclusive sets of features are certainly possible. The Perspective 2 rules increase DRF of individual patients.

To test the quality of the decision rules in tables I and II, the test set of 13 patients has been considered. These patients were not included in the test data set due to missing information. Additional testing was performed for ten randomly selected patients 2, 3, 14, 15, 24, 27, 28, 33, 42, and 44 from the 50-patient set. These patients were selected according to the cross-validation guidelines discussed in [24]. For each of the 49-patient data set, decision rules were derived and the patient deleted from the training set was tested. In all cases, the diagnosis produced by the primary decision-making algorithm agreed with the diagnosis provided by an invasive test.

B. Confirmation Algorithm Results

An interesting observation concerns the ability of each feature to uniquely represent objects (patients). A measure associated with this ability is called a classification quality ratio. For example, feature F1 (Patient's age) uniquely identifies 15% of all patients.

TABLE III. CLASSIFICATION QUALITY OF INDIVIDUAL FEATURES IN THE 50-PATIENT DATA SET

F1: 15%	F2: 14%	F3: 42%	F4: 100%	F5: 38%	F6: 50%
F7: 0%	F8: 50%	F9: 23%	F10: 24%	F11: 40	F12: 0%

The classification quality ratios in table III will have some impact on the selection of features to be used by the confirmation algorithm. To test the confirmation algorithm, we have randomly formed 10 feature sets, each with 2–-12 features. Some of these feature sets meet the definition of reduct and some are random modifications of the reducts. The last (10th) feature set includes all 12 features. Of course, all reducts and their supersets have the classification quality ratio of 100%, while some of patients were selected according to the cross-validation guidelines discussed in [24]. For each of the 49-patient data set, decision the feature sets have classification quality less than 100%. The selected feature sets and their classification quality are shown in table IV.

TABLE IV. CLASSIFICATION QUALITY OF FEATURE SETS

No.	Feature sets	Classification quality
1	(F1, F8)	60%
2	(F3, F6)	100%
3	(F1, F3, F8, F10)	89%
4	(F2, F3, F4, F6, F9, F11)	100%
5	(F4, F6, F8, F12)	88%
6	(F1, F3, F5, F7, F8, F9, F10, F12)	100%
7	(F2, F3, F5, F7, F8, F9, F10, F12)	100%
8	(F1, F5, F8, F10, F11)	92%
9	(F1, F2, F6, F7, F8, F10, F11, F12)	100%
10	(F1 – F12)	100%

To test the confirmation algorithm, a subset of 10 patients have been selected from the 50-patient data set. Each patient was removed, one at a time, from the 50 patient data set and the confirmation algorithm diagnosed this patient using the remaining 49-patient data set.

In addition to testing the ten randomly selected patients, we considered 13 additional patients.

C. Observations

The following observations can be made from the experiments:

- The lowest classification quality of the confirmation algorithm is 60%, which is the highest accuracy rate reported in clinical diagnosis [13].
- The highest classification quality of the confirmation algorithm is 100%.
- The feature sets which are not reducts have classification quality less than 100%. They are considered as "inferior" in the machine learning literature.
- The classification quality by the confirmation algorithm with all features is only 70%.
- Other feature sets ensuring 100% classification quality are likely to be found, however, the training set of 50 is too small for further generalizations.
- Some patients produced the largest number of errors in the algorithmic classification. They are the ones diagnosed with benign tumors, respectively malignant tumors, which cannot be entirely differentiated radiologically. For a certain diagnosis, a biopsy fragment should be taken.

Based on the training and test data sets used in this research, the classification quality by the combined primary and confirmation algorithm is 91.3% and the diagnostic accuracy is 100%. This means that 91.3% of all patients tested have been correctly diagnosed. The concepts presented needs further testing on larger and broader data sets.

V. DISCUSSIONS

None of the features may be considered important for all pulmonary diseases. However, certain features are significant for specific illnesses. To determine that, we carried out a factor analysis.

The table below summarizes the loadings of the features, respectively the accuracy of the classification for the whole set of patients.

TABLE V. CLASSIFICATION ACCURACY VS. FEATURES LOADINGS

Factor	Accuracy	Loading
F1	8%	7.2%
F2	7%	7.4%
F3	28%	25.3%
F4	62%	61.7%
F5	21%	23.1%
F6	38%	43.5%
F7	12%	15.3%
F8	45%	44.6%
F9	19%	21.2%
F10	18%	19.2%
F11	32%	30.1%
F12	14%	13.4%

Comparing the classification accuracy and the feature loadings, one may notice that they are quite close to each other, but globally no one influences the diagnosis in an important way.

Bram van Ginneken [5] classified the tumors as benign or malignant using the k-nearest neighborhood (kNN) algorithm.

The k-nearest neighbors of the feature vector are extracted from the training set, leaving out the feature vector to be classified, if necessary (which is easy to implement by simply ignoring neighbors at zero distance whenever they occur). Each neighbor votes for the region to be normal or abnormal. He used a fast algorithm for finding the k-nearest neighbors developed by Arya and Mount [2]. Instead of a binary normal/abnormal decision when classifying feature vectors, a probability measure that a region is abnormal is computed using weighted voting among the k-nearest neighbors.

The classification is a number in the range from 0 (normal) to 1 (abnormal). Given these classifications, region of interest (ROC) analysis can be performed.

Comparing the algorithm proposed by us with the one proposed by Ginneken, the former one is more explicit in the way that gives the possibility to explain the underlying relationships between the factors and the diagnosis. Moreover, the algorithm described here outputs not only benign/malignant (crisp) results, but more complex diagnoses, involving various possible pulmonary diseases.

VI. CONCLUSIONS

The research reported in this paper opens new avenues for medical decision-making. The proposed idea of combining different decision modes is novel and offers a viable concept for many applications. The primary decision-making and confirmation algorithms when combined generate decisions of high accuracy. The diagnosis by the two algorithms was of perfect accuracy for the clinical data reported in the paper. Additional developments of the algorithms and large-scale testing will be the ultimate proof of diagnostic accuracy for lung cancer and other diseases. The number of features necessary for high-accuracy autonomous diagnosis was smaller than in the original data set. This reduced number of features should lower testing costs. Because data from noninvasive tests were used for diagnosis, patients mortality and morbidity risks should be significantly reduced.

It would be a great interest the application of the proposed algorithm to larger databases of radiographs, involving more data mining techniques. On the other hand a combination of rough set theory and k-NN algorithm may lead to interesting results.

No	Age	Radius	Shape	Calcification	Lung	Location	Gender	Homogeneity	Oneness	Edge	Atelectasis	Evolution in time	Diagnosis
No	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	D
1	23	50	3	0	R	1	M	H	U	T	NA	1	pneumococcal pneumonia
2	65	150	5	0	R	2,4,5,6	M	I	U	S	NA	1	interstitial pneumonia
3	60	35	3	0	R	1	F	H	U	T	NA	1	secondary tuberculosis
4	13	45	3	0	L	5	F	H	U	T	NA	1	pneumococcal pneumonia
5	45	52	1	0	L	3	M	H	U	T	A	1	undrained abscess
6	55	60	2	2	R	1	M	H	U	T	NA	1	pneumococcal pneumonia
7	1	25	1	0	B	2,3,4,5	M	H	M	T	NA	3	bronchopneumonia
8	54	22	6	0	R	5	M	H	U	T	A	0	atelectasis
9	36	40	3	0	L	3	F	H	U	T	A	1	bacterial pneumonia
10	88	200	5	0	B	3,4,5,6,7	M	I	U	S	A	2	pulmonary edema
11	73	143	5	0	R	2,3,4	F	I	U	S	NA	1	interstitial pneumonia
12	4	56	2	1	R	3	M	H	U	T	NA	1	primary tuberculosis
13	68	75	2	0	R	1,3	F	I	M	S	NA	1	secondary tuberculosis
14	74	34	1	0	B	2,3,4,5	M	I	M	S	NA	3	bronchopneumonia
15	44	64	3	0	L	6,7	F	H	U	T	A	1	atelectasis
16	57	34	5	0	B	6, 7	F	I	U	S	NA	2	pulmonary edema
17	5	46	5	0	L	4, 5	M	I	U	S	NA	1	interstitial pneumonia
18	28	42	1	3	R	2	F	H	U	T	NA	4	benign tumor
19	48	46	2	0	L	7	M	H	U	S	NA	1	bacterial pneumonia
20	26	22	3	0	R	3	F	I	U	s	NA	1	pneumococcal pneumonia
21	3	49	5	0	R	2,4	M	I	U	S	NA	1	interstitial pneumonia
22	35	42	1	0	B	2,3,4,5,7	F	I	M	T	NA	3	bronchopneumonia
23	14	21	2	2	B	1	F	I	M	T	NA	1	primary tuberculosis
24	1	146	5	0	B	5,6,7	M	I	U	S	NA	2	pulmonary edema (congenital cardiopathy)
25	7	2	4	0	B	5,7	M	H	M	T	NA	1	primary tuberculosis (miliaria)
26	31	40	1	1	L	7	M	H	U	T	NA	4	benign tumor
27	60	65	2	0	R	4	M	I	U	T	A	5	malignant tumor
28	75	160	5	0	B	4,5,6,7	F	I	U	S	NA	2	pulmonary edema
29	26	80	5	0	R	5,7	F	I	U	S	NA	1	interstitial pneumonia
30	22	30	2	0	R	1	F	I	U	T	NA	1	primary tuberculosis
31	53	51	2	2	L	2	M	I	M	S	A	5	malignant tumor
32	8	25	3	0	R	3	M	H	U	T	NA	1	pneumococcal pneumonia
33	61	65	1	0	R	5	M	H	U	T	A	0	pulmonary abscess
34	72	40	1	0	R	3,5	M	H	M	T	NA	1	multiple pulmonary abscess
35	63	110	5	0	B	5,6,7	M	I	U	S	NA	2	pulmonary edema
36	27	25	1	3	R	2	F	H	U	T	NA	4	benign tumor
37	17	55	2	0	R	2	M	H	U	T	NA	1	primary tuberculosis
38	33	65	5	0	L	2,3	F	I	U	S	NA	1	interstitial pneumonia
39	40	33	1	0	R	1,2,3,5	M	I	M	S	NA	3	bronchopneumonia
40	57	38	2	2	L	3	M	I	U	T	NA	1	secondary tuberculosis
41	62	33	1	0	B	2,3,4,5	M	I	M	S	NA	3	bronchopneumonia
42	38	200	5	0	L	2,4,5	F	I	U	S	NA	1	interstitial pneumonia
43	20	48	3	0	R	1	M	H	U	T	NA	1	pneumococcal pneumonia
44	44	44	3	0	1	2,3	M	H	U	T	NA	1	pneumococcal pneumonia
45	42	123	5	0	L	4,5,6	M	I	U	S	NA	1	interstitial pneumonia
46	66	24	1	0	B	2,3,5	M	I	M	T	NA	3	bronchopneumonia
47	56	33	1	0	R	4	F	H	U	T	NA	1	pulmonary abscess
48	32	30	6	-	L	3,4	M	H	U	T	A	0	atelectasis
49	88	125	5	0	B	4,5,6,7	M	I	U	S	NA	2	pulmonary edema
50	19	24	1	3	L	4	M	H	U	T	NA	4	benign tumor

VII. APPENDIX A

The 50 patients whose characteristics have been used in this article are shown in the image below.

REFERENCES

[1] S.G. Armato, F. Li, M.L. Giger, H. MacMahon, S. Sone, and Doi K. Lung cancer: performance of automated lung nodule detection applied to cancers missed in a ct screening program. Radiology, (225):685–692, 2002.

[2] S. Arya and D.M. Mount. Approximate nearest neighbor queries inxed dimensions. In Proceedings of the 7th ACM-SIAM Symposium on Discrete Algorithms, pages 271–280, 2002.

[3] L. Chandra, S.R. Leclair, J.A. Meech, B. Varma, M. Smith, and B. Balachandran. Using rough sets as tools for knowledge discovery. In Proc. Australiasia-Pacific Forum on Intelligent Processing and Manufacturing of Materials, volume I, pages 663–667, 2002.

[4] C.H. Chen, L.F. Pau, and P.S.P. Wang. The Handbook of Pattern Recognition and Computer Vision (2nd Edition),. World Scientific Publishing Co., 2003.

[5] Bram van Ginneken. Computer-Aided Diagnosis in Chest Radiography. PhD thesis, Image Sciences Institute, 2004.

[6] J.M. Goo, Lee J.W., Lee H.J., Kim S., Kim J.H., and Im J.G. Automated lung nodule detection at low-dose ct: preliminary experience. Korean Jornal of Radiology, (4):211–216, 2003.

[7] J.M. Goo, T. Tongdee, R. Tongdee, K. Yeo, Hildebolt C.F., and Bae K.T. Volumetric measurement of synthetic lung nodules with multidetector row ct: effect of various image reconstruction parameters and segmentation thresholds on measurement accuracy. Radiology, (235):850–856, 2005.

[8] J. W. Grzymala-Busse. A new version of the rule induction system leers. Fundamenta Informaticae, (31):27–39, 2000.

[9] C.I. Henschke, D.F. Yankelevitz, D.P. Naidich, and Libby DM Mc-Cauley D.I., McGuinness G. Ct screening for lung cancer: suspiciousness of nodules according to size on baseline scans. Radiology, (231):164–168, 2004.

[10] B. Julesz, E. N. Gilbert, L. A. Shepp, and H. L. Frisch. Inability of humans to discriminate between visual textures that agree in second-order statistics - revisited. Perception, (2):391–405, 2000.

[11] I. Kononenko, I. Bratko, and M. Kokar. Machine Learning in Data Mining: Methods and Applications, chapter Application of machine learning to medical diagnosis, pages 389–428. Wiley, 2004.

[12] W. Kowalczyk and F. Slisser. Modeling customer retention with rough data models. In Proc. First Eur. Symp. PKDD '97, pages 4–13, 1997.

[13] Andrew Kusiak, Jeffrey A. Kern, Kemp H. Kernstine, and Bill T. L. Tseng. Autonomous decision-making: A data mining approach. IEEE Transactions on Information Technology in Biomedicine, 4(4), 2000.

[14] I.J. Lee, G. Gamsu, J. Czum, N. Wu, R. Johnson, and S. Chakrapani. Lung nodule detection on chest ct: evaluation of a computer aided detection (cad) system. Korean Jornal of Radiology, (6):89–93, 2005.

[15] Z. Pawlak. Rough Sets: Theoretical Aspects of Reasoning About Data. MA: Kluwer, 1998.

[16] Z. Pawlak. Cross-validatory choice and assessment of statistical predictions MA: Kluwer, 2000.

[17] [Werner Platzer. Pernkopf Anatomy: Atlas of Topographic and Applied Human Anatomy : Thorax, Abdomen and Extremities. Urban and Schwarzenberg; 3rd edition, 1998.

[18] Charles Putman. Diagnostic Imaging of the Lung (Lung Biology in Health and Disease). Marcel Dekker Inc., 2001.

[19] G.D. Rubin, J.K. Lyo, D.S. Paik, A.J. Sherbondy, L.C. Chow, and A.N. Leung. Pulmonary nodules on multi-detector row ct scans: performance comparison of radiologists and computer-aided detection. Radiology, (234):274–283, 2005.

[20] G. Ruhe. Qualitative analysis of software engineering data using rough sets. In Proc. Fourth Int. Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery, pages 292–299, 1999.

[21] D. Ruprecht and H. Mueller. Image warping with scattered data interpolation. IEEE Computer Graphics and Applications, (2):37–46, 2003.

[22] A. Skowron. Data filtration: A rough set approach. In Proc. Int. Workshop on Rough Sets and Knowledge Discovery, pages 108–118, 2000.

[23] J. Stefanowski and K. Slowinski. Rough Sets and Data Mining: Analysis and Imprecise Data, chapter Rough sets as a tool for studying attribute dependencies in the urinary stones treatment data set, pages 177–196. Kluwer, 2003.

[24] M. Stone. Cross-validatory choice and assessment of statistical predictions. Journal of Royal Statistics Society, (36):11–47, 1998.

[25] S. Tsumoto. Extraction of experts decision process from clinical databases using rough set model. In Proc. First Eur. Symp. PKDD 97, pages 58–67, 1997.

[26] Z. M. Wojcik. Edge detector free of the detection/localization tradeoff using rough sets. In Proc. Int. Workshop on Rough Sets and Knowledge Discovery, pages 421–438, 2001.