

# Autonomous Decision-Making: A Data Mining Approach

Andrew Kusiak, Jeffrey A. Kern, Kemp H. Kernstine, and Bill T. L. Tseng

**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. Data mining 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. The approach has been tested on a medical data set for patients with lung abnormalities referred to as solitary pulmonary nodules (SPNs). 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**—Data mining, lung cancer diagnosis, machine learning, medical decision making, rough set theory.

## 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).

Manuscript received September 24, 1999; revised, April 10, 2000.

A. Kusiak and B. T. L. Tseng are with the Intelligent Systems Laboratory, 4312 Seamans Center for the Engineering Arts and Sciences, The University of Iowa, Iowa City, IA 52242-1527 USA.

J. A. Kern is with the Department of Internal Medicine, The University of Iowa, Iowa City, IA 52242-1527 USA.

K. H. Kernstine is with the Department of Surgery, The University of Iowa, Iowa City, IA 52242-1527 USA.

Publisher Item Identifier S 1089-7771(00)08189-9.

- 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 [12] 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 [13]. An important aspect of the approach proposed in this paper is that the decisions (diagnoses) are  $(100 - \epsilon)\%$  accurate for  $(100 - \delta)\%$  of objects with unknown outcomes, where  $\epsilon$  and possibly  $\delta$  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 solitary pulmonary nodules (SPN's) using information from noninvasive tests. An SPN is a lung abnormality that could potentially become cancerous. The

new approach significantly reduces patients' risks and costs. In a typical SPN disease occurrence scenario, a nodule is detected on a patient's chest radiograph. As this SPN may be either benign or malignant, further testing is required to determine its exact nature. The diagnosis is perceived to depend on many features, such as the SPN diameter, border character, presence of calcification, patient's age, smoking history, results of CT densitometry, and overall prevalence of malignancy within the population [11]. Multiple medical disciplines are involved in collecting a large volume of clinical data at different times and locations, with varying accuracy and consistency. Therefore, an approach that fuses information from different sources and intelligently processes large volumes of data is needed.

The research presented in this paper shows that the number of features (results of noninvasive tests, patient's data, etc.) necessary to diagnose an SPN is smaller than the number used in current medical practice. At the same time, the decision-making accuracy is significantly improved.

## II. LITERATURE SURVEY

The researchers and practitioners of today create formulas, models, algorithms, and other constructs defined in abstract spaces. In the world of tomorrow, large volumes of symbolic and numeric data will define spaces for processing by the new and existing tools. Data mining 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. The growing volume of data that is available in a digital form spurs this accelerated interest.

One of the few theories developed specifically for data mining is the rough set theory [12]. It has found applications in industry, service organizations, healthcare [8], software engineering [15], edge detection [24], data filtration [17], and clinical decision-making [19], [22].

A comprehensive comparative analysis of prediction methods included in [7] 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 SPN patients.

### A. Background

To date, numerous models and algorithms have been developed for decision-making using medical data (called here features). The existing approaches share a common thread—the outcome is predicted with an error. Such form of an outcome is not acceptable in many applications. For example, a patient with an SPN would like to be either certain whether the nodule is benign or malignant or be told that accurate diagnosis cannot be made without additional tests. These additional tests (e.g., lung biopsy) are often invasive and involve higher risks to the

TABLE I  
FIVE-OBJECT DATA SET

Object No.	F1	F2	F3	F4	D
1	0	1	0	2	0
2	1	1	0	2	2
3	0	0	0	1	0
4	0	1	1	0	1
5	0	0	1	3	0

patient and are expensive. The patient and physician would feel more comfortable with an algorithm-generated outcome if the autonomous decision could be confirmed in more than one way. The proposed algorithms are designed to make predictions of high accuracy for a large percentage of cases being diagnosed. For cases in which an accurate decision could not be automatically generated, other decision-making modalities could be used, e.g., higher-level prediction systems or humans.

The basic construct of rough set theory is called a reduct [12]. The reduct can be loosely defined as a minimal subset of features uniquely identifying all objects (patients) in a data set (see [16] for a formal definition of the reduct).

By definition, reduct expresses an alternative and simplified way of representing a set of objects (patients). It is easy to see that reduct has the same properties as a key defined in relational database theory (with respect to a specific instance of a relation only). In this context, reduct can be considered an empirical key.

The term "reduct" was originally defined for sets rather than objects (patients) with input and output features or decision tables with features (tests) and decisions (diagnoses). Reducts of the objects in a decision table have to be computed with the consideration given to the value of the decision (diagnosis). In this paper, each reduct is viewed from two perspectives: feature and decision rule perspectives.

To illustrate the reduct, consider the data set in Table I for four features (F1–F4) and a three-level decision ( $D = 0, 1, \text{ and } 2$ ).

The meaning of the two perspectives is illustrated with the reduct represented as

$$1xxx2 \quad (1)$$

for object 2 of Table I. The first four elements in this reduct represent the features and the value 2 represents the decision (diagnosis). The value of feature F1 is 1, and this reduct is referred to as a single-feature reduct as it contains only one feature. The features F2–F4 marked with "x" are not included in the reduct. As this reduct has been derived from object 2, it is referred to as an o-reduct (object-reduct). The reduct in (1) can be also expressed as the following decision rule:

$$\text{IF the value of feature F1} = 1 \text{ THEN the decision } D = 2 \quad (2)$$

and therefore it is called an r-reduct (rule-reduct).

The reduct generation algorithm presented in [9] produces single- and multi-feature reducts. In [3], an algorithm was developed for extraction of decision rules from data sets—a concept complementary to the reduct generation.

The reduct generation algorithm essentially applies the reduct definition to each object. The entry 'x' in each reduct [see (1)]

TABLE II  
THE MINIMAL SET OF FEATURES

Object No.	o-Reduct No.	F1	F2	F3	F4	D
1	1	0	x	0	x	0
	2	0	x	x	2	0
3	3	x	0	x	x	0
	4	x	x	x	1	0
5	5	x	0	x	x	0
	6	x	x	x	3	0
4	7	x	x	x	0	1
2	8	1	x	x	x	2

TABLE III  
ALTERNATIVE SOLUTION

Object No.	o-Reduct-No.	F1	F2	F3	F4	D
1	2	0	x	x	2	0
4	7	x	x	x	0	1
3,5	3,5	x	0	x	x	0
2	8	1	x	x	0	2

implies that the corresponding feature is not considered in determining the object's output. To obtain a reduct, one input feature at a time is considered and it is determined whether this feature uniquely identifies the corresponding object. A conflict in the decision of any two objects disqualifies that feature from producing a single-feature reduct.

One of the objectives of feature extraction is to select reducts with the minimum number of features representing all original objects (patients). An example of a minimal set of features is shown in Table II. Note that only two input features F1 and F4 (reduct F1xF4 or, for short, F1, F4) out of four are needed for unambiguous representation of all objects. An alternative solution is presented in Table III where three (F1, F2, and F3) out of four features are selected. Table III contains the merged o-reduct (3, 5) that has been obtained from two identical reducts 3 and 5. Tables II and III are called decision tables.

The results in Table III can be expressed as decision rules, e.g., the decision rule (r-reduct) corresponding to object 2 is

$$\text{IF feature } F1 = 1 \text{ THEN decision } D = 2. \quad (3)$$

The two tables illustrate the data reduction aspect of data mining, specifically feature reduction as well as rule (r-reduct) reduction. The power of data reduction becomes apparent for data sets with large numbers of objects and features.

In order to express the robustness of representing objects with the selected features, a measure called decision redundancy factor (DRF) is introduced. 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, the DRF =  $k - 1$ , where  $k$  is the number of features included in the reduced objects. This measure will also reflect the user's confidence in predictions [26].

The DRF is explained with the data set in Table IV, which was transformed by the feature extraction algorithm in the form shown in Table V.

It is clearly visible in Table V that column F2 can be deleted (all entries are "x"), as it does not add value to the proper clas-

TABLE IV  
TEST DATA

Object No.	F1	F2	F3	F4	D
1	1	0	0	2	2
2	0	1	1	0	1
3	1	1	0	2	2
4	1	2	0	2	2
5	0	2	1	0	1
6	0	0	1	0	1
7	1	3	0	2	2
8	0	3	1	0	1

TABLE V  
THE DECISION TABLE FOR THE DATA IN TABLE IV

Object No.	F1	F2	F3	F4	D
2, 5, 6, 8	0	x	1	0	1
1, 3, 4, 7	1	x	0	2	2

sification of objects. In fact, the decision  $D$  could be uniquely identified based on each of the three features F1, F3, and F4 in Table V. The eight rows in Table V merge into two, thus significantly reducing the number of rules. Since we have single-feature reducts only, the number of input features in a row that are different than "x" indicate the number of times an object can be uniquely identified. In other words, the DRF for the objects in Table V is  $3 - 1 = 2$ .

### III. ALGORITHM DEVELOPMENT

The feature extraction algorithm can generate multiple feature sets (reducts). These feature sets are used for predicting an object's outcome with the primary and confirmation algorithms.

The primary decision-making algorithm compares the feature values of a new object with the features of decision rules. When the matching criterion is met, the object's decision is assigned equal to that of the matching decision rule, and the confirmation algorithm is invoked. Only when the decisions reached by the two algorithms are identical may we say that a final decision has been made.

#### A. The Primary Decision-Making Algorithm

The steps of the primary decision-making algorithm are as follows.

- Step 1) Match the new object's feature values with the features of the alternative decision rules generated by the rule extraction algorithm. If the match is satisfactory, go to Step 2; otherwise, go to Step 3.
- Step 2) Assign the primary decision to the new object equal to the decision associated with the matching decision rules and go to Step 4.
- Step 3) Output "No primary decision is made—More feature values are needed" and go to Step 4.
- Step 4) If all new objects have been considered, stop; otherwise, go to Step 1.

The alternative rules (perspectives) used in Step 1 are to increase the value of DRF. However, users may not be satisfied with the outcome generated by the primary decision-making algorithm, regardless of the value of DRF. Common reasons for

this lack of confidence are errors in the data, missing data, and so on. To overcome this application barrier, an “orthogonal” decision is generated with the confirmation algorithm presented in the next section.

### B. The Confirmation Algorithm

To present the confirmation algorithm, absolute distance measure  $d_{ij}$  between objects  $i$  and  $j$  defined

$$d_{ij} = \sum_{k=1}^n |f_{ik} - f_{jk}| \quad (4)$$

where  $f_{ik}$  is the value of feature  $Fk$  for object  $i$  and  $n$  is the number of features in a feature set.

The steps of the confirmation algorithm are as follows.

- Step 0) Define proper feature sets.  
For each proper feature set:
- Step 1) Cluster objects with equal outcome in groups.
- Step 2) Compute distance  $d_{ij}$  between a new object and every object in each group of Step 1.
- Step 3) For each group, compute the average distance of the distances  $d_{ij}$  obtained in Step 2.
- Step 4) Assign the new object a decision corresponding to the cluster with the minimum average distance.
- Step 5) Continue until all new objects have been considered.

Step 0 is key to getting high-accuracy results generated by the algorithm. The computational experience reported in Section IV-C proves that some feature sets ensure high-accuracy decisions. We call these feature sets proper. The proper feature sets are formed through experimentation by beginning with a small set of features, often with low individual classification quality, and adding new features until acceptable classification accuracy is achieved.

The accuracy of the results generated by the two algorithms was tested using a medical data set for SPN patients collected at the University of Iowa Hospitals and Clinics.

## IV. SPN CASE STUDY

Computational results will be illustrated with the data set collected at the University of Iowa Hospital and Clinics for 118 SPN patients with known diagnoses, confirmed by pathology tests. Eighteen randomly selected features for each patient were used in the computational study. The 118 patients’ records were checked for completeness and reduced to 50. Each of the 68 records rejected was missing at least one of the 18 features used in our study.

The selected 18 features are listed next.

- F1 Age (Patient’s age)
- F2 CT max r (Computed Tomography maximum radius, computed as  $\max\{\text{Ctr1}, \text{Ctr2}\}$ )
- F3 CT max area (Computed Tomography maximum area)
- F4 Borders (1 = sharp and smooth, 2 = smoothly lobulated, 3 = large irregular spiculation, 4 = many small spiculations)
- F5 Calcification type (1 = central calcification, 2 = laminated, 3 = dense)

- F6 Location in thorax (1 = central, 2 = mediastinal, 3 = peripheral)
- F7 Nodes (0 = none, 1 = less than 1 cm, 2 = larger than 1 cm without calcification)
- F8 Other sus lesions, 0 = No, 1 = Yes (Other suspected lesions)
- F9 1 = M, 0 = F (1 = Male, 0 = Female)
- F10 PET PN image > bg, 1 = Yes, 2 = No (Positron Emission tomography Pulmonary Nodule image is greater than background)
- F11 Pk/yrs (packet - years)
- F12 BMI (Body Mass Index)
- F13 hx of ca, Yes = 1, No = 0 (history of cancer)
- F14 FEV<sub>1%</sub> (Forced Expiratory Volume, 1 sec, percent predicted)
- F15 DLCO% ADJ (Adjusted DLCO, percent predicted)
- F16 FVC% (Forced Vital Capacity, percent predicted)
- F17 FEF<sub>25-75%</sub> (Forced Expiratory Flow, 25–75%)
- F18 SUV (Positron Emission Tomography Standard Uptake Value)

D Diagnosis (M = Malignant, B = Benign)  
The 50-patient data set is shown in Table XIII.

### A. Case Study Description

A solitary pulmonary nodule (SPN) is defined as a discrete pulmonary opacity less than 3 cm in size that is surrounded by normal lung tissue and is not associated with atelectasis or adenopathy [4], [2]. Approximately 150 000 SPNs are identified in the U.S. per year [21]. Benign processes, such as inflammation, fibrosis, or granulomas can manifest as SPNs. Some 40%–50% of SPNs are malignant [10]. Therefore, an early and correct diagnosis is critical for optimal therapy of malignant SPNs and avoidance of unnecessary diagnostic tests, treatment costs, and risks for benign SPNs. In the current medical practice, more than one hundred data points from noninvasive tests might be considered during the diagnosis of an SPN. For an individual patient, it is almost impossible for a human decision-maker to determine which of these data points are critical for the correct diagnosis. Physicians have difficulty in differentiating benign from malignant cases based on clinical features, including symptoms, physical examination, and laboratory results [6]. Several diagnostic procedures, including chest radiography, chest computed tomography (CT), and bronchoscopy are used to differentiate benign from malignant nodules. Unfortunately, these features have poor sensitivity and specificity [10]. Even the choice of diagnostic options is complicated by variances in test effectiveness, cost, accuracy, and patient risks. Considering the uncertainty and the probability of malignancy, biopsies are often performed on SPNs. However, approximately 50%–60% of SPNs are not malignant and could be monitored clinically and radiographically [5], [4]. Hubner *et al.* [5] conject that appropriate follow-up by CT scan could reduce biopsies or resections of benign nodules by 20%–40%. The exact number of cases that go to surgical resection is not reported. To estimate potential savings from implementing the algorithms, it is conservatively estimated that 50% of the cases proceed to a surgical resection to verify the diagnosis, yet only 40% of

TABLE VI  
PERSPECTIVE 1 DECISION RULES

---

Decision rule 1. IF (F10 <= 1) AND (F17 <= 89) THEN (D = B); [Patients 3, 4, 26, 40, 46]
Decision rule 2. IF (F2 in [0.9, 1.55]) AND (F18 in [2.75, 8.25]) FND (F6 >= 2) FND (F5 <= 0) THEN (D = B); [Patients 19, 33, 43]
Decision rule 3. IF (F14 >= 78) AND (F18 <= 4.1) AND (F17 <= 54) THEN (D = B); [Patients 3, 26, 43, 47]
Decision rule 4. IF (F14 in [94, 109]) AND (F16 >= 103) THEN (D = B); [Patients 9, 33, 40]
Decision rule 5. IF (F6 <= 2) FND (F10 >= 1) FND (F16 <= 94) THEN (D = M); [Patients 7, 8, 13, 14, 15, 18, 20, 22, 27, 28, 30, 31, 32, 34, 38, 48]
Decision rule 6. IF (F2 <= 0.9) AND (F10 >= 1) THEN (D = M); [Patient 16]
Decision rule 7. IF (F10 >= 1) AND (F14 <= 74) AND (F13 <= 0) THEN (D = M); [Patients 5, 7, 10, 18, 22, 23, 25, 31, 32, 35, 45]
Decision rule 8. IF (F15 <= 52) THEN (D = M); [Patients 7, 11, 31, 37, 45]
Decision rule 9. IF (F14 <= 80) AND (F17 >= 31) THEN (D = M); [Patients {5, 14, 23, 28, 34, 39, 42, 50}]
Decision rule 10. IF (F14 in [81, 87]) THEN (D = M); [Patients 6, 38]
Decision rule 11. IF (F6 <= 2) AND (F10 >= 1) AND (F14 >= 84) AND (F15 >= 63) THEN (D = M); [Patients 1, 6, 8, 20, 21, 30, 38, 44, 48, 49]
Decision rule 12. IF (F16 <= 93) AND (F17 >= 49) THEN (D = M); [Patients 2, 8, 12, 15, 20, 23, 28, 34, 36, 38]
Decision rule 13. IF (F6 >= 2) AND (F14 >= 91) AND (F15 <= 97) THEN (D = M); [Patients 11, 24, 29]
Decision rule 14. IF (F2 >= 1.65) AND (F15 >= 116) THEN (D = M); [Patients 17, 41, 49]

---

TABLE VII  
PERSPECTIVE 2 DECISION RULES

---

Decision rule 1. IF (F8 <= 0) AND (F10 <= 1) THEN (D = B); [Patients 40, 46]
Decision rule 2. IF (F8 >= 0) AND (F9 <= 0) AND (F18 <= 4.1) THEN (D = B); [Patients 26, 43, 47]
Decision rule 3. IF (F2 in [0.9, 1.1]) AND (F4 >= 2) AND (F18 <= 8.25) THEN (D = B); [Patients 33, 43]
Decision rule 4. IF (F2 in [1.1, 1.55]) AND (F6 >= 2) AND (F8 <= 0) THEN (D = B); [Patients 3, 19]
Decision rule 5. IF (F2 in [2.25, 2.75]) AND (F4 <= 1) THEN (D = B); [Patient 9]
Decision rule 6. IF (F2 <= 2.1) AND (F6 <= 2) AND (F10 >= 1) THEN (D = M); [Patients 1, 8, 10, 13, 14, 16, 18, 21, 22, 30, 31, 32, 49]
Decision rule 7. IF (F2 >= 2.75) AND (F18 <= 8.6) THEN (D = M); [Patients 11, 15, 20, 27, 34, 36, 37, 38, 44]
Decision rule 8. IF (F2 <= 2.75) AND (F13 <= 0) AND (F18 >= 5.3) THEN (D = M); [Patients 1, 18, 23, 25, 29, 30, 32, 39, 41, 45]
Decision rule 9. IF (F2 <= 3.25) AND (F18 >= 8.45) THEN (D = M); [Patients 1, 2, 6, 8, 12, 13, 18, 23, 25, 29, 39, 45, 48]
Decision rule 10. IF (F2 >= 3.75) THEN (D = M); [Patients 7, 15, 17, 28, 34, 35, 37, 38]
Decision rule 11. IF (F3 <= 6) AND (F4 <= 1) AND (F6 >= 2) THEN (D = M); [Patients 2, 24, 29, 50]
Decision rule 12. IF (F7 >= 1) AND (F8 <= 0) THEN (D = M); [Patients 5, 8, 16, 29, 36, 42]

---

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

---

F1 = 70%, F2 = 30%, F3 = 56%, F4 = 0%, F5 = 0%, F6 = 16%, F7 = 6%, F8 = 0%, F9 = 0%, F10 = 10%, F11 = 62%, F12 = 56%, F13 = 0%, F14 = 76%, F15 = 86%, F16 = 88%, F17 = 82%, F18 = 60%.
--

---

TABLE IX  
CLASSIFICATION QUALITY OF FEATURE SETS

Feature Set	Classification Quality
Feature Set I (F15, F16)	100%
Feature Set II (F1, F13)	100%
Feature Set III (F1, F3, F15, F16)	100%
Feature Set IV (F1, F3, F11, F12, F15, F16, F18)	100%
Feature Set V (F4, F8, F9, F13)	68%
Feature Set VI (F4, F7, F8, F9, F10, F13)	94%
Feature Set VII (F4, F5, F6, F7, F8, F9, F10, F13)	100%
Feature Set VIII (F1, F7, F10, F12, F18)	100%
Feature Set IX (F2, F4, F6, F7, F8, F9, F10, F13)	100%
Feature Set X (F1 through F18)	100%

these cases are malignant [10]. The cost of diagnosis and treatment of a patient with SPN may exceed \$30 000 [25], and the patient is exposed to surgical risks. Thus, the costs of making benign diagnoses are in excess of \$900 million per year in

hospital charges alone. The costs to society become larger if we consider patient productivity losses and costs to insurers. Therefore, there is a need to improve the diagnostic accuracy of SPNs from patient, physician, and society perspectives.

### B. 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 VI includes 14 decision rules generated with the rule extraction algorithm partially based on the concepts presented in [3].

The rules in Table VI accurately describe the 50 patients in the training data set of Table XIII. Each decision rule indicates the patients that it represents, e.g., rule 1 describes patients 3, 4, 26, 40, and 46. Some patients are described by more than one decision rule, e.g., patient 49 is included in decision rules 11 and

TABLE X  
RESULTS PRODUCED BY THE CONFIRMATION ALGORITHM FOR TEN PATIENTS FROM THE TRAINING DATA SET (DIAGNOSED PATIENTS 2, 14, 15, 24, 27, 28, 42, 44, 3 AND 33)

Feature Set	I (F15, F16)			II (F1, F3)			III (F1, F3, F15, F16)			IV (F1, F3, F11, F12, F15, F16, F18)		
	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}
Patient 2 (D = M)	1392/39 (35.69)	367/10 (36.7)	1	986.4/39 (25.29)	200.08/10 (20.01)	0	2953.4/39 (75.73)	591.8/10 (59.18)	0	5341.05/39 (136.95)	993.7/10 (99.37)	0
Patient 14 (D = M)	1597/39 (40.95)	635/10 (63.5)	1	928.8/39 (23.82)	181.8/10 (18.18)	0	2525.8/39 (64.76)	806.8/10 (80.68)	1	4162.25/39 (106.72)	1283.1/10 (128.31)	1
Patient 15 (D = M)	1491/39 (38.23)	579/10 (57.9)	1	1681.2/39 (43.11)	588.8/10 (58.88)	1	3172.2/39 (72.73)	1157.8/10 (115.8)	1	6018.65/39 (154.32)	2102.1/10 (210.21)	1
Patient 24 (D = M)	1829/39 (46.90)	435/10 (43.5)	0	1007.4/39 (25.83)	165.8/10 (16.58)	0	2836.4/39 (72.73)	590.8/10 (59.08)	0	5216.45/39 (133.76)	909.3/10 (90.93)	0
Patient 27 (D = M)	1973/39 (50.59)	729/10 (72.9)	1	1176.8/39 (30.17)	361.8/10 (36.18)	1	3149.8/39 (80.76)	1080.8/10 (108.08)	1	4771.85/39 (122.36)	1423.5/10 (142.35)	1
Patient 28 (D = M)	1979/39 (50.74)	741/10 (74.1)	1	2788.4/39 (71.50)	814.8/10 (81.48)	1	4767.4/39 (112.24)	1545.8/10 (154.58)	1	7102.45/39 (182.11)	1914.7/10 (191.47)	1
Patient 42 (D = M)	2133/39 (54.69)	769/10 (76.9)	1	966.8/39 (24.79)	291.8/10 (29.18)	1	3099.8/39 (79.48)	1050.8/10 (105.08)	1	4698.65/39 (120.48)	1316.5/10 (131.65)	1
Patient 44 (D = M)	2235/39 (57.31)	407/10 (40.7)	0	1470/39 (37.69)	325.8/10 (32.58)	0	3705/39 (95)	722.8/10 (72.28)	0	5218.65/39 (133.81)	1037.1/10 (103.71)	0
Patient 3 (D = B)	3903/40 (97.58)	1101/9 (122.33)	1	1232.6/40 (30.82)	91.8/9 (10.2)	0	5095.6/40 (127.39)	1192.8/9 (132.53)	1	7017.75/40 (175.44)	1420.9/9 (157.88)	0
Patient 33 (D = B)	2819/40 (70.48)	515/9 (57.22)	0	1081.4/40 (27.04)	119.8/9 (13.31)	0	3860.4/40 (96.51)	634.8/9 (70.53)	0	6225.25/40 (155.63)	994.7/9 (110.52)	0
<b>C-Quality</b>			<b>70%</b>			<b>60%</b>			<b>60%</b>			<b>70%</b>

(a)

Feature Set	V (F4, F8, F9, F13)			VI (F4, F7, F8, F9, F10, F13)			VII (F4, F5, F6, F7, F8, F9, F10, F13)			VIII (F1, F7, F10, F12, F18)		
	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}
Patient 2 (D = M)	118/39 (3.03)	38/10 (3.8)	1	159/39 (4.08)	53/10 (5.3)	1	195/39 (5)	57/10 (5.7)	1	962.65/39 (24.68)	374.9/10 (37.49)	1
Patient 14 (D = M)	100/39 (2.56)	34/10 (3.4)	1	161/39 (4.13)	55/10 (5.5)	1	189/39 (4.85)	61/10 (6.1)	1	789.45/39 (20.24)	203.3/10 (20.33)	1
Patient 15 (D = M)	78/39 (2)	30/10 (3)	1	119/39 (3.05)	45/10 (4.5)	1	171/39 (4.38)	61/10 (6.1)	1	1007.45/39 (25.83)	329.3/10 (32.93)	1
Patient 24 (D = M)	96/39 (2.46)	34/10 (3.4)	1	133/39 (3.41)	45/10 (4.5)	1	169/39 (4.33)	49/10 (4.9)	1	765.05/39 (19.62)	239.5/10 (23.95)	1
Patient 27 (D = M)	106/39 (2.72)	36/10 (3.6)	1	143/39 (3.67)	47/10 (4.7)	1	171/39 (4.38)	53/10 (5.3)	1	1023.05/39 (26.23)	371.7/10 (37.17)	1
Patient 28 (D = M)	90/39 (2.31)	32/10 (3.2)	1	127/39 (3.26)	43/10 (4.3)	1	179/39 (4.59)	59/10 (5.9)	1	818.05/39 (20.98)	241.9/10 (24.19)	1
Patient 42 (D = M)	106/39 (2.72)	36/10 (3.6)	1	167/39 (4.28)	57/10 (5.7)	1	203/39 (5.21)	61/10 (6.1)	1	773.85/39 (19.84)	264.7/10 (26.47)	1
Patient 44 (D = M)	92/39 (2.36)	34/10 (3.4)	1	129/39 (3.31)	45/10 (4.5)	1	157/39 (4.03)	51/10 (5.1)	1	1030.65/39 (26.43)	201.3/10 (20.13)	0
Patient 3 (D = B)	134/40 (3.35)	24/9 (2.67)	0	211/40 (5.28)	35/9 (3.89)	0	247/40 (6.18)	39/9 (4.33)	0	1141.15/40 (28.53)	145.1/9 (16.12)	0
Patient 33 (D = B)	156/40 (3.9)	28/9 (3.11)	0	193/40 (4.83)	39/9 (4.33)	0	29/40 (72.5%)	5/9 (55.56)	0	823.85/40 (20.60)	234.9/9 (26.1)	1
<b>C-Quality</b>			<b>100%</b>			<b>100%</b>			<b>100%</b>			<b>80%</b>

(b)

14. To increase the value of the DRF, it is desirable that each object in the training set be represented by multiple rules. As the decision rules in Table VI 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, Perspective 3, and so on, will increase the value of DRF for the objects (patients) in the training set and the objects in the test data set.

The rules in Table VII use features that partially overlap with the features used in Perspective 1 of Table VI. Mutually exclusive sets of features are certainly possible.

The Perspective 2 rules in Table VII increase DRF of individual patients. For example, the previously mentioned patient 49 that was described with rules 11 and 14 of Perspective 1 is also represented with rule 6 of Perspective 2.

To test the quality of the decision rules in Tables VI and VII, the test set of 13 patients has been considered (see Table XIV). These patients were not included in the test data set due to missing information. The decision rules of Tables VI and VII generated diagnoses for all 13 patients that have agreed with the diagnoses shown in Table XIV. Additional testing was performed for ten randomly selected patients 2, 3, 14, 15, 24, 27,

TABLE X (Continued)  
RESULTS PRODUCED BY THE CONFIRMATION ALGORITHM FOR TEN PATIENTS FROM THE TRAINING DATA SET (DIAGNOSED PATIENTS 2, 14, 15, 24, 27, 28, 42, 44, 3 AND 33)

Feature Set	IX (F2, F4, F6, F7, F8, F9, F10, F13)			X (F1 – F18)		
	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}
Patient 2 (D = M)	239.1/39 (6.13)	61.75/10 (6.18)	1	8163.15/39 (209.3)	1612.45/10 (161.25)	0
Patient 14 (D = M)	226.3/39 (5.8)	68.75/10 (6.88)	1	6151.35/39 (157.73)	1770.85/10 (177.09)	1
Patient 15 (D = M)	238.9/39 (6.13)	87.35/10 (8.74)	1	8256.25/39 (211.70)	2638.45/10 (263.8)	1
Patient 24 (D = M)	224.5/39 (5.76)	53.15/10 (5.32)	0	7977.95/39 (204.56)	1507.45/10 (150.75)	0
Patient 27 (D = M)	216.3/39 (5.55)	69.35/10 (6.94)	1	7627.15/39 (195.56)	2195.85/10 (219.59)	1
Patient 28 (D = M)	278.9/39 (7.15)	95.35/10 (9.54)	1	9352.35/39 (239.80)	2507.05/10 (250.71)	1
Patient 42 (D = M)	243.3/39 (6.24)	72.35/10 (7.24)	1	7046.95/39 (180.69)	1925.85/10 (192.5)	1
Patient 44 (D = M)	206.3/39 (5.29)	69.35/10 (6.94)	1	7199.95/39 (184.61)	1475.45/10 (147.5)	0
Patient 3 (D = B)	297.7/40 (7.44)	43.15/9 (4.79)	0	9038.45/40 (225.96)	1839.05/9 (204.34)	0
Patient 33 (D = B)	284.5/40 (7.11)	47.15/9 (5.24)	0	8516.75/40 (212.92)	1452.85/9 (161.42)	0
<b>C-Quality</b>			<b>90%</b>			<b>70%</b>

(c)

TABLE XI  
RESULTS PRODUCED BY THE CONFIRMATION ALGORITHM FOR 13 TEST SET PATIENTS

Feature Set	V (F4, F8, F9, F13)			VI (F4, F7, F8, F9, F10, F13)			VII (F4, F5, F6, F7, F8, F9, F10, F13)		
	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}	D = M	D = B	Diagnosis: Min {(D = M), (D = B)}
Patient 51 (D = M)	132/40 (3.3)	44/10 (4.4)	1	173/40 (4.33)	59/10 (5.9)	1	209/40 (5.23)	63/10 (6.3)	1
Patient 52 (D = M)	108/40 (2.7)	38/10 (3.8)	1	185/40 (4.63)	49/10 (4.9)	1	221/40 (5.53)	53/10 (5.3)	0
Patient 53 (D = M)	122/40 (3.05)	42/10 (4.2)	1	163/40 (4.08)	57/10 (5.7)	1	191/40 (4.78)	63/10 (6.3)	1
Patient 54 (D = M)	108/40 (2.7)	36/10 (3.6)	1	169/40 (4.23)	57/10 (5.7)	1	241/40 (6.03)	71/10 (7.1)	1
Patient 55 (D = B)	146/40 (3.65)	26/10 (2.6)	0	223/40 (5.58)	37/10 (3.7)	0	259/40 (6.48)	41/10 (4.1)	1
Patient 56 (D = B)	128/40 (3.2)	22/10 (2.2)	0	165/40 (4.13)	33/10 (3.3)	0	201/40 (5.03)	37/10 (3.7)	1
Patient 57 (D = M)	108/40 (2.7)	36/10 (3.6)	1	149/40 (3.72)	51/10 (5.1)	1	185/40 (4.63)	55/10 (5.5)	1
Patient 58 (D = M)	102/40 (2.55)	36/10 (3.6)	1	143/40 (3.58)	51/10 (5.1)	1	171/40 (4.28)	57/10 (5.7)	1
Patient 59 (D = M)	80/40 (2)	32/10 (3.2)	1	169/40 (4.22)	61/10 (6.1)	1	205/40 (5.13)	65/10 (6.5)	1
Patient 60 (D = M)	502/40 (12.55)	134 (13.4)	1	225/40 (5.63)	73/10 (7.3)	1	253/40 (6.32)	79/10 (7.9)	1
Patient 61 (D = M)	104/40 (2.6)	36/10 (3.6)	1	141/40 (3.53)	47/10 (4.7)	1	177/40 (4.43)	51/10 (5.1)	1
Patient 62 (D = B)	118/40 (2.95)	20/10 (2)	0	159/40 (3.97)	35/10 (3.5)	0	187/40 (4.68)	41/10 (4.1)	0
Patient 63 (D = B)	148/40 (3.7)	36/10 (3.6)	0	185/40 (4.63)	51/10 (5.1)	1	221/40 (5.53)	55/10 (5.5)	0
<b>C-Quality</b>			<b>100%</b>			<b>92.3%</b>			<b>92.3%</b>

TABLE XII  
SUMMARY OF ACCURACY RESULTS

Test Set	Feature Set	V	VI	VII
10 internal patients	Classification quality	100%	100%	100%
	Diagnostic accuracy	100%	100%	100%
13 additional patients	Classification quality	100%	92.3%	92.3%
	Diagnostic accuracy	100%	100%	100%
Total test set (10 + 13)	Classification quality	100%	95.6%	95.6%
	Diagnostic accuracy	100%	100%	100%
Total test set (10 + 13)	<b>Combined algorithm classification quality</b>	<b>91.3%</b>		
	<b>Combined algorithm diagnostic accuracy</b>	<b>100%</b>		

TABLE XIII  
TRAINING DATA SET FOR 50 PATIENTS

No.	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	D
51	79	2	12.6	4	0	3	1	1	1	2		22	0	113	109	83	144	16.8	M
52	64	1.5	4.9	1	0	3	0	1	1	1		26	0	83	104	71	93	0	M
53	76	5	78.5	4	0	2	1	0	1	2	40	27	0	77		72	52	10.8	M
54	69	1	3.1	1	1	3	2	0	0	2		22	1	40	42	85	13	4.2	M
55	37	1.8	10.2	3	0	3	0	0	1	1		25	1	78	87	108	30	0	B
56	54	2	12.6	2	0	3	0	1	0	2	20		0	89	100			2.5	B
57	75	3	28.3	1	0	3	1	0	0	2	0	38	1					5.5	M
58	81	3	28.3	2	0	2	1	1	1	2		33	1	24	67	44	13	12.8	M
59	76	4	50.3	2	0	3	3	0	1	2		35	0	52	48	62	26	12.3	M
60	59	4	36.3	3	0	2	4	1	0	2		26	0	89		104	38	7.25	M
61	57	1	3.1	3	0	3	0	1	1	2	30	26	0	82	74	87		8.7	M
62	76	3.5	33.2	2	0	2	1	0	0	2		20	0	96	60	110	31	9.6	B
63	65	1.5	7.1	1	0	3	0	0	0	2		25	1	49	64	71	16	2.9	B

28, 33, 42, and 44 from the 50-patient set of Table XIII. These patients were selected according to the cross-validation guidelines discussed in [20]. 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.

C. 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 70% of all patients. The classification quality for each of the 18 features is shown in Table VIII.

The classification quality ratios in Table VIII 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 2–18 features. Some of these feature sets meet the definition of reduct and some are random modifications of the reducts. The last (Xth) feature set includes all 18 features. Of course, all reducts and their supersets have the classification quality ratio of 100%, while some of

the feature sets have classification quality less than 100%. The selected feature sets and their classification quality are shown in Table IX.

To test the confirmation algorithm, a subset of 10 patients (2, 3, 14, 15, 24, 27, 28, 33, 42, and 44), we randomly 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.

The results produced by the confirmation algorithm are shown in Table X(a)–(c).

Under each feature set in Table X(a), e.g., set I with features 15 and 16, there are three numbers. The numbers under D = B and D = M are the average distances of Step 3 of the confirmation algorithm. The number under ‘Diagnosis: Min{(D = B), (D = M)}’ is the diagnosis corresponding to the group with the minimum average distance from Step 4 of the same algorithm. The test patients, each with the diagnosis D confirmed with pathology test, are listed as rows of Table X. The last row includes the classification quality (C-Quality) of each feature set. At the end of Table X, the algorithm classification quality and diagnostic accuracy are provided. The algorithm classification quality is calculated as the minimum of the classification quality of the three individual feature sets. The two algorithms recommend no diagnostic decision whenever a disagreement occurs

TABLE XIV  
TEST DATA SET FOR 13 PATIENTS

Patient No.	Age		CT max r		CT max area		Borders		Calcification type		Location in thorax		Nodes		Other sus. Lesions		1 = M, 0 = F		PET PN imag > bg, 1 = Yes, 2 = No		Pk/yr		BMI		hx of ca, Yes = 1, No = 0		FEV1%		DLCO% ADJ		FVC%		FEF25-75%		SUV		M = Malignant, B = Benign	
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	D																			
1	68	1	3.1	1	3	2	3	1	1	2	150	27	0	104	79	97	79	12	M																			
2	73	1.5	5.7	1	0	3	1	1	0	2	5	37	1	99	108	92	91	9.4	M																			
3	57	1.2	3.8	3	0	3	0	0	1	1	20	24	0	90	58	12	41	0	B																			
4	71	0.75	1.8	1	0	2	0	0	1	1	20	23	0	55	118	78	25	0	B																			
5	85	1.6	6.2	2	0	3	4	0	0	2	140	19	0	60	100	61	45	3.4	M																			
6	65	3	28.3	4	0	2	0	0	0	2	60	30	0	87	71	103	34	11.4	M																			
7	39	4	38.5	2	0	1	0	0	1	2	40	37	0	33	41	25	18	10.1	M																			
8	57	1.8	10.2	1	0	2	3	0	0	2	30	26	1	97	104	85	98	11	M																			
9	49	2.5	19.6	1	0	3	0	0	1	2	30	42	0	104	106	106	91	3.6	B																			
10	72	1.5	7.1	2	0	2	0	0	1	2	20	30	0	64	71	113	31	4.6	M																			
11	77	3	28.3	2	0	3	0	0	1	2	50	26	1	116	45	107	80	5.5	M																			
12	73	1	3.1	2	0	3	0	1	1	2	0	25	1	96	113	91	67	9.1	M																			
13	62	2	12.6	1	0	2	2	1	1	2	100	21	1	50	89	68	22	10.3	M																			
14	64	2	12.6	2	0	2	2	1	0	2	60	23	1	75	58	82	37	5.1	M																			
15	71	4	50.3	2	0	1	1	0	0	2	100	39	0	94	62	88	73	4.6	M																			
16	76	0.8	2.0	3	0	2	4	0	0	2	25	20	1	76	69	112	16	1.2	M																			
17	34	4	38.5	2	0	3	0	0	0	2	0	23	1	104	117	102	92	18	M																			
18	72	2	9.6	2	0	2	1	1	0	2	30	27	0	37	59	43	20	9.3	M																			
19	70	1.5	4.9	2	0	3	0	0	0	2	0	20	0	109	109	102	84	4.3	B																			
20	74	3	28.3	2	0	2	0	0	1	2	20	27	0	106	86	92	87	2.25	M																			
21	34	2	9.6	3	0	1	0	0	0	2	0	19	1	99	96	101	78	1.8	M																			
22	72	2	7.1	4	0	2	3	1	1	2	100	26	0	56	56	77	24	3.5	M																			
23	73	2	7.1	4	0	3	0	1	1	2	30	24	0	68	90	59	66	9.9	M																			
24	70	1	3.1	1	0	3	0	0	0	2	0	29	0	112	84	109	77	2.6	M																			
25	75	2	12.6	2	0	3	0	1	1	2	24	26	0	32	56	57	15	9.2	M																			
26	55	1	3.1	2	0	2	2	1	0	1	70	22	1	79	77	97	31	0	B																			
27	79	3	28.3	3	0	2	0	0	1	2	35	33	1	44	61	61	24	6.4	M																			
28	61	5	78.5	2	0	1	0	0	0	2	0	29	1	74	58	63	74	9.5	M																			
29	69	1	3.1	1	0	3	4	0	1	2	60	24	0	93	88	99	50	13.7	M																			
30	72	2	12.6	4	0	2	1	0	1	2	50	24	0	102	78	80	31	7.1	M																			
31	69	1.2	3.8	2	0	2	0	0	0	2	50	32	0	74	51	92	25	2.8	M																			
32	64	1.5	4.9	3	0	1	0	0	0	2	75	37	0	53	63	80	21	6.1	M																			
33	64	1	3.1	3	0	3	0	1	1	2	0	31	1	104	115	116	56	7.8	B																			
34	49	5	50.3	1	0	1	1	0	1	2	30	27	0	75	114	69	65	6.6	M																			
35	64	4	38.5	1	0	3	1	1	0	2	92	25	0	70	80	88	24	17.6	M																			
36	66	3.5	33.2	4	0	3	2	0	1	2	81	37	0	88	67	83	59	3.6	M																			
37	66	6	95.0	2	0	2	0	0	0	2	50	20	0	90	52	98	59	3.3	M																			
38	76	5	50.3	1	0	1	1	1	0	2	60	30	1	87	79	87	79	7.1	M																			
39	62	1	3.1	3	0	3	0	0	0	2	40	28	0	79	61	93	32	17.6	M																			
40	56	1.5	4.9	1	0	2	0	0	0	1	0	21	1	109	105	107	85	0	B																			
41	68	2.5	19.6	2	1	3	0	1	1	2	60	26	0	115	136	114	76	6.3	M																			
42	72	2.5	19.6	3	0	3	2	0	1	2	30	28	1	56	59	57	36	3.2	M																			
43	53	1	3.1	3	0	3	3	1	0	2	15	19	0	88	87	94	52	4	B																			
44	52	3.2	30.2	2	0	2	0	0	1	2	35	25	1	92	109	105	46	7.6	M																			
45	71	1	3.1	3	0	3	0	1	0	2	50	22	0	52	48	81	15	11.4	M																			
46	54	1	3.1	2	0	3	1	0	0	1	41	32	0	44	81	44	28	0	B																			
47	47	2.2	10.8	4	0	2	0	1	0	2	30	22	1	81	89	100	37	2.4	B																			
48	70	3	28.3	4	0	1	0	1	1	2	90	20	0	89	82	90	34	13.6	M																			
49	45	2	9.6	1	0	2	0	0	0	2	11	23	0	121	128	124	86	2.2	M																			
50	64	0.5	0.8	1	0	3	1	1	0	2	90	46	0	75	69	84	33	4.2	M																			

in the diagnosis based on any of the feature sets V, VI, and VII, called here proper feature sets.

In addition to testing the ten randomly selected patients, we considered 13 additional patients from the 118-patient data set

(see Table XIV for the original data). These patients had sufficient data points to be tested on three feature sets V, VI, and VII that produced 100% classification quality in Table X. The test results of the 13 additional patients are shown in Table XI.

#### D. Observations

The following observations can be made from the information included in Tables X and XI.

- The lowest classification quality of the confirmation algorithm is 60%, which is the highest accuracy rate reported in clinical diagnosis.
- The highest classification quality of the confirmation algorithm is 100%, which corresponds to the three proper feature sets V, VI, and VII in Table X and feature set V in Table XI.
- Note that the following relationship holds for the three feature sets in Tables X and XI:  $V \subseteq VI \subseteq VII$ .
- The feature sets V and VI are not reducts and have classification quality less than 100% in Table IX. They are considered as “inferior” in the data mining literature, yet the feature set V with the classification rate of 68% in Table IX resulted in 100% classification quality with the confirmation algorithm (see Tables X and XI). This observation might be key to future developments in data mining.
- The classification quality of individual features included in the feature set V and the feature sets VI and VII is low (see Table VIII). In fact, the classification quality of each individual feature in set V is 0%.
- The classification quality produced by the confirmation algorithm for the feature sets I and II in Table X(a), which are reducts, is only 60% and 70%. Combining the two reducts in feature set III in Table X(a) has resulted in 60% classification quality by the confirmation algorithm.
- The classification quality by the confirmation algorithm with all features [see Table X(c)] 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.
- Patient 24 and 44 produced the largest number of errors in the algorithmic classification. The feature values of these patients need attention. Removing the two patients from the training set would certainly improve the classification quality of the confirmation algorithm.

The accuracy results produced by the primary and confirmation algorithms are summarized in Table XII.

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. 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.

#### REFERENCES

- [1] P. Adriaans and D. Zantinge, *Data Mining*. New York: Addison Wesley, 1996.
- [2] T. J. Gross, Solitary Pulmonary Nodule, Virtual Hospital, The University of Iowa, Iowa City, IA, 1997.
- [3] J. W. Grzymala-Busse, “A new version of the rule induction system LEERS,” *Fundamenta Informaticae*, vol. 31, pp. 27–39, 1997.
- [4] N. C. Gupta, J. Maloof, and E. Gunel, “Probability of malignancy in solitary pulmonary nodules using fluorine-18-FDG and PET,” *J. Nucl. Medicine*, vol. 37, no. 6, pp. 943–948, 1996.
- [5] K. F. Hubner, E. Buonocore, H. R. Gould, J. Thie, G. T. Smith, S. Stephens, and J. Dickey, “Differentiating benign from malignant lung lesions using “Quantitative” parameters of FDG-PET images,” *Clin. Nucl. Medicine*, vol. 21, no. 12, pp. 941–949, 1996.
- [6] H. F. Hourini, M. A. Meziane, and E. A. Zerhouni *et al.*, “The solitary pulmonary nodule: Assessment, diagnosis and management,” *Chest*, vol. 91, p. 128, 1987.
- [7] I. Kononenko, I. Bratko, and M. Kokar, “Application of machine learning to medical diagnosis,” in *Machine Learning in Data Mining: Methods and Applications*, R. S. Michalski, I. Bratko, and M. Kubat, Eds. New York: Wiley, 1998, pp. 389–428.
- [8] W. Kowalczyk and F. Slisser, “Modeling customer retention with rough data models,” in *Proc. First Eur. Symp. PKDD '97*, Trondheim, Norway, 1997, pp. 4–13.
- [9] A. Kusiak, *Computational Intelligence in Design and Manufacturing*. New York: Wiley, 2000.
- [10] S. A. Landis, T. Murray, S. Bolden, and P. A. Wingo, *Cancer J. Clinicians*, vol. 49, pp. 8–31, 1999.
- [11] G. A. Lillington, “Management of the solitary pulmonary nodule,” *Hospital Practice*, pp. 41–48, May 1993.
- [12] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Boston, MA: Kluwer, 1991.
- [13] —, “Rough sets and data mining,” in *Proc. Australasia-Pacific Forum on Intelligent Processing and Manufacturing of Materials*, vol. 1, T. Chandra, S. R. Leclair, J. A. Meech, B. Varma, M. Smith, and B. Balachandran, Eds., Gold Coast, Australia, 1997, pp. 663–667.
- [14] Z. Pawlak, K. Slowinski, and R. Slowinski, “Rough classification of patients after highly selective vagotomy for duodenal ulcer,” *Int. J. Man-Machine Studies*, vol. 24, pp. 413–433, 1998.
- [15] G. Ruhe, “Qualitative analysis of software engineering data using rough sets,” in *Proc. Fourth Int. Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery*, Tokyo, Japan, 1996, pp. 292–299.
- [16] N. Shan, W. Ziarko, H. J. Hamilton, and N. Cercone, “Using rough sets as tools for knowledge discovery,” in *Proc. First Int. Conf. Knowledge Discovery and Data Mining*, U. M. Fayyad and R. Uthurusamy, Eds. Menlo Park, CA, 1995, pp. 263–268.
- [17] A. Skowron, “Data filtration: A rough set approach,” in *Proc. Int. Workshop on Rough Sets and Knowledge Discovery*. Alberta, Canada, 1994, pp. 108–118.
- [18] R. Slowinski and J. Stefanowski, “Rough classification with valued closeness relation,” in *New Approaches in Classification and Data Analysis*. New York: Springer-Verlag, 1993, pp. 482–489.
- [19] J. Stefanowski and K. Slowinski, “Rough sets as a tool for studying attribute dependencies in the urinary stones treatment data set,” in *Rough Sets and Data Mining: Analysis and Imprecise Data*, T. Y. Lin and N. Cercone, Eds. Dordrecht, The Netherlands: Kluwer, 1997, pp. 177–196.
- [20] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *J. Royal Statistic. Soc.*, vol. 36, pp. 11–147, 1974.

- [21] S. J. Swensen, M. D. Silverstein, E. S. Edell, V. F. Trastek, G. L. Augenbaugh, D. M. Ilstrup, and C. D. Scheleck, "Solitary pulmonary nodules: Clinical prediction model versus physicians," in *Mayo Clinic Proc.*, vol. 74, 1999, pp. 319–329.
- [22] S. Tsumoto, "Extraction of experts' decision process from clinical databases using rough set model," in *Proc. First Eur. Symp. PKDD '97*, Trondheim, Norway, 1997, pp. 58–67.
- [23] S. Tsumoto and W. Ziarko, "The application of rough sets-data mining based technique to differential diagnosis of meningococcal meningitis," in *Proc. 9th Int. Symp. Foundations on Intelligent Systems*, Zakopane, Poland, 1996, pp. 438–447.
- [24] Z. M. Wojcik, "Edge detector free of the detection/localization tradeoff using rough sets," in *Proc. Int. Workshop on Rough Sets and Knowledge Discovery*. Alberta, Canada, 1993, pp. 421–438.
- [25] UIHC, internal cost data, .
- [26] G. A. Klein, "A recognition-primed decision (RPD) model of rapid decision making," in *Decision Making in Action: Models and Methods*, G. A. Klein, Ed. Norwood, NJ: Ablex, pp. 138–147.



**Andrew Kusiak** is Professor of Industrial Engineering at the University of Iowa, Iowa City. He is interested in product development, manufacturing, healthcare informatics and technology, and applications of computational intelligence and optimization. He has published research papers in journals sponsored by AAAI, ASME, IEEE, IIE, INFORMS, ESOR, IFIP, IFAC, IPE, ISPE, and SME. He speaks frequently at international meetings, conducts professional seminars, and consults for industrial corporations. He serves on the editorial boards of sixteen journals, edits book series, and is the Editor-in-Chief of the *Journal of Intelligent Manufacturing*.



**Jeffrey A. Kern** is Professor of Medicine at the University of Iowa, Iowa City. He studies the regulation of epithelial cell growth as it relates to lung development and tumorigenesis. His clinical expertise is in the diagnosis and treatment of patients with lung cancer. He has published research papers on both the basic science and clinical problems of lung cancer in journals sponsored by the American Association of Immunology, National Academy of Science, American Association of Cancer Research, American Thoracic Society, American Association of Chest Physicians. He speaks frequently at international meetings and academic centers and serves as a consultant to biotechnology corporations.

**Kemp H. Kernstine** is a graduate of Duke University Medical School in 1982. He completed his internship and residency in General Surgery in 1990 from the University of Minnesota Hospitals and Clinics.

He is a Cardiothoracic Surgeon and the director of the General Thoracic Surgery Service at the University of Iowa Hospitals and Clinics. After completing his surgical residency, he received specialty training in Cardiothoracic Surgery at the Brigham and Women's Hospital, Harvard Medical School, completed in 1994 with further subspecialty training in General Thoracic Surgery, Lung Transplantation and Thoracic Oncology. He currently is involved in numerous projects concerning the diagnosis, staging and treatment of cancers of the lung, esophagus and other chest cancers.



**Tzu-Liang (Bill) Tseng** received the M.S. degree in industrial engineering from The University of Wisconsin at Madison and the Ph.D. degree in industrial engineering from the University of Iowa, Iowa City, in 1999.

He is Assistant Professor of Industrial Technology at Western Kentucky University, Bowling Green, Kentucky. His current research focuses on the data mining, fuzzy set theory, and their applications in manufacturing and concurrent engineering.